A BAYESIAN SEMIPARAMETRIC APPROACH TO ESTIMATING A BACTERIUM'S WILD-TYPE DISTRIBUTION AND PREVALENCE: ACCOUNTING FOR CONTAMINATION AND MEASUREMENT ERROR

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Will A. Eagan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Bruce A. Craig, Chair

Department of Statistics

Dr. Chuanhai Liu

Department of Statistics

Dr. George P. McCabe

Department of Statistics

Dr. Arman Sabbaghi

Department of Statistics

Dr. John Turnidge

European Committee on Antimicrobial Susceptibility Testing

Approved by:

Dr. Jun Xie

Head of the School Graduate Program

This dissertation would not be possible without the dedicated and talented medical staff of Brigham and Women's Hospital in Boston, Massachusetts. In particular, I would like to thank Dr. Ronald Bleday and Dr. Melissa Murphy for their outstanding work.

ACKNOWLEDGMENTS

I have had the tremendous privilege to interact with so many enthusiastic students, dedicated staff, and erudite faculty throughout my graduate career. I wish to highlight and thank a select few that have had significant impact on my graduate education.

First and foremost, I would like to thank my advisor, Professor Bruce A. Craig. His encouragement and guidance allowed me to not only survive my time, but thrive as an emerging researcher. This thesis is an embodiment of his mentorship, guidance, and encouragement.

I would like to thank the past and present leadership of the Department of Statistics during my tenure. (Now) Dean Rebecca Doerge was so welcoming to me when I matriculated and was never afraid of showing the tough love of a mentor. Professor Thomas Sellke was a constant source of encouragement, understanding, and humor. Professor Hao Zhang served as a great source of advice and generously supported me with funding alongside Dean Doerge. I thank the Department of Statistics, College of Science, the Graduate School, the Purdue Research Foundation, and the Elihu Root Fellowship from Hamilton College for funding. For travel funds to attend conferences, I thank the Department of Statistics, the American Statistical Association (ASA), Midwest Biopharmaceutical Statistical Workshop (MBSW), the ASA Biopharmaceutical Section, and the Institute of Mathematical Statistics (IMS).

For technical support, Doug Crabill was invaluable in teaching me to utilize Purdue's clusters for my simulation studies. He was always there to offer help. I certainly enjoyed my chances to participate in the Wednesday night applied probability seminar.

I would like to thank my committee members: Professors George McCabe, Chuanhai Liu, Arman Sabbaghi, and John Turnidge. I would like to thank Professor Sabbaghi for very generously allowing me to present at his group's meeting regularly. The

feedback from his group and him was critical to my success. I would like to thank John Turnidge for very generously serving as an external member of my committee. He provided the data sets used in my real data analysis. Also, he provided helpful insight into designing my realistic simulations as a subject matter expert.

There are several other faculty I would like to acknowledge. My masters committee comprised of Professors Michael Zhu, Xiao Wang, and Jun Xie (Chair) were always there to provide help and guide me. Professor Mark Daniel Ward was critical to my success on the qualifying exams. He was a major source of inspiration in terms of statistical leadership and his efforts elevated the quality of my solutions. He provided me with a greater appreciation for probability. I also like to thank Professor Min Zhang who provided me an opportunity work for her Data Science boot camp and allowed me to present my research in the Bioinformatics Seminar. I was very fortunate to have many great discussions on qualifying exam problems with the late Professor Herman Rubin.

There are several other individuals I would like to thank. Dr. Stijn Jaspers very kindly shared his R code for the the methodology in Jaspers et al. [2016a]. Marcy Hackenbrack, a Senior Project Manager, at the Clinical and Laboratory Standards Institute (CLSI) very kindly directly me to related AST material available from the CLSI. I would like to thank the European Food Safety Authority (EFSA) for providing me with their most recent annual reports. Dr. John Stelling, the Co-Founder of the World Health Organization (WHO) Collaborating Centre for Surveillance of Antimicrobial Resistance established at Brigham and Women's Hospital and Harvard Medical School, provided valuable discussion on the background material on the domain side.

There are several past and present graduate students I would like to thank. Several of my classmates were common project collaborators of mine or studied with me for the qualifying exams: Eric Gerber, Timothy Keaton, Nathan Hankey, Timothy Reese, Dominique (Williams) McDaniel, Min Ren, Evidence Matangi, Ryan Murphy, Bingrou (Alice) Zhou, Barret Schloerke, Putu Ayu Sudyanti, and Yuying (Rice) Song.

There are several other past graduate students who mentored me in many ways on a regular basis or served as regular sources of encouragement: Jingyi Zhu, Ching-Yun (Veavi) Chang, Lin-Yang (Mike) Cheng, Jeremy Troisi, Glenn DePalma, Xiaoguang Wang, Faye Zheng, Bing Yu, Jihwan Oh, and Wutao Wei. There are several peers who were always there to add advice namely Tracy Gonzalez, Siqi Liang, Tallis Bowers, Jeanine Gnang, Jungmin Shin, Ahmad Hakeem Abdul Wahab, Raquel De Souza Borges Ferreira, Wenbin Zhu, Yumin Zhang, and Yueyun Zhang. This list is by no means exhaustive.

In my time at Purdue, several lecturers have done a terrific job supervising my teaching. I would like to think Laura Cayon, Leonore Findsen, Piyas Chakraborty, Sarah Sellke, and Celeste Furtner for being such kind and encouraging bosses. Ellen Gundlach very generously served as my mentor for my teaching certificate and introduced me to a more formal view of statistics education.

There are several faculty I would like to thank that played a major part in my decision to pursue a Ph.D. I would like to thank Professors Larry Knop, Tim Kelly, and Zachariah Dietz for their encouragement to pursue statistics for graduate study. There are two researchers who introduced me to applied Bayesian statistics: Doctors Brian and Natalia Connolly. They were truly so generous with their time, effort, and patience. They convinced very early in my career the power of statistics in addressing meaningful scientific questions.

Finally, I would like to thank my family. My parents and sister were a huge source of encouragement, support, and even inspiration to me. I could not have done it without you. Lastly, I would like to thank one particular special lady, my wife, Sophie Sun, who encouraged, supported, and stood by me throughout this whole exciting adventure.

TABLE OF CONTENTS

Page

LIST OF TABLES x

LIST OF FIGURES xi

SYMBOLS xviii

ABBREVIATIONS xix

GLOSSARY xx

ABSTRACT xxi

1 **INTRODUCTION: ASSESSING THE PREVALENCE OF ANTIMICROBIAL RESISTANCE** 1
- 1.1 Background 1
- 1.2 Monitoring AMR 2
- 1.3 Minimum Inhibitory Concentration (MIC) Assay 4
  - 1.3.1 Assay Properties: Interval Censoring 6
  - 1.3.2 Assay Properties: Variability 6
- 1.4 Semiparametric Mixture Model 8
  - 1.4.1 Wild-Type Distribution 8
  - 1.4.2 Non-Wild-Type Distribution 9
  - 1.4.3 Inclusion of Measurement Error 9
- 1.5 The Epidemiological Cut-Off Value 11
- 1.6 Layout of dissertation 12

2 **OVERVIEW OF EXISTING METHODS** 1
- 2.1 Subset Methods 1
  - 2.1.1 Turnidge et al. Method 1
  - 2.1.2 Jaspers et al. Method 3
  - 2.1.3 Uncertainty Quantification of Estimates 4
  - 2.1.4 Limitations of Subset Methodologies 5
- 2.2 Semiparametric Mixture Model Methods 5
  - 2.2.1 Frequentist Mixture Model 6
  - 2.2.2 Bayesian Mixture Models 7
- 2.3 Summary 9

Page

3 **WILD-TYPE DISTRIBUTION AND PREVALENCE ESTIMATION VIA A BAYESIAN SEMIPARAMETRIC MODEL: ACCOUNTING FOR CONTAMINATION AND MEASUREMENT ERROR (BAYESACME)** . . . 11
3.1 BayesACME: The Motivation . . . 11
3.2 Model . . . 12
3.3 Estimation . . . 13
3.3.1 Priors for the WT Parameters . . . 14
3.3.2 Prior on σ^2_δ . . . 14
3.3.3 Prior for the NWT Parameters . . . 16
3.4 Computation . . . 19
3.4.1 Setting Initial Values . . . 19
3.4.2 Setting the Base Distribution . . . 20
3.4.3 Overview of MCMC Algorithm . . . 21
3.5 Choosing Between WT Models . . . 22
3.6 Single-Year Simulation Study . . . 23
3.6.1 Study Settings . . . 24
3.7 Single-Year Results . . . 27
3.7.1 Normal Case . . . 28
3.7.2 Log2gamma Case . . . 37
3.8 Real Data Set Application . . . 45
3.8.1 Conclusion . . . 48
4 **MULTIYEAR AST MONITORING** . . . 51
4.1 Previous Literature . . . 51
4.2 Multiyear Extensions of the Subset Methods . . . 52
4.3 Multiyear Extension of BayesACME . . . 53
4.3.1 Prior Selection . . . 55
4.4 Computation . . . 55
4.4.1 Setting Initial Values . . . 55
4.4.2 Updates to the Model Parameters . . . 56
4.5 Multiyear Simulation Study . . . 57
4.5.1 Normal Results . . . 61
4.5.2 Log2gamma Results . . . 65
4.6 Real Multiyear Data Set Application . . . 69
4.7 Conclusion . . . 72
5 **CONCLUDING REMARKS** . . . 74
5.1 Limitations . . . 75
5.2 Future Work . . . 76
5.2.1 Explicitly Including Between-Lab Variability . . . 77
5.2.2 Allowable Overlap Between the NWT and WT Distributions . . . 77
5.2.3 WT Model Comparison . . . 78

Page
5.2.4 Time Dependent Dirichlet Process Mixture Models 79
5.2.5 Considering a Hypersusceptible Subpopulation 80
5.2.6 Potential Changes to the MIC Assay 81
REFERENCES 82
A LATENT WT DISTRIBUTIONS 87
B DERIVATION OF MAXIMUM LIKELIHOOD ESTIMATES FOR JASPERS ET AL. SUBSET METHOD 90
C SINGLE-YEAR COMPUTATION 94
D SINGLE-YEAR SIMULATION RESULTS 99
E MULTIYEAR EXTENSIONS TO SUBSET METHODS 136
F MULTIYEAR COMPUTATION 138
G MULTIYEAR PREVALENCE ESTIMATION RESULTS 143
VITA 191

LIST OF TABLES

Table Page

1.1 Repeated measurements of the same quality control isolate of *E. coli* ATCC 25922 at 10 different laboratories. This table from Annis and Craig [2005b] illustrates the existence of assay variability. There is both within-lab (within row) and between-lab (within column) variability. 7

3.1 Contamination for the Continuous Normal Case 26

3.2 Contamination for the Censored Normal Case 26

3.3 Contamination for the Continuous log2gamma Case 27

3.4 Contamination for the Censored log2gamma Case 27

3.5 Numerical Summaries for Figure 3.4 30

3.6 Numerical Summaries for Figure 3.5 32

3.7 Numerical Summaries for Figure 3.6 34

3.8 Numerical Summaries for Figure 3.7 35

3.9 Numerical Summaries for Figure 3.9 38

3.10 Numerical Summaries for Figure 3.10 40

3.11 Numerical Summaries for Figure 3.11 42

3.12 Numerical Summaries for Figure 3.12 43

3.13 Results from the *N. gonorrhoeae* vs. Penicillin from a single Dutch lab in 2013 46

4.1 Changes in Wild-Type Prevalence over Four Years 58

4.2 Contamination for the Multiyear Simulation Study with ME 59

4.3 Numerical Summaries for Normal Distribution at Slow Rate 63

4.4 Numerical Summaries for Normal Distribution at Fast Rate 65

4.5 Numerical Summaries for log2gamma Distribution at Slow Rate 67

4.6 Numerical Summaries for log2gamma distribution at Fast Rate 69

4.7 Results from the *N. gonorrhoeae* vs. Penicillin from a Single Dutch Lab from 2011 to 2014 72

LIST OF FIGURES

Figure Page

1.1 Clinical results from the European Committee on Antimicrobial Susceptibilty Testing (EUCAST) for *E. coli* treated with Ampicillin up until 2020. The vertical axis displays the count and the horizontal axis corresponds with the observed $\log_2(MIC)$ results. 3

1.2 In the figure above, each row represents a different “drug/bug” combination. The first non-cloudy well in the row (read right to left in the figure) is declared the MIC. 5

1.3 This figure shows how the pdf of a log2gamma with $\alpha = 2.9686$ and $\beta = 4.0526$ becomes convolved with Normal ME. The different densities correspond with increasing values of σ_δ. As σ_δ grows the distribution becomes more Normal. 10

2.1 The shaded green area denotes the selected bins used in estimation of the wild-type parameters and prevalence. 2

2.2 This histogram produced using from Simulation 1 of Jaspers et al. [2016a] at size 5000 with estimated density curves. 8

3.1 Proposed prior density and the previous prior from Craig [1999] used to describe the ME standard deviation σ_δ. 15

3.2 This figure demonstrates the worst case contamination occurring on a latent true level in the Normal case. 17

3.3 Stacked histogram for contamination by bin where each row corresponds to a level of contamination. 25

3.4 This set of boxplots contrasts the results of the ECOFF values of the discussed methods at size 600 and $\sigma_\delta = 0.5$. 29

3.5 This set of boxplots contrasts the results of the prevalence values of the discussed methods at size 600 and $\sigma_\delta = 0.5$. 31

3.6 The boxplots show the results of the obs. ECOFF estimates of the methods at medium contamination and $\sigma_\delta = 0.5$. 33

3.7 These boxplots display the results of the prevalence values of the discussed methods at medium contamination and $\sigma_\delta = 0.5$. 35

Figure Page

3.8 The boxplots show the bias of the ME variance, latent WT variance, and the obs. WT standard deviation, each when $\sigma_\delta = 0.4, 0.5, 0.6$, respectively at low contamination and size 600. 36

3.9 The results of the ECOFF values of the compared methods at medium contamination and $\sigma_\delta = 0.5$. 37

3.10 Boxplots of results for prevalence of the compared methods at medium contamination and $\sigma_\delta = 0.5$ as sample size varies. Notice that the methods that misspecified the observed WT distribution are able to produce adequate estimates of prevalence. 39

3.11 Boxplots of estimates of the observed ECOFF at size 1200 with $\sigma_\delta = 0.5$ across contamination levels. Notice the drastic changes in TURNM versus the other methods. 41

3.12 Boxplots of estimates of prevalence at size 1200 with $\sigma_\delta = 0.5$. 42

3.13 Boxplots of bias in the obs. ECOFF at size 600 with low contamination across the three levels of σ_δ. 44

3.14 This histogram visualizes the 2013 collection of *N. gonorrhoeae* treated with Penicillin. 45

3.15 These are kernel densities of the resulting posterior distributions of the test statistic. The Normal case is the black dashed curve and the log2gamma distribution is the red-dashed curve. 47

3.16 For the Normal case case, this plot illustrates the probability of an isolated with the recorded MIC labelled a WT. The blue error bars are the 95% credible intervals. 48

4.1 This plot shows the trend for the fast (in blue) and slow (in red) rates of decline for the WT prevalence. 58

4.2 Normal distribution at low contamination at Size 300 at the slow rate of WT prevalence decline 61

4.3 Normal distribution at high contamination at size 300 at the slow rate of WT prevalence decline 62

4.4 Normal distribution at low contamination at Size 300 at the fast rate of WT prevalence decline 64

4.5 Normal distribution at high contamination at size 300 at the fast rate of WT prevalence decline 64

4.6 Log2gamma distribution at low contamination at size 300 at the slow rate of WT prevalence decline 66

Figure Page

4.7 Log2gamma distribution at high contamination at size 300 at the slow rate of WT prevalence decline . . . 66

4.8 Log2gamma distribution at low contamination at size 300 at the fast rate of WT prevalence decline . . . 68

4.9 Log2gamma distribution at high contamination at size 300 at the fast rate of WT prevalence decline . . . 68

4.10 Each year's MIC distribution is shown as a histogram. . . . 70

4.11 This plot is the estimate of prevalence versus time for the three extended methods from 2011 to 2014. The left column is the Normal case and the right column is the log2gamma case. . . . 71

A.1 The pdf of the Normal distribution for different parameters values. . . . 88

A.2 The pdf of the log2gamma distribution with different parameters values. The parameter α solely determines the spread. Both α and β determine the mean and mode. . . . 89

D.1 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.5$. . . . 100

D.2 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.5$. . . . 101

D.3 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.5$. . . . 102

D.4 This set of boxplots contrasts the estimates of the prevalence estimates of the discussed methods at size 300 and $\sigma_\delta = 0.5$. . . . 103

D.5 This set of boxplots contrasts the estimates of the prevalence estimates the discussed methods at size 600 and $\sigma_\delta = 0.5$. . . . 104

D.6 This set of boxplots contrasts the estimates of the prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.5$. . . . 105

D.7 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$. . . . 106

D.8 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$. . . . 107

D.9 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$. . . . 108

D.10 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$. . . . 109

Figure Page

D.11 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$. 110

D.12 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$. 111

D.13 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$. 112

D.14 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$. 113

D.15 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$. 114

D.16 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$. 115

D.17 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$. 116

D.18 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$. 117

D.19 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.5$. 118

D.20 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.5$. 119

D.21 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.5$. 120

D.22 This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 300 and $\sigma_\delta = 0.5$. 121

D.23 This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 600 and $\sigma_\delta = 0.5$. 122

D.24 This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 1200 and $\sigma_\delta = 0.5$. 123

D.25 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$. 124

D.26 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$. 125

D.27 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$. 126

Figure Page

D.28 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$. 127

D.29 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$. 128

D.30 This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$. 129

D.31 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 300 and $\sigma_\delta = 0.4$. 130

D.32 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 300 and $\sigma_\delta = 0.6$. 131

D.33 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 600 and $\sigma_\delta = 0.4$. 132

D.34 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 600 and $\sigma_\delta = 0.6$. 133

D.35 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.4$. 134

D.36 This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.6$. 135

G.1 Normal with Low Contamination at Size 300 143

G.2 Normal with Low Contamination at Size 600 144

G.3 Normal with Low Contamination at Size 1200 145

G.4 Normal with Medium Contamination at Size 300 146

G.5 Normal with Medium Contamination at Size 600 147

G.6 Normal with Medium Contamination at Size 1200 148

G.7 Normal with High Contamination at Size 300 149

G.8 Normal with High Contamination at Size 600 150

G.9 Normal with High Contamination at Size 1200 151

G.10 Trend for Normal case with Slow Rate of Decline at size 300 152

G.11 Trend for Normal case with Slow Rate of Decline at size 600 153

G.12 Trend for Normal case with Slow Rate of Decline at size 1200 154

G.13 Normal with Low Contamination at Size 300 155

Figure Page

G.14 Normal with Low Contamination at Size 600 156
G.15 Normal with Low Contamination at Size 1200 157
G.16 Normal with Medium Contamination at Size 300 158
G.17 Normal with Medium Contamination at Size 600 159
G.18 Normal with Medium Contamination at Size 1200 160
G.19 Normal with High Contamination at Size 300 161
G.20 Normal with High Contamination at Size 600 162
G.21 Normal with High Contamination at Size 1200 163
G.22 Trend for Normal case with Fast Rate of Decline at size 300 164
G.23 Trend for Normal case with Fast Rate of Decline at size 600 165
G.24 Trend for Normal case with Fast Rate of Decline at size 1200 166
G.25 log2gamma with Low Contamination at Size 300 167
G.26 log2gamma with Low Contamination at Size 600 168
G.27 log2gamma with Low Contamination at Size 1200 169
G.28 log2gamma with Medium Contamination at Size 300 170
G.29 log2gamma with Medium Contamination at Size 600 171
G.30 log2gamma with Medium Contamination at Size 1200 172
G.31 log2gamma with High Contamination at Size 300 173
G.32 log2gamma with High Contamination at Size 600 174
G.33 log2gamma with High Contamination at Size 1200 175
G.34 Trend for log2gamma case with Slow Rate of Decline at size 300 176
G.35 Trend for log2gamma case with Slow Rate of Decline at size 600 177
G.36 Trend for log2gamma case with Slow Rate of Decline at size 1200 178
G.37 log2gamma with Low Contamination at Size 300 179
G.38 log2gamma with Low Contamination at Size 600 180
G.39 log2gamma with Low Contamination at Size 1200 181
G.40 log2gamma with Medium Contamination at Size 300 182
G.41 log2gamma with Medium Contamination at Size 600 183

Figure Page

G.42 log2gamma with Medium Contamination at Size 1200 184
G.43 log2gamma with High Contamination at Size 300 185
G.44 log2gamma with High Contamination at Size 600 186
G.45 log2gamma with High Contamination at Size 1200 187
G.46 Trend for log2gamma case with Fast Rate of Decline at size 300 188
G.47 Trend for log2gamma case with Fast Rate of Decline at size 600 189
G.48 Trend for log2gamma case with Fast Rate of Decline at size 1200 190

SYMBOLS

α	Shape Parameter for the Gamma Distribution
α_{conc}	Concentration Parameter for the Dirichlet Process Mixture Model
β	Rate Parameter for the Gamma Distribution
δ	Measurement Error Term
ϵ	Between-Lab Error Term
μ	Mean parameter for the Normal Distribution
σ	Standard Deviation Parameter for the Normal Distribution
σ_δ	Standard Deviation of Measurement Error Term
σ_ϵ	Standard Deviation of Between-Lab Error Term
$\lceil\cdot\rceil$	Ceiling Function

ABBREVIATIONS

ACME	Accounting for Contamination and Measurement Error
AMR	Antimicrobial Resistance
AST	Antimicrobial Susceptibility Testing
CDC	Centers for Disease Control and Prevention
cdf	Cumulative Distribution Function
CLSI	Clinical and Laboratory Standards Institute
ECDC	European Centre for Disease Prevention and Control
EUCAST	European Committee on Antimicrobial Susceptibility Testing
MCMC	Markov Chain Monte Carlo
ME	Measurement Error
MIC	Minimum Inhibitory Concentration
NARMS	National Antimicrobial Resistance Monitoring System for Enteric Bacteria
NCCCLS	National Committee for Clinical Laboratory Standards
NWT	Non-Wild-Type
pdf	Probability Density Function
pmf	Probability Mass Function
QC	Quality Control
WHO	World Health Organization
WT	Wild-Type

GLOSSARY

Assay The analysis performed in this case to determine the MIC

Convolution The addition of two random variables

Prevalence Prevalence is defined as the proportion of the bacterium that are susceptible to a particular antimicrobial agent at a specific time. In some other sources, prevalence may refer to the proportion of a bacterium demonstrating resistance. In that case, the amount is one minus the stated prevalence

Wild-Type Organisms devoid of acquired resistance mechanisms. Non-Wild-Type is the opposite.

ABSTRACT

Eagan, Will A. Ph.D., Purdue University, December 2020. A Bayesian Semiparametric Approach to Estimating a Bacterium's Wild-Type Distribution and Prevalence: Accounting for Contamination and Measurement Error. Major Professor: Bruce A. Craig.

Antimicrobial resistance (AMR) is a major challenge to modern medicine and of grave concern to public health. To monitor AMR, researchers analyze "drug/bug" collections of clinical assay results to estimate AMR prevalence and the distribution of susceptible (wild-type) strains. This estimation is challenging because (a) the collection of assay results is a mixture of susceptible and resistant (non-wild-type) strains and (b) the most commonly used dilution assay produces interval-censored readings. To limit the effects of contamination from non-wild-type strains, methods have focused on using the counts in the K left-most bins, with K based on different heuristics. This limited use of the available data can result in the loss of precision and accuracy of model parameters. More recent methods have fit all the bin counts using a mixture model. These methods, however, struggle with identifiability and rely on penalization or informative priors to obtain reasonable estimates. In addition, none of the methods specifically account for the inherent assay variability, which has been shown to encompass a three-fold dilution range.

To account for this measurement error and utilize the full data set of bin counts, we propose a Bayesian semiparametric method to handle both single-year and multiyear studies. Similar to the previous mixture model methods, we model the wild-type distribution parametrically. Because less is known about the non-wild-type distribution, the proposed method uses a Dirichlet Process mixture model for the non-wild-type distribution. By accounting for measurement error we are able to impose biological constraints on the degree of overlap between the two underlying true distributions.

In doing this, we maintain identifiability. The feasibility of this approach and its improved precision and accuracy are demonstrated through simulation studies and an application to a real data set.

1. INTRODUCTION: ASSESSING THE PREVALENCE OF ANTIMICROBIAL RESISTANCE

1.1 Background

Antimicrobial resistance (AMR) is a natural phenomenon that has been accelerated through the overuse and misuse of antimicrobials in both humans and animals. AMR is now a major threat to the effective prevention and treatment of bacterial infections. In fact, the World Health Organization (WHO) recently predicted that the number of deaths from AMR will increase from 700,000 to 10 million by the year 2050 [Freedman, 2019]. If correct, AMR would become the leading cause of death among humans surpassing cancer, heart disease, and diabetes.

Relative to higher level organisms, bacteria evolve at a very fast rate. Each bacterium is made up of many different genetic variations, or strains. As these strains evolve through mutations, they develop resistance to various antibiotics. In fact, some strains of *Staphyloccus aureus*, *Escherichia coli*, and *Clostridium difficle* show resistance to virtually all tolerable antibiotic treatments.

The increase in AMR is due in part to the limited development of novel antimicrobials. According to Jinks [2017], the last new class of antibiotic was developed in 1984, with all "new" antibiotics since then just variations of previously developed drugs. Much of this can be attributed to the economic challenges that discourage the development of novel antibiotics. In a recent podcast, comparative pathobiologist Mohammed Seleem claimed that it can cost up to $2 billion and require as many as 15 years of research to produce a new antibiotic. Given that patients typically only take an antibiotic for a short amount of time, usually in intervals of 5, 10, or 15 days, recouping these development costs is slow. In contrast to therapies treating long-term chronic conditions, the economic model for antibiotics is not lucrative [Seleem, 2020].

Because novel antibiotic development lacks the economic incentives found in other areas of drug development, other approaches have been considered. For example, antibiotic "cocktails" have emerged as an effective treatment approach [Nield, 2018]. With this approach, a clinician prescribes several different antibiotics simultaneously to combat a single infection. Unfortunately this strategy can backfire. A recent study involving a two-drug cocktail revealed an increased chance of resistance to the second drug when already resistant to the first [Liu et al., 2020, Weintraub, 2020].

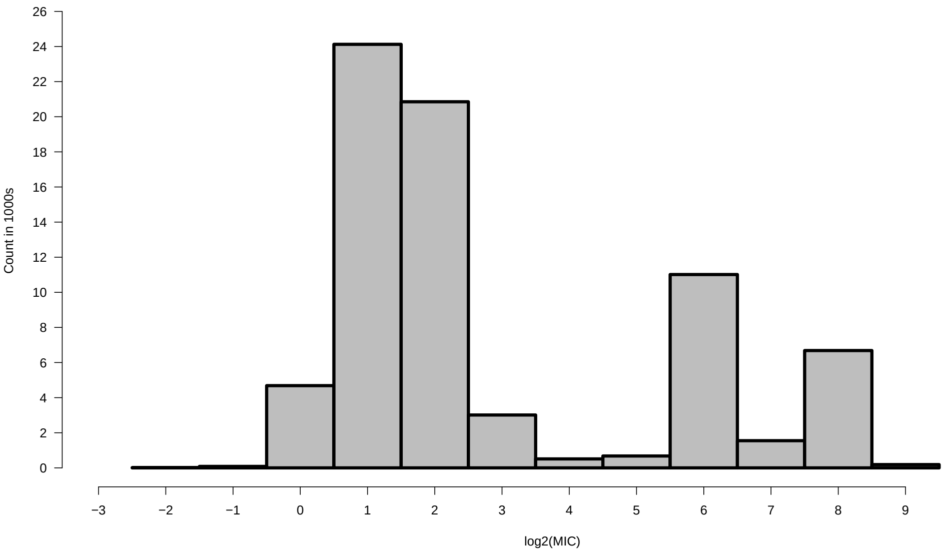
Given the dire predictions and the relative shortage of novel antibiotics, AMR monitoring is paramount. According to Fuhrmeister and Jones [2019], the three motivations for AMR monitoring are to (1) define the scope of AMR, (2) develop interventions to improve antimicrobial use, and (3) decrease the resistance selection pressure. Each of these motivations relies on there being adequate methods to measure AMR. That is the focus of this dissertation.

1.2 Monitoring AMR

When a patient arrives at a clinic or hospital suffering from an unknown bacterial illness, an antimicrobial assay is typically performed. This assay determines the potency levels of a variety of drugs needed to kill off or deter the growth of the cultured microorganism, or isolate. The smallest concentration of a drug that will deter the growth of the isolate is known as the minimum inhibitory concentration, or MIC. Some assays directly estimate this concentration, while others provide a proxy for this concentration. This general process is called antimicrobial susceptibility testing (AST).

There are two general goals of AST: (1) to predict the outcome of treatment using common dose levels of antimicrobial agents and (2) to guide the clinician in the selection of the most appropriate agent for a particular clinical problem [Turnidge, 2015]. In turn, collectively monitoring the growth of resistance among the isolates tested helps to assess the effectiveness of certain treatment policies and decide whether additional mitigation measures need to be taken. It also provides information to the government and could be used to enhance incentives. The Centers for Disease Control and Prevention (CDC), for example, uses lab results collected from hospitals and clinics to detect new infectious threats, to track trends, and to collaborate with appropriate responders [CDC, 2020].

Figure 1.1 is an example of one such collection of assay results for *E. coli* treated with Ampicillin. The assay summarized here considers two-fold dilutions of the drug so the results are usually summarized on the log base 2 scale as integers. Each bar in Figure 1.1 represents the count of isolates whose growth was first inhibited at this concentration (i.e., the observed MIC). It is very common for this collection of observed MICs to be multimodal.



Source: EUCAST (03−19−2020)

Figure 1.1. Clinical results from the European Committee on Antimicrobial Susceptibilty Testing (EUCAST) for *E. coli* treated with Ampicillin up until 2020. The vertical axis displays the count and the horizontal axis corresponds with the observed $\log_2(MIC)$ results.

When studying AMR, the underlying distribution of observed MIC results is considered a mixture of two subpopulations. On the left, is the wild-type (WT) subpopulation, which is a collection of isolates made up of those strains with no acquired resistance mechanisms (i.e., the natural population). The non-wild-type (NWT) subpopulation is a collection of isolates on the right, which represent the mutant strains that have developed antibiotic resistance.

If we denote the probability distribution function for the wild-type as $f_{WT}(Y)$ and the non-wild-type probability distribution function $f_{NWT}(Y)$. The mixture distribution is

$$f(Y) = \pi f_{WT}(Y) + (1 - \pi) f_{NWT}(Y)$$

where π represents the prevalence of the WT isolates. The goals of AMR monitoring depend on the accurate estimation of π and f_{WT}.

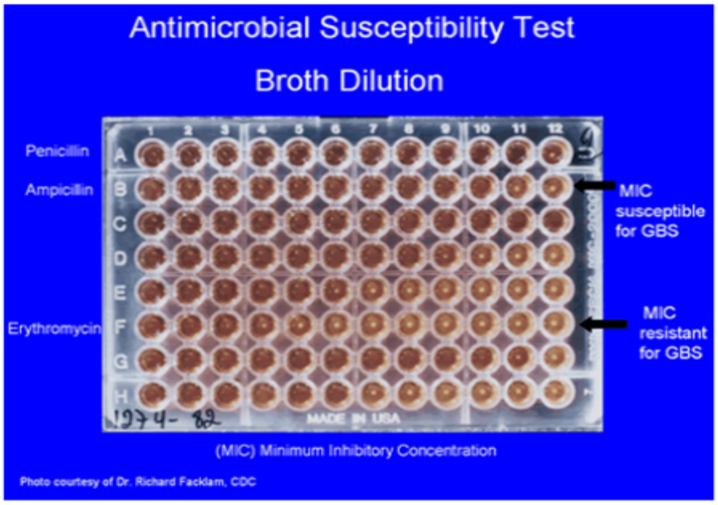
Over time, mutations cause phenotypic changes, thereby altering this mixture of observed assay results. The natural (WT) distribution remains the same [Kahlmeter et al., 2003], but the WT prevalence likely decreases. This, in turn, means that the NWT distribution gains new isolates that may alter its shape. We will use this description of evolution in Chapter 4 when we consider methods for monitoring AMR over time.

Note that these evolutionary changes occur on two levels: Genotypic and phenotypic. Our focus is on phenotypic changes detected by the assay. Results may appear identical phenotypically, even though the strains differ genetically. Also note that this mixture model is not appropriate for a species that is intrinsically resistant to a drug or shows "hypersusceptibility" creating a subpopulation to the left of the WT distribution [Harrison et al., 2019, Roemhild et al., 2020]. Discussion on how to handle the latter, albeit rare, situation can be found in the concluding chapter.

1.3 Minimum Inhibitory Concentration (MIC) Assay

To model a collection of assay results, we must first understand the properties of the assay being used. This dissertation focuses on the most common AST assay, the dilution assay. This assay considers several drugs at once and for each assesses the inhibitory strength of a set of two-fold drug concentrations. The result of the assay is the smallest two-fold dilution that visibly inhibits growth. This is an estimate of the MIC, explaining why it is often called the MIC assay [Zhou et al., 2009].

The assay is performed using a 96-well plate (Figure 1.2). Each well contains the isolate and some broth for sustenance. The wells in each row are serially diluted with a specific drug and then the plate is incubated for 16-20 hours in a temperature of 35° − 37°C. Afterward, each row in the plate is read by a clinician or machine to determine the dilution with the first non-cloudy well in the row. This is the dilution that has hindered growth.



Antimicrobial Susceptibility Test

Broth Dilution

(MIC) Minimum Inhibitory Concentration

Photo courtesy of Dr. Richard Facklam, CDC

Figure 1.2. In the figure above, each row represents a different "drug/bug" combination. The first non-cloudy well in the row (read right to left in the figure) is declared the MIC.

There are two types of panels for the MIC assay: "limited" and "broad." This dissertation only focuses on results from the "broad" panel. Vendors offer "limited" panels that consider a smaller number of concentrations (i.e., 2 to 5) in order to look at more antimicrobials at once. The "broad" panel typically considers 12 concentrations; the same as the number of columns of a 96-well plate.

1.3.1 Assay Properties: Interval Censoring

For each tested isolate, we can consider there being an underlying, continuous concentration X^* that will hinder growth. Because only 12 concentrations, $C_1, C_2, ..., C_{12}$, are used in the assay, the first concentration greater than this value of X^* is the assay result Y. Thus, the observed MIC value is linked to this continuous concentration as follows:

$$Y = \begin{cases} C_1 & \text{if } X^* \leq C_1 \\ C_j & \text{if } C_{j-1} < X^* \leq C_j \text{ for } j = 2, ..., 11 \\ C_{12} & \text{if } X^* > C_{11} \end{cases}$$

Except for the two extreme bins, the relationship is simply

$$Y = \lceil X^* \rceil$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

1.3.2 Assay Properties: Variability

To further understand Figure 1.1, we must consider the inherent variability of the MIC assay. Table 1.1 is an example of quality control (QC) data collected by the National Committee on Clinical Laboratory Standards (NCCLS), a precursor to the Clinical and Laboratory Standards Institute (CLSI). It summarizes the assay results of the same isolate analyzed 50 times in each of 10 labs around the United States. There is a common three-fold dilution range within a lab and comparable, but not identical, results across labs.

This three-fold range of results suggests that measurement error (ME) is a non-ignorable component of the MIC assay. Its incorporation into MIC analyses traces back to Craig [2000] and its impacts are discussed in Annis and Craig [2005a,b]. They consider X^* to be the sum of a true MIC value X and a Normal random variable δ

Table 1.1.
Repeated measurements of the same quality control isolate of *E. coli* ATCC 25922 at 10 different laboratories. This table from Annis and Craig [2005b] illustrates the existence of assay variability. There is both within-lab (within row) and between-lab (within column) variability.

	Observed MIC			
Lab	−8	−7	−6	−5
I	8	36	6	–
II	6	41	3	–
III	7	32	11	–
IV	–	48	2	–
V	2	48	–	–
VI	–	33	17	–
VII	7	41	2	–
VIII	–	15	35	–
IX	–	33	16	1
X	1	35	14	–
Combined	**31**	**362**	**106**	**1**

that represents this within-lab variability. The observed MIC can then be expressed as

$$Y = \lceil X + \delta \rceil$$

where $\delta \sim N(0, \sigma_\delta)$. Current monitoring methods focus on the mixture distribution in terms of X^*. The inclusion of ME and describing the mixture distribution in terms of X is a major motivator for this work.

The results of Table 1.1 also suggest that there is between-lab variability. At this time, however, MIC collections do not identify the lab so this and other sources of variability (e.g., technician and day) are all confounded with the within-lab effect. Because of this, we treat the collected results as if they are from one lab in this dissertation.

1.4 Semiparametric Mixture Model

In Section 1.2, we described the mixture distribution under consideration on the observed MIC (integer) scale. Having now linked a latent variable X^* to each Y, we can also describe the mixture model in terms of continuous densities. Specifically,

$$f(\mathbf{X}^*) = \pi f_{WT}(\mathbf{X}^* | \boldsymbol{\theta_{WT}}) + (1 - \pi) f_{NWT}(\mathbf{X}^* | \boldsymbol{\theta_{NWT}})$$

For the remainder of this dissertation, we use $\boldsymbol{\theta_{WT}}$, $\boldsymbol{\theta_{NWT}}$, and π to represent the WT distribution parameters, NWT distribution parameters, and WT distribution prevalence, respectively. It is on this scale that all current methods fit their models.

Given that $\mathbf{X}^* = \mathbf{X} + \boldsymbol{\delta}$, we describe the mixture density on the true latent scale as

$$f(\mathbf{X}) = \pi f_{WT}(\mathbf{X} \mid \boldsymbol{\theta_{WT}}, \sigma_\delta) + (1 - \pi) f_{NWT}(\mathbf{X} \mid \boldsymbol{\theta_{NWT}}, \sigma_\delta)$$

Notice that the ME standard deviation σ_δ is now an additional parameter of both the observed WT and observed NWT distributions. We conclude this section with a brief discussion on the difference modelling the mixture on the X^* and X scales but first provide some details on the models used to represent the two subpopulation distributions.

1.4.1 Wild-Type Distribution

On the concentration scale, the observed WT distribution is typically modeled using a logNormal distribution. The distribution's flexibility is described in both Lee and Whitmore [1999] and Craig [2000]. More recently, other distributions have been proposed, specifically the gamma distribution [Jaspers et al., 2014a]. This distribution can be justified as the selective sampling from a logNormal where those infected with low MIC isolates are less likely to get tested.

Because we will focus our analysis on the log base 2 scale, we will consider the Normal and log2gamma distributions for the WT distribution. Properties of each distribution that we use in this research are detailed in Appendix A.

1.4.2 Non-Wild-Type Distribution

The first modelling of the true latent MIC distribution was done as part of a diagnostic test calibration method [Craig, 2000, DePalma and Craig, 2018]. They used a mixture of Normals, where the number of components was an additional unknown parameter. Later Qi [2008] considered modeling this distribution using M-splines. These two versions of nonparametric modeling were used because the researchers had little *a priori* knowledge of the distribution shape and wanted a flexible approach to describe the multimodal distribution.

Although none of these methods specifically considered the true latent MIC distribution as a mixture of a WT and NWT distribution, the reasoning and approaches for modelling the NWT distribution in the context of monitoring AMR are similar. Both Jaspers et al. [2014b] and Grazian [2019] use a mixture of Normals with an unknown number of components to describe the NWT distribution of X^*. Jaspers et al. [2016a] use penalized B-splines (i.e., P-splines). In our approach will also consider the true NWT distribution as a mixture of Normals. Details of this model can be found in Chapter 3.

1.4.3 Inclusion of Measurement Error

As mentioned in the beginning of this subsection, current methods focus on estimating the WT distribution and its prevalence on the X^* scale. We, on the other hand, separate out ME and model the WT distribution on the X scale. The impact that this deconvolution has on the resulting estimates depends on the choice of distribution for the WT.

When the true WT distribution is Normal or $X \sim N(\mu_{WT}, \sigma_{WT})$, X^* becomes a Normal with mean μ_{WT} and standard deviation $\sqrt{\sigma^2_{WT} + \sigma^2_{\delta}}$. There is no change in distributional form but the variance of the resulting Normal distribution is larger. Estimation of π and μ_{WT} are therefore unaffected.

When the true WT distribution is log2gamma or $X \sim log2gamma(\alpha, \beta)$, inclusion of Normal ME means that the resulting distribution of X^* is neither log2gamma nor Normal. It is an example of a convolution whose density function is an intractable integral where g_X is the latent true WT density and h is the ME density:

$$f_{X^*}(z) = \int_{-\infty}^{\infty} g_X(s)h(z-s)ds$$

This means accounting for ME is necessary to estimate $\boldsymbol{\theta_{WT}}$ accurately. Approximating $f(X^*)$ as a log2gamma will lead to biased results. In Figure 1.3, densities of the convolution of a log2gamma with various levels of measurement error show that the distribution becomes increasingly Normal with larger σ_δ.

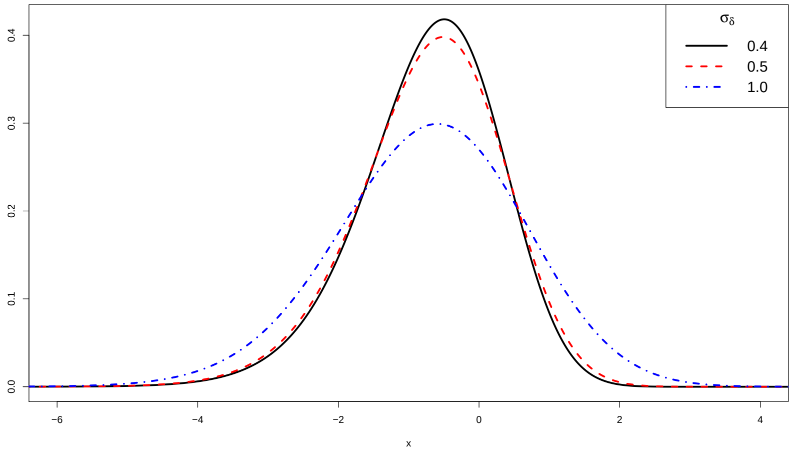


Figure 1.3. This figure shows how the pdf of a log2gamma with $\alpha = 2.9686$ and $\beta = 4.0526$ becomes convolved with Normal ME. The different densities correspond with increasing values of σ_δ. As σ_δ grows the distribution becomes more Normal.

1.5 The Epidemiological Cut-Off Value

In addition to monitoring changes in prevalence over time, microbiologists are interested in determining a value that effectively distinguishes the WT isolates on the left from the NWT isolates on the right. Initial efforts to establish this breakpoint resulted in different organizations and experts proposing competing values that partitioned the range of observed MIC values. These "eye-balled" partitions demarcated where the WT distribution ended and the NWT distribution began. In 2003, Kahlmeter et al. discussed the need for a unified value that was better suited for AMR monitoring. They proposed the Epidemiological Cut-off value (ECOFF) to be the threshold MIC value for declaring an isolate as phenotypically resistant [Kahlmeter et al., 2003]. It has become the standard, although in the CLSI documentation is often abbreviated as ECV.

It is important to distinguish the ECOFF from what is known as a clinical breakpoint. Clinical breakpoints divide the MIC distribution into regions of susceptible and resistant isolates. They take into account how the human body handles the drug (i.e., the pharmacodynamics). ECOFFs are used to detect emerging resistance mechanisms but do not necessarily imply resistance. In the EUCAST system, for example, a clinical breakpoint is a concentration that is at least an ECOFF. Monitoring of resistance can use either or both values, depending on the purpose of surveillance. As this dissertation is concerned with statistical estimation of the mixture distribution, its focus is on the ECOFF.

The development of the ECOFF marks a movement away from eye-balling histograms to one of modelling a distribution. For example, Turnidge et al. [2006] proposed the ECOFF as the 99.9^{th} percentile of the observed WT distribution. Determining this quantity provides motivation for statistical estimation of the WT distribution [Turnidge et al., 2006]. Thus the observed ECOFF is a function of $\boldsymbol{\theta_{WT}}$ and σ_δ. This is a scientific improvement over visual inspection of histograms as it is more reproducible and less subjective.

There is a model-based alternative to the ECOFF [Jaspers et al., 2014b, 2016a]. It relies on an estimate of the entire mixture distribution rather than just the WT distribution. Denoting the observed continuous WT distribution cumulative distribution function (cdf) as F_{WT} and the cdf of the overall mixture as F, the probability that isolate i with an observed MIC value of C_j is WT is

$$P(WT \mid Y_i = C_j) = \frac{\pi[F_{WT}(C_j) - F_{WT}(C_{j-1})]}{F(C_j) - F(C_{j-1})}$$

Through a threshold probability, such as 0.5, an isolate with observed MIC can be classified as WT or NWT.

1.6 Layout of dissertation

In Chapter 1, the importance and challenge of monitoring AMR using the MIC assay has been explained along with a description of the underlying mixture model. In Chapter 2, we describe the monitoring methods currently in the literature, discussing both their strengths and weaknesses. This is followed by a detailed description of our proposed method for both single-year (Chapter 3) and multiyear (Chapter 4) analyses. Each chapter outlines the model and estimation algorithm. Each chapter also includes a simulation study to compare our approach with those described in Chapter 2. We conclude with a summary of the research contributions along with some future research in Chapter 5.

2. OVERVIEW OF EXISTING METHODS

In this chapter, we detail the current AMR monitoring methods. We start with subset methods, which use only a fraction of the bin counts to estimate the parameters of interest. This is followed by a discussion of methods that use all the bin counts and estimate an entire mixture distribution. These latter methods all consider a semi-parametric mixture model like the one outlined in Chapter 1. The differences in the methods are the estimation and the approach to nonparametric density estimation.

When describing all methods, we denote the collection of clinical results as the bin counts, $m_1, m_2, ..., m_J$ ($\sum_j m_j = N_{tot}$) for concentrations $C_1, C_2, ..., C_J$, respectively. It is assumed that these counts result from the censoring of a continuous observed MIC mixture distribution, $f(X^*)$, with the WT distribution on the left. The goal is to estimate the parameters for the WT distribution and the WT prevalence.

2.1 Subset Methods

The subset methods take advantage of the fact that the left-most bin counts are almost entirely from the WT distribution. By focusing on these bins, these methods avoid the contamination by the NWT distribution on the right. The key differences in the two approaches are the method of estimation and the heuristic used to determine how many bins on the left to include.

2.1.1 Turnidge et al. Method

This method is by far the most straightforward and widely-used AMR monitoring approach, largely because it is available for download as an EXCEL macro on the European Committee on Antimicrobial Susceptibility Testing (EUCAST) and Clinical Laboratory and Standards Institute (CLSI) websites. The method fits a scaled continuous WT distribution to the cumulative counts of the K left-most bins using non-linear least squares. Denoting $B_j = \sum_{j'=1}^{j} m_{j'}$ as the cumulative count from the left for bin j, this approach finds $\boldsymbol{\theta_{WT}}$ and N (rather than π) that minimizes:

$$h(N, \boldsymbol{\theta_{WT}}) = \sum_{j=1}^{K} \left[B_j - N \cdot F_{WT}(C_j; \boldsymbol{\theta_{WT}})\right]^2$$

where F_{WT} is the cdf of the continuous WT distribution. The prevalence estimate is $\hat{\pi} = \frac{\hat{N}}{N_{tot}}$. The Excel macro, ECOFFfinder, uses the Solver function available in the Excel Analysis add-on to do this estimation/minimization. We, instead use the R programming language and a constrained BFGS optimization procedure.

The heuristic for determining K is to search over incremental subsets of bins and choose the subset that minimizes $|B_K - \hat{N}|$. Turnidge et al. [2006] mention that typically, but not always, the K that minimizes $|B_K - \hat{N}|$ corresponds with the subset that maximizes the absolute values of the standardized WT parameter estimates. When this was not the case, it was argued that differences in the estimated $\boldsymbol{\theta_{WT}}$ are minimal.

The method starts with those bins on the left up to the first bin to the right of the mode. It then iteratively adds a bin on the right and reestimates. For example, Figure 2.1 highlights the bins in green that are used in the first subset. The additional

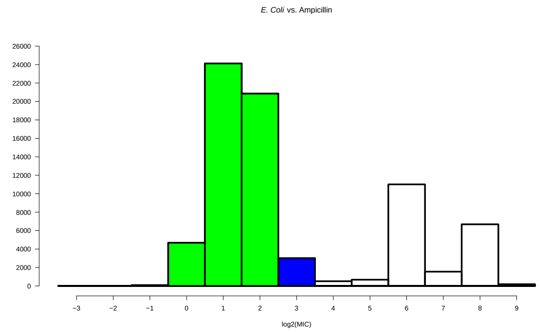


Figure 2.1. The shaded green area denotes the selected bins used in estimation of the wild-type parameters and prevalence.

bin highlighted in blue is added for the second subset, and so on. All remaining bins on the right are omitted from estimation.

Because N is an estimate of the number of isolates that are WT, this heuristic does very well when the WT and NWT distributions do not overlap and struggles when there is overlap. An alternative heuristic we consider in our simulation study is to always choose K to be the first bin after the first mode. This heuristic is less susceptible to contamination.

2.1.2 Jaspers et al. Method

This subset method cleverly converts the subset selection problem into a model selection problem. It selects among a series of multinomial models where the underlying bin probabilities depend on the WT and NWT distributions. Similar to Turnidge et al. [2006], the probabilities for the first K bins are based on the WT distribution and prevalence. The remaining bin probabilities have no restrictions. Maximum likelihood estimation is used to fit each model and the one with the lowest AIC is selected [Akaike, 1974]. Jaspers et al. [2014a] also use the AIC to choose between the Normal and log2gamma distributions. Once the distribution is chosen, they recommend averaging models for different values of K using AIC weights to better estimate $\boldsymbol{\theta_{WT}}$.

For each model, the log-likelihood to maximize is

$$\sum_{j=1}^{K} m_j \log(\tilde{p_j}) + \sum_{j=K+1}^{J} m_j \log(p_j) + \lambda(1 - \sum_{j=1}^{K} \tilde{p_j} - \sum_{j=K+1}^{J} p_j) \tag{2.1}$$

where the first K probabilities $\tilde{p_j}$ are

$$\tilde{p_j} = \begin{cases} \pi F_{WT}(C_j; \boldsymbol{\theta_{WT}}) & j = 1 \\ \pi \left[F_{WT}(C_j; \boldsymbol{\theta_{WT}}) - F_{WT}(C_{j-1}; \boldsymbol{\theta_{WT}})\right] & j = 2, 3, ..., K \end{cases} \tag{2.2}$$

and the remaining p_j have no constraints except $\sum_{j=1}^{K} \tilde{p_j} + \sum_{j=K+1}^{J} p_j = 1$. The full details of the derivation of the maximum likelihood estimates are in Appendix B.

Unlike the Turnidge et al. [2006] subset method, this approach also uses the total count in the bins to the right of K in its estimate of π.

In our simulation study, we alter the starting value of K to be the maximum of four bins and the number of bins that include the left-most mode. Without this modification, there is a chance of considering subsets that only use bins to the left of the first mode.

NWT Density Estimation

Motivated by the desire to use a model-based alternative to the ECOFF, Jaspers et al. [2014b] also propose a second step that utilizes the remaining $J - K$ bins to estimate a continuous, observed NWT distribution. This is done using a penalized mixture of Normals. This approach considers a large number of Normal densities centered at equidistant locations μ_q between C_K and C_J, each with standard deviation $\sigma_q = \frac{2}{3}(\mu_q - \mu_{q-1})$. To avoid overfitting, they introduce a penalty term based on the finite (m^{th}-order) differences of adjacent coefficients. The estimated observed MIC density is then

$$\hat{f}(x) = \hat{\pi} f_{WT}(x \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{WT}}) + (1 - \hat{\pi}) f_{NWT}(x \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{NWT}})$$

2.1.3 Uncertainty Quantification of Estimates

Both subset methods rely upon asymptotic theory or bootstrapping for uncertainty quantification. Bootstrapping is likely to be preferred in practice [Efron, 1981]. In fact, Jaspers et al. [2014a] quantify uncertainty in a simulation study using bootstrapping.

Bootstrapping would involve resampling with replacement L data sets of size N_{tot} and estimating the WT parameters. This is very easy to implement and makes quantification of the observed ECOFF straightforward. While one may be comfortable relying on asymptotic theory to obtain standard errors of the WT parameters, obtaining a standard error for the observed ECOFF would require implementing the delta method.

2.1.4 Limitations of Subset Methodologies

Both subset methods address NWT contamination by trying to avoid it. In Turnidge et al. [2006], that is literally what they do. In Jaspers et al. [2014a] they include the last $J-K$ bins, but without any distributional constraints. We argue that these approaches under-utilize the data, resulting in a loss of precision and possibly increased bias. In fact, Jaspers et al. [2014a] makes this same point arguing their approach is best implemented when the WT and NWT distributions have "clear separation."

Another crucial omission is that neither of these methods addresses measurement error. Both methods focus estimation on the latent observed MIC distribution assuming this distribution is either Normal or log2gamma. Outside of the Normal case, the inclusion of Normal measurement error alters the resulting latent observed MIC distribution. Not taking this into account will likely lead to biased estimates of $\boldsymbol{\theta_{WT}}$.

2.2 Semiparametric Mixture Model Methods

The remaining methods all consider fitting the mixture model

$$f(\mathbf{X}^*) = \pi f_{WT}(\mathbf{X}^* \mid \boldsymbol{\theta_{WT}}) + (1-\pi) f_{NWT}(\mathbf{X}^* \mid \boldsymbol{\theta_{NWT}})$$

to the collections of observed MIC results. In each approach, f_{NWT} is modelled nonparametrically and f_{WT} is considered Normal. The approaches again differ in terms of estimation approach and choice of nonparametric model.

2.2.1 Frequentist Mixture Model

Recognizing the limitations of their earlier work where separate bin regions were used to estimate the WT and NWT distributions, Jaspers et al. [2016b] consider joint estimation. The joint estimation is accomplished through an iterative algorithm of updating the estimates for $\boldsymbol{\theta_{WT}}$ and the weights (including π) for the NWT distribution.

Similar to Jaspers et al. [2014b], they consider a "generous" number of equidistant Normal components to describe the NWT distribution, but this time they span the entire range of concentrations. Nonparametric maximum likelihood estimation (NPMLE) is used to determine the mixing weights and WT prevalence given the estimates for $\boldsymbol{\theta_{WT}}$. Standard maximum likelihood is used to estimate $\boldsymbol{\theta_{WT}}$ given the mixing weights. They recommend using the Jaspers et al. subset method to obtain initial values of $\boldsymbol{\theta_{WT}}$ and start with the estimation of WT prevalence and the NWT mixing weights.

Limitations

From the paper discussion, it appears identifiability is a key issue with this method. Jaspers et al. [2016b] write that "...upon convergence, [WT prevalence] was occasionally decreased to zero and replaced with several nonparametric components." An *ad hoc* remedy is proposed that uses the knowledge of $\hat{\sigma}_{obs}$ (the observed WT standard deviation) from the subset method to add a penalization term. It is not clear how well their recommended penalization weight generalizes beyond the scope of the examples they discuss in their paper.

Currently, this approach is only described for the Normal WT distribution case. While measurement error does not impact the estimation of π and μ_{WT} and the observed ECOFF in the Normal setting, it would in other cases. Finally, the means and standard deviations in the components for the NWT distribution are pre-specified.

To offer more flexible modelling these parameters should be estimated jointly with the mixture weights.

2.2.2 Bayesian Mixture Models

Stijn Jaspers also led the development of a Bayesian semiparametric mixture model method. Following a similar set-up to the frequentist approach, this method focuses on a Normal WT distribution. Instead of considering a mixture of Normals for the NWT distribution, this method uses penalized B-splines (i.e., P-splines). They create a finer grid on the interval of C_1 to C_J, where each subinterval goes from χ_{i-1} to χ_i, $i = 1$ to I (e.g., $I = 100$). While the probability of the observed WT distribution in each subinterval is known, the probability associated with the NWT distribution is not. Thus, they use equally-spaced B-splines to fit a smoothed distribution over this grid. These bin probabilities are:

$$\pi[\Phi(\chi_i; \mu_{WT}, \sigma_{obs}) - \Phi(\chi_{i-1}; \mu_{WT}, \sigma_{obs})] + (1-\pi)\frac{\exp(\eta_i)}{\sum_{i=1}^{I}\exp(\eta_i)}$$

where the η_i are a product of the B-splines evaluated at the midpoints of the finer grid and the spline coefficient for subinterval i denoted ϕ_i. A penalty is placed on the vector of spline coefficients, $\boldsymbol{\phi}$, to have a smoothing effect on adjacent spline coefficients. Following the literature of Lang and Brezger [2004], the P-spline penalty is based on the r^{th} order differences of the spline coefficients, $\Delta^r\boldsymbol{\phi} \sim N(0, \tau^{-1/2})$. For the NWT distribution a Gamma prior placed on τ and an improper prior is specified for $\boldsymbol{\phi} \mid \tau$.

Similar to the frequentist mixture model, this model has an identifiability problem in regards to ensuring that the WT distribution accounts for almost all the density on the left. As a remedy, they consider “relatively” informative priors on μ_{WT}, σ_{obs}, and π. Specifically,

$$\mu_{WT} \sim N(\mu_\mu, \sigma_\mu)$$

$$\sigma_{obs} \sim InvGamma(\alpha_{\sigma}, \beta_{\sigma})$$

$$\pi \sim Beta(\alpha_{\pi}, \beta_{\pi})$$

where each of these prior parameters are based on the estimates obtained from the Jaspers et al. [2014a] subset method. The point estimates determine the mean of the corresponding prior distribution and the standard deviations were set to be slightly larger than the estimates of the standard errors at a sample size of 500.

The computation uses a Langevin-Hastings algorithm within Gibbs scheme. A Langevin-Hastings algorithm is akin to a Metropolis-Hastings algorithm, but brings in gradient information about the target posterior with the hopes of improving efficiency [Liang et al., 2011]. Jaspers et al. [2016a] direct the specific computational details to Atchadé et al. [2005], Haario et al. [2001], Lambert and Eilers [2009].

For illustration, an example is shown in Figure 2.2. This example involved 5000 MIC values from a mixture of three Normals with mean vector, $\boldsymbol{\mu} = (2.0, 4.5, 7.5)$, standard deviation vector, $\boldsymbol{\sigma} = (0.8, 0.7, 0.6)$, and weights $\mathbf{w} = (0.6, 0.2, 0.2)$.

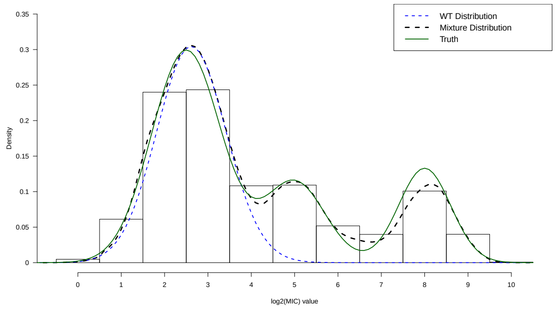


Figure 2.2. This histogram produced using from Simulation 1 of Jaspers et al. [2016a] at size 5000 with estimated density curves.

The first component in each of the three vectors refers to the WT distribution. In Figure 2.2, the blue dashed curve is the posterior mean for the WT distribution. The black dashed curve is the estimate of the mixture density. The green curve is the true density.

There are other Bayesian mixture models, but the approaches are fully nonparametric. The general idea is to fit the distribution with a finite Normal mixture model first proposed in Craig [2000] where the number of components was either known or estimated. Grazian [2019] places a default prior on the number of components to allow for a “flexible parametric” model. Thus, it follows the framework of modelling the NWT distribution as a mixture of Normals. As this method is described in a pre-print, the details of this method are likely to change. It does use a different prior than is required in other set-ups [DePalma and Craig, 2018].

Limitations

Others have noted that the use of P-splines can struggle to fit in a mixture with a peaked long tail. In that case, it may require additional smoothing parameters [Jullion and Lambert, 2007, Lambert and Eilers, 2009]. Similar to the other methods, this approach ignores ME. Because it only considers a Normal WT distribution, ME has no effect but the approach cannot be easily generalized.

There is an identifiability issue with the location of the WT and NWT distributions. This method makes the identifiablility of the WT component dependent solely on the priors for π and $\boldsymbol{\theta_{WT}}$. There is no condition preventing a spurious NWT component from forming to the left of the observed WT distribution.

Given that the NWT distribution is only defined on a finite grid, the WT distribution should be as well. This means they ought to consider a truncated Normal between C_1 and C_J. This is not done in the current algorithm so this method will struggle if the WT distribution left tail goes considerably off the grid.

2.3 Summary

The methods available for AMR monitoring can be broken into two groups: subset methods and mixture model methods. Subset methods, while simple and easy to implement, do not take full use of the data and are focused solely on estimation of the

WT distribution and its prevalence. These approaches work very well when the two subpopulations are clearly separated, but may struggle when there is contamination.

Mixture models, on the other hand, attempt to fit the entire distribution. Thus, they are better equipped to handle NWT contamination. The hopes are in that fitting the entire distribution, there can also be improvements in the estimation of the WT distribution and the ability to make comparisons across the two subpopulations.

Modelling this mixture, however, is complicated by the discrete nature of the data and current methods all have issues with identifiability. In fact, Jaspers et al. [2016a] discuss this issue for their two semiparametric approaches. Current workarounds require penalties and informative priors.

Finally, all the approaches described in this chapter focus on modelling the latent observed MIC distribution X^*. Given that ME is present in this assay, it makes more sense to deconvolve this ME from the true latent WT distribution. This guarantees a proper modelling of the distribution of X^* when X is non-Normal.

3. WILD-TYPE DISTRIBUTION AND PREVALENCE ESTIMATION VIA A BAYESIAN SEMIPARAMETRIC MODEL: ACCOUNTING FOR CONTAMINATION AND MEASUREMENT ERROR (BAYESACME)

In this chapter, we discuss our novel Bayesian semiparametric approach to single-year AST monitoring. We begin with a description of the model, followed by an outline of our estimation approach and a description of our WT distribution selection method. We then summarize an extensive simulation study comparing the performance of our method with the methods described in Chapter 2. The study considers various combinations of sample size, degree of measurement error, and amount of overlap between the latent true wild-type (WT) and non-wild-type (NWT) distributions. We conclude the chapter with an application to a real data set.

3.1 BayesACME: The Motivation

Similar to the mixture model approaches in Chapter 2, we consider there to be a latent continuous mixture distribution of observed MICs. Unlike these previous methods, we directly account for measurement error (ME), thereby allowing us to deconvolve this observed distribution into a latent true MIC distribution and a continuous ME distribution.

This deconvolution is important because it allows us to (1) properly model the continuous observed MIC distribution when the true WT distribution is non-Normal and (2) incorporate biological information into the prior of the mixture model. Specifically, we can maintain the proper ordering of the WT and NWT distributions, with overlap only occurring in their right and left tails, respectively. We can also limit the amount of overlap. We do this through the prior of the NWT distribution, thereby allowing us to maintain weak prior information on $\boldsymbol{\theta_{WT}}$ and π; something the Bayesian method proposed in Jaspers et al. [2016a] could not do.

3.2 Model

As with the other mixture model approaches, the observed bin counts, $m_1, ..., m_J$ are assumed to arise from a multinomial distribution:

$$m_1, ..., m_J \sim Multinomial(N_{tot}, p_1, ..., p_J)$$

where

$$p_j = \int_{C_{j-1}}^{C_j} (\pi f_{WT}(s) + (1-\pi) f_{NWT}(s)) ds$$

The C_j represent the assay concentrations with $C_0 = -\infty$ and $C_J = \infty$.

As with other methods, the densities f_{WT} and f_{NWT} represent the continuous observed MIC distributions that have been convolved with measurement error. The subset methods described in Chapter 2 assume f_{WT} is either Normal or log2gamma. The mixture model methods assume f_{WT} is Normal with the NWT distribution described nonparametrically as either a mixture of Normals or a combination of P-splines.

We, however, deconvolve this observed continuous response into a combination of signal and noise, specifically $X^* = X + \delta$. This deconvolution means the observed distribution $f(\cdot)$ is now convolution of a ME density $h(\cdot)$ and true MIC distribution $g(\cdot)$. In other words,

$$f(X^*) = \int_{-\infty}^{\infty} h(s) g(X^* - s) ds$$

Thus, the $\boldsymbol{\theta_{WT}}$ and the $\boldsymbol{\theta_{NWT}}$ in our approach represent the parameters of the true MIC distributions. We consider both the Normal and log2gamma distributions for g_{WT} and assume g_{NWT} is a mixture of Normals.

This decomposition also motivates an alternative expression of the model. For computational convenience, instead of X_i^* we introduce latent values X_i and δ_i to correspond with each Y_i. We also consider an indicator c_i, which denotes if isolate i is WT or not.

This allows us to specify the model as follows:

$$c_i \sim Bernoulli(\pi)$$

$$X_i \sim (g_{WT}(X_i))^{c_i}(g_{NWT}(X_i))^{1-c_i}$$

$$\delta_i \sim N(0, \sigma_\delta)$$

$$Y_i = \lceil X_i + \delta_i \rceil$$

The benefit of this model specification is that it makes working with non-Normal WT distributions far more straightforward. The likelihood can be expressed as

$$\prod_{i=1}^{N_{tot}} h^*(Y_i \mid X_i, \sigma_\delta) g(X_i \mid \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}, c_i) p(c_i \mid \pi)$$

where p is the Bernoulli pmf, g is either the Normal or log2gamma pdf, and h^* is a truncated Normal density. Because the truncated Normal is awkward to work with, we include the latent $\boldsymbol{\delta}$ (with the restrictions $Y_i = \lceil X_i + \delta_i \rceil$) in our estimation approach. As a result, the set of unknowns, $\{\mathbf{X}, \mathbf{c}, \boldsymbol{\delta}, \pi, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}, \sigma_\delta^2\}$, is large but estimation is much more computationally straightforward.

3.3 Estimation

For estimation, we take a Bayesian approach. The hierarchical structure of our model naturally fits this paradigm and the use of priors allows us to ensure biological realism, such as the WT distribution is to the left of the NWT distribution. We are also able to utilize quality control data to set an informative prior for σ_δ^2. The other advantage of a Bayesian approach is that the uncertainty of the parameters and functions of parameters (i.e., the observed ECOFF) is readily available.

3.3.1 Priors for the WT Parameters

For the WT distribution, we consider the Normal and log2gamma distributions. In both cases, we consider the use of Jeffreys priors. For the Normal distribution, this means

$$p(\mu_{WT}) \propto 1$$

and

$$p(\sigma^2_{WT}) \propto \frac{1}{\sigma^2_{WT}}$$

For the log2gamma distribution, this means

$$p(\alpha) \propto \sqrt{\psi^{(1)}(\alpha)}$$

and

$$p(\beta) \propto \frac{1}{\beta}$$

where $\psi^{(1)}$ is the trigamma function.

For the prior on the WT prevalence, π, we consider

$$\pi \sim Unif(0.5, 1.0)$$

The lower limit of 0.5 was chosen because all the species considered in this dissertation are majority WT. The prior, however, can be generalized to $\pi \sim Unif(0.0, 1.0)$.

3.3.2 Prior on σ^2_δ

As discussed in Chapter 1, the variance for the latent true WT distribution and the ME variance are confounded when the true WT distribution is Normal. The skewness of the log2gamma distribution and symmetry of the Normal measurement error enables a nonparametric deconvolution, but the information about the latent true distribution is rather weak [Delaigle et al., 2016]. One strategy for modelling ME is treating σ_δ as an additional unknown parameter. Thus, it requires an appropriate prior.

The method detailed in van de Kassteele et al. [2012] proposes non-informative priors on the two variance components, but identifiability issues will emerge between the two components. Quality control data, like that in Table 1.1, provide information on the between-lab and within-lab variabilities one might expect to see in a study. By leveraging QC data, an informative prior for σ^2_δ is justifiable.

For the purposes of a realistic simulation study, plausible values of σ_δ must be considered. The first example of QC data is *E. coli* ATCC 25922 and they produce an estimated value of $\hat{\sigma}^2_\delta = 0.19$. The second example of QC data is *S. aureus* ATCC 29213 and they produce an estimated value of $\hat{\sigma}^2_\delta = 0.32$. Thus, we consider the prior $\sigma^2_\delta \sim InvGamma(\alpha_\delta = 17, \beta_\delta = 4)$. This distribution has a mode at 0.25. A value of 0.16 corresponds roughly with the 30^{th} percentile and a value of 0.36 is the 86^{th} percentile. In contrast, the prior proposed in Craig [1999] was $\sigma_\delta \sim Gamma(\alpha = 6.007, \beta = 12.015)$. These two densities for σ_δ are displayed in Figure 3.1. The density for the prior from Craig [1999] has a mode slightly to the left of the proposed prior's mode and much heavier tails.

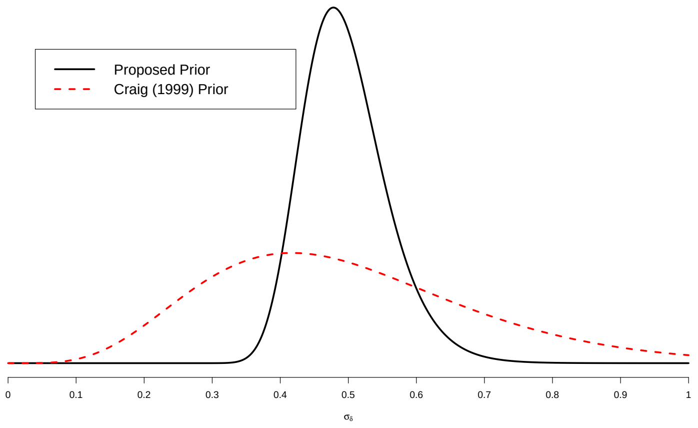


Figure 3.1. Proposed prior density and the previous prior from Craig [1999] used to describe the ME standard deviation σ_δ.

3.3.3 Prior for the NWT Parameters

The NWT distribution is often multimodal with an unknown number of components. We address this by using a mixture of an unknown number of Normals. Specifically, we use a Dirichlet Process Mixture Model (DPMM) for the NWT distribution [Escobar and West, 1995]. There are, however, modifications made for this particular application. We discuss these as they arise in our description of the DPMM.

A DPMM with a Normal kernel denoted as $N(\cdot \mid \boldsymbol{\theta_i})$ can be expressed as [Ross and Markwick, 2018]

$$X_i \sim F$$

$$F = \sum_{i=1}^{k} w_i N(X_i \mid \boldsymbol{\theta_i})$$

$$\boldsymbol{\theta_i} \sim G$$

$$G \sim DP(\alpha_{conc}, G_0)$$

$$\alpha_{conc} \sim Gamma(\alpha_0, \beta_0)$$

$$G_0 = N(\mu_B \mid \mu_{G_0}, \sigma_0) InvGamma(\sigma_B^2 \mid \alpha_1, \beta_1)$$

Here $\boldsymbol{\theta_i}$ represents the mean and variance of mixture component i. The collection of these k $\boldsymbol{\theta_i}$'s and their associated weights comprise the parameters in $\boldsymbol{\theta_{NWT}}$. To implement a DPMM, one must choose a suitable base distribution G_0 and either a set value or prior for the concentration parameter α_{conc}. For the base distribution, we utilize a semi-conjugate Normal-Inverse Gamma distribution. This allows for greater flexibility than the fully-conjugate Normal-Inverse Gamma distribution by allowing the variance of μ_B to be independent of the variance σ_B^2 [Görür and Rasmussen, 2010].

To ensure the NWT distribution is to the right of the WT distribution and to address the non-identifiability of components within a mixture model, conditions are built into the base distribution. First, we set a lower limit, referred to as A, for the means of these components. In doing so, this guarantees that the NWT distribution is to the right of the WT distribution. Second we provide some control on how large the variances of these components can be through the prior on σ_B^2. Thus our base distribution is

$$G_0 = TrN(\mu_B \mid A, \infty, \mu_{G_0}, \sigma_0) InvGamma(\sigma_B^2 \mid \alpha_1, \beta_1)$$

We build in these constraints because we assume that most of the apparent overlap between the WT and NWT distributions is due to ME and censoring. For this reason, we choose a value for A that is far in the WT right tail. Currently, we use the ECOFF of the true WT distribution. For the Normal case this means

$$A = \mu_{WT} + \Phi^{-1}(0.999)\sigma_{WT} \approx \mu_{WT} + 3.09\sigma_{WT}$$

and in the log2gamma case with gamma distribution cdf F,

$$A = \log_2(F^{-1}(0.999; \alpha, \beta))$$

We also restrict the variances of the Normal components through their prior to limit the degree of overlap. Specifically, we restrict the overlap to primarily occur in the left tail beyond one standard deviation from the mean. These restrictions are displayed graphically in Figure 3.2 using a single Normal distribution for the NWT distribution. Only 1% of the Normal component is within one standard deviation of the WT mean.

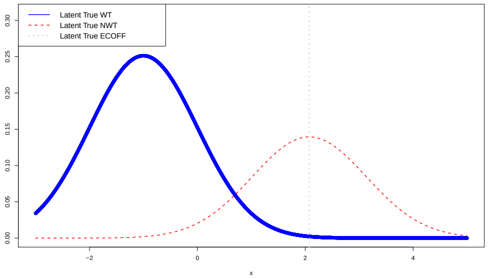


Figure 3.2. This figure demonstrates the worst case contamination occurring on a latent true level in the Normal case.

In the Normal case, we can determine this upper limit for the variance by finding σ such that

$$\frac{\mu_{WT} + \sigma_{WT} - A}{\sigma} = \Phi(0.01)$$

This turns out to be

$$\sigma = \left(\frac{\Phi^{-1}(0.999) - 1}{\Phi^{-1}(0.99)} \right) \sigma_{WT} \approx 0.9\sigma_{WT}$$

which we set as the 97.5th percentile of the prior.

The log2gamma case adheres to the same approach but considers the 84th percentile Q rather than one standard deviation from the mean. Thus, the upper limit for the variance is the solution to the equation

$$\frac{Q - A}{\sigma} = \Phi(0.01)$$

and the prior for variance can be set accordingly.

In both WT distribution cases, we also want to ensure these variances from being too small. Although this to some degree is controlled by the concentration parameter α_{conc}, we also set the 7.5th percentile of the variance prior to be roughly $\sigma_\delta/2$.

Both of these restrictions on G_0 require knowledge of the WT distribution and σ_δ, unknowns we are trying to estimate. As a result, we use the data to obtain estimates that are then held fixed in our algorithm. Our current estimation procedure is described in the next section.

A gamma prior for the concentration parameter α_{conc} is common. Balancing a low level of certainty in the base distribution with a desire to avoid having too many NWT mixture components (for computational efficiency), a gamma prior with mean 1.250 and standard deviation 0.559 is currently recommended.

3.4 Computation

The general computational strategy for updating the parameters is a Metropolis-within-Gibbs scheme. First, initial values are determined for $\boldsymbol{\theta_{WT}}$, $\mathbf{c}$, σ_δ^2, $\mathbf{X}$, and $\boldsymbol{\delta}$. We then apply our heuristic to set the base distribution of the DPMM. Then we run through the updates of our MCMC scheme.

3.4.1 Setting Initial Values

For the starting value of σ_δ, the prior mode is used. For initial values of π and $\boldsymbol{\theta_{WT}}$, we desire a method that is computationally efficient and robust to contamination. For this purpose, we consider our modification of the Turnidge et al. method that uses only the bins on the left up to the mode plus one as the subset for estimation. The initial values for the WT parameters are denoted as $\pi^{(0)}$ and $\boldsymbol{\theta^{(0)}_{WT}}$. For the variance in the Normal case, we take the modified Turnidge et al. estimate and subtract our initial estimate of σ_δ^2 from it. In the log2gamma case, measurement error is ignored in the initial values of α and β because any adjustment to α requires an updated estimate of β. It is possible to modify these estimates for ME, but the computationally affordable approach is favored for setting initial values that will be later discarded.

We then generate starting values for $\boldsymbol{\delta}$ from a $N(0, \sigma_\delta^{(0)})$. The $\mathbf{Y}$ are ordered from smallest to largest and the first $\lceil N_{tot}\pi^{(0)}\rceil$ values are designated as WT ($c_i = 1$) and the remainder are NWT ($c_i = 0$). The initial values of $\mathbf{X}$ for those designated as WT are generated from a truncated WT distribution using the corresponding values of $\boldsymbol{\delta}^{(0)}$ and $\mathbf{Y}$ with parameters $\boldsymbol{\theta^{(0)}_{WT}}$.

For the NWT distribution, we start with a single Normal component. The initial estimates for the NWT $\mathbf{X}$ are generated from a truncated Normal using the corresponding values of $\boldsymbol{\delta}$ and $\mathbf{Y}$. The parameters for this generation are the sample mean of the NWT identified $\mathbf{Y}$ minus 0.5 and their standard deviation. For the NWT distribution, there is no adjustment based on σ_δ because the number of components will likely increase.

3.4.2 Setting the Base Distribution

Because the information we need for these restrictions depends on the deconvolution of $\mathbf{X}^*$ into $\mathbf{X} + \boldsymbol{\delta}$, we estimate the restriction parameters using the observed distribution $\mathbf{X}^*$. While the modified Turnidge et al. approach provides adequate estimates of μ_{WT} and σ_{obs}, we need σ_{WT} to compute the limit A. Based on our prior for σ_δ and estimates of σ_{obs} found in the literature, the observed WT standard deviation σ_{obs} is typically 7.5% to 15.5% larger than σ_{WT}. Thus, we would want to multiply our estimate of σ_{obs} by a number between 2.61 and 2.86 to estimate A adequately. We decided on 2.75 so

$$\hat{A} = \hat{\mu}_{WT}^{(0)} + 2.75\hat{\sigma}_{obs}^{(0)}$$

Similarly, for the variance prior, we set the 97.5th percentile equal to 0.9 $\hat{\sigma}_{WT}^{(0)}$ and the 7.5th percentile roughly equal to $\sigma_\delta/2$. To eliminate the dependency on the starting values, we run a 5000 iteration burn-in with BayesACME. This enables a re-calculation of $\hat{A}$ and the prior for the variance components using

$$\hat{\hat{A}} = \hat{\mu}_{WT} + 2.75\sqrt{\hat{\sigma}_{WT}^2 + \hat{\sigma}_\delta^2}$$

We then use the updated estimates of $\boldsymbol{\theta_{WT}}$, σ_δ, and $\hat{\hat{A}}$ to determine updated values of α_1 and β_1.

An analogous process exists for the log2gamma distribution. Using the modified Turnidge et al. approach, we obtain initial estimates of α and β that ignore measurement error. Now we use estimates of the WT mode and standard deviation to calculate A. The WT mode plus 2.5 times the WT standard deviation roughly approximates the latent true ECOFF (i.e., the 99.9th percentile). We replace $\hat{A}$ above with

$$\hat{A} = \log_2(\frac{\hat{\alpha}^{(0)}}{\hat{\beta}^{(0)}}) + 2.5\sqrt{\frac{\Psi^{(1)}(\hat{\alpha}^{(0)})}{\ln(2)^2}}$$

Then run a burn-in to use estimates from BayesACME to reduce the potential impact from contamination. Then

$$\hat{\hat{A}} = \log_2(\frac{\hat{\alpha}}{\hat{\beta}}) + 2.5\sqrt{\frac{\Psi^{(1)}(\hat{\alpha})}{\ln(2)^2} + \hat{\sigma}_\delta^2}$$

and again we update our choice for the prior of the component variances.

3.4.3 Overview of MCMC Algorithm

Details of each step of our algorithm are presented in Appendix C. Here we provide a general outline. In the Normal case, μ_{WT} and σ^2_{WT} are each updated in separate Gibbs steps. In the log2gamma case, α is updated in a Metropolis step and β is updated in a Gibbs step. Beyond those updates the algorithms for the Normal and log2gamma case are very similar. The DPMM is updated using Algorithm 8 from Neal [2000] and α_{conc} is updated in a Gibbs step using West [1992].

Steps in the Algorithm

1. Set initial values for parameters

2. Determine the base distribution of DPMM

3. Update the Model Parameters for a Set Number of Iterations

    - Update $\boldsymbol{\theta_{WT}} \mid \mathbf{X}, \mathbf{c}$ (this varies for choice of WT distribution)

        - Normal

            - Update $\mu_{WT} \mid \sigma^2_{WT}, \mathbf{X}, \mathbf{c}$

            - Update $\sigma^2_{WT} \mid \mu_{WT}, \mathbf{X}, \mathbf{c}$

        - log2gamma

            - Update $\alpha \mid \beta, \mathbf{X}, \mathbf{c}$

            - Update $\beta \mid \alpha, \mathbf{X}, \mathbf{c}$

- Update $\pi \mid \mathbf{c}$
- Update $\sigma_\delta^2 \mid \boldsymbol{\delta}$
- Update $\boldsymbol{\theta_{NWT}} \mid \mathbf{X}, \mathbf{c}$
- Update $\mathbf{X} \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\delta}$
- Update $\boldsymbol{\delta} \mid \mathbf{X}, \mathbf{c}, \mathbf{Y}, \sigma_\delta^2$
- Update $\mathbf{c} \mid \mathbf{X}, \pi, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}$

4. Discard burn-in

3.5 Choosing Between WT Models

In the Bayesian paradigm, a popular model selection approach is using a Bayes Factor [Jeffreys, 1998, Kass and Raftery, 1995]. Our use of Jeffreys priors renders the Bayes Factor directly inapplicable. It is possible that Intrinsic or Fractional Bayes Factors could be applied [Berger and Pericchi, 1996a,b, O'Hagan, 1995]. The major drawback to their use are the non-straightforward computations. There are MCMC approximations to Bayes Factors available [Carlin and Chib, 1995, Newton et al., 1999], but they may not work well in practice.

The AST data in their crudest form are modelled by a multinomial distribution. One straightforward approach to assessing fit is a goodness-of-fit test to this multinomial distribution. We propose using the MCMC results for each model to assess how well they fit the observed counts. For each iteration of the Normal and log2gamma MCMC chains, we calculate the goodness-of-fit of the fitted multinomial model

$$\sum_{j=1}^{J} \left[\frac{m_j - N_{tot} p_j(\tilde{\boldsymbol{\theta}})}{\sqrt{N_{tot} p_j(\tilde{\boldsymbol{\theta}})}} \right]^2$$

where $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}, \pi, \sigma_\delta)$. This gives us approximations to the posterior distributions of goodness-of-fit. Johnson et al. [2004] proved under mild conditions that the asymptotic posterior distribution has a $\chi^2(J-1)$ distribution so each model could be assessed for fit. In terms of selecting the model, we choose the distribution that has the smaller posterior median.

The chief benefit of this goodness-of-fit approach is that it enables clever use of the MCMC chain and avoids additional computation. One drawback of this approach is it may be sensitive to low bin counts; a problem especially for small data sets. In that case, combining bins may be required. Typically, this means combining bins on the right together as the counts are typically lower there and places a greater focus on the WT distribution. In the concluding chapter, an alternative model selection approach is discussed.

3.6 Single-Year Simulation Study

This section summarizes the relative performance of BayesACME and the methods described in Chapter 2, specifically Turnidge et al. [2006] and Jaspers et al. [2014a]. We chose these two methods because of their popularity and flexibility to handle both the Normal and log2gamma WT distributions. In addition to these approaches, we consider our modification of the Turnidge et al. method. The other Bayesian approach [Jaspers et al., 2016a] is also considered but only for a subset of the Normal WT distribution scenarios.

We investigate both the Normal and log2gamma WT distributions using a full factorial design involving the following three levels for each of three factors:

- Sample size: $N_{tot} = 300, 600, 1200$
- Measurement error: $\sigma_\delta = 0.4, 0.5, 0.6$
- Levels of contamination: Low, Medium, and High

The sample sizes are much smaller than those considered in the simulation study of Jaspers et al. [2014a]. These sample sizes were based on the sizes of clinical results from a single lab provided by Professor John Turnidge. The levels of measurement error were chosen to represent the range we observe in QC data.

To quantify the amount of overlap between the WT and NWT distributions, we use the the Bhattacharyya coefficient [Bhattacharyya, 1943]. The Bhattacharyya coefficient for distributions p and q is defined for both continuous and discrete distributions. Thus, we can quantify our contamination at the true MIC distribution scale, at the observed continuous MIC distribution scale, or at the observed MIC scale.

In the continuous case,

$$BC(p,q) = \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}dx$$

In the discrete case,

$$BC(p,q) = \sum_{x} \sqrt{p(x)q(x)}$$

Note that $0 \leq BC(p,q) \leq 1$ where a value of 1 implies complete overlap and a value of 0 implies no overlap. The exact values of the Bhattacharyya coefficients that make up our different contamination levels are discussed in the following subsection.

The Bhattacharyya coefficient is intended to describe the overlap between two distributions. We are using this for the two components of a mixture model that introduces a mixture weight π. A limitation of Bhattacharyya coefficient is it does not account for π and its practical importance in the overlap between the components. For example, for the same Bhattacharyya coefficient it is much harder to estimate the WT distribution when $\pi = 0.65$ compared to $\pi = 0.85$.

3.6.1 Study Settings

For each of the $3^3 = 27$ combinations, we simulated 1000 data sets and analyzed each data set using the various methods. For the Normal distribution case, the latent WT distribution parameters are $\mu_{WT} = -1.0$ and $\sigma_{WT} = 1.0$. The parameters for the log2gamma distribution are $\alpha = 2.9686$ and $\beta = 4.0526$ (rate). This distribution was based on approximating a Normal distribution with selection probability $\Phi(\frac{(x-(-2))}{.6})$.

For the NWT distribution, we consider a mixture of two Normals, with components that change so that the contamination in the right tail of the WT distribution varies. For the NWT distribution, the weight vector, mean vector, and standard deviation vector are denoted $\mathbf{w_{NWT}}$, $\boldsymbol{\mu_{NWT}}$, and $\boldsymbol{\sigma_{NWT}}$, respectively.

- Low Contamination:

$$\mathbf{w_{NWT}} = (\tfrac{5}{37}, \tfrac{32}{37}) \qquad \boldsymbol{\mu_{NWT}} = (4.60, 5.10) \qquad \boldsymbol{\sigma_{NWT}} = (0.95, 1.10)$$

- Medium Contamination:

$$\mathbf{w_{NWT}} = (0.60, 0.40) \qquad \boldsymbol{\mu_{NWT}} = (2.60, 5.00) \qquad \boldsymbol{\sigma_{NWT}} = (0.80, 1.10)$$

- High Contamination:

$$\mathbf{w_{NWT}} = (0.80, 0.20) \qquad \boldsymbol{\mu_{NWT}} = (2.20, 5.00) \qquad \boldsymbol{\sigma_{NWT}} = (0.70, 1.00)$$

Figure 3.3 is a stacked histogram showing the contributions of the latent true WT (in blue) and latent true NWT (in red) distributions in each bin. The left column is the Normal case and the right column is the log2gamma case. Each successive row

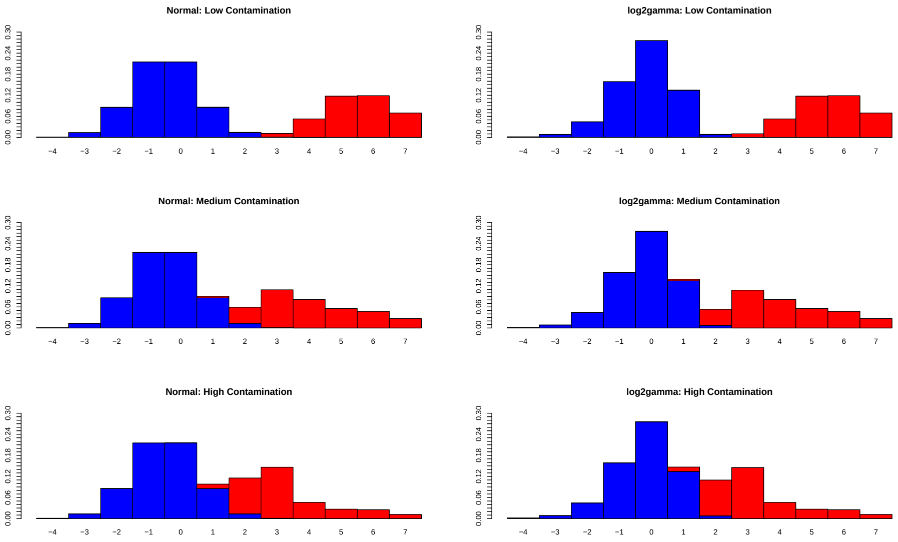


Figure 3.3. Stacked histogram for contamination by bin where each row corresponds to a level of contamination.

is an increase in contamination. While the amount of overlap is relatively small, it is increasing over the levels.

Table 3.1 summarizes the Bhattacharrya coefficients for the different contamination levels and values of σ_δ in the Normal setting. The inclusion of measurement error (ME) serves as an "amplifier" of contamination because it increases the observed WT distribution variance. Table 3.2 reports the Bhattacharyya coefficients for the observed MIC distribution. The censoring also increases the degree of overlap.

Table 3.1.
Contamination for the Continuous Normal Case

Contamination	**No ME**	$\boldsymbol{\sigma_\delta = 0.4}$	$\boldsymbol{\sigma_\delta = 0.5}$	$\boldsymbol{\sigma_\delta = 0.6}$
Low	0.0150	0.0259	0.0332	0.0430
Medium	0.1070	0.1489	0.1717	0.1987
High	0.1554	0.2137	0.2440	0.2785

Table 3.2.
Contamination for the Censored Normal Case

Contamination	**No ME**	$\boldsymbol{\sigma_\delta = 0.4}$	$\boldsymbol{\sigma_\delta = 0.5}$	$\boldsymbol{\sigma_\delta = 0.6}$
Low	0.0201	0.0323	0.0403	0.0509
Medium	0.1282	0.1695	0.1918	0.2179
High	0.1855	0.2411	0.2699	0.3025

Table 3.3 is similar to Table 3.1, but for the log2gamma distribution. The Bhattacharyya coefficients are generally lower because the right tail of the log2gamma decays at a faster rate than the Normal. Similar to the Normal case, the censoring increases the degree of overlap (Table 3.4).

For the simulation studies only A is updated, and the prior for the NWT variances is fixed. In future work, both A and the prior for the NWT variances will be automatically updated. In the Normal case, $\sigma_B^2 \sim InvGamma(\alpha_1 = 6, \beta_1 = 2)$. In the log2gamma case, $\sigma_B^2 \sim InvGamma(\alpha_1 = 9, \beta_1 = 2.5)$. In the Normal case, 0.9 corresponds with the 97.4$^{\text{th}}$ percentile and the 7.5$^{\text{th}}$ percentile is around 0.204. For the log2gamma case, the true values of α and β correspond roughly to the 95.9$^{\text{th}}$ per-

Table 3.3.
Contamination for the Continuous log2gamma Case

Contamination	**No ME**	$\boldsymbol{\sigma_\delta = 0.4}$	$\boldsymbol{\sigma_\delta = 0.5}$	$\boldsymbol{\sigma_\delta = 0.6}$
Low	0.0071	0.0178	0.0257	0.0368
Medium	0.0788	0.1373	0.1677	0.2026
High	0.1204	0.2051	0.2463	0.2913

Table 3.4.
Contamination for the Censored log2gamma Case

Contamination	**No ME**	$\boldsymbol{\sigma_\delta = 0.4}$	$\boldsymbol{\sigma_\delta = 0.5}$	$\boldsymbol{\sigma_\delta = 0.6}$
Low	0.0113	0.0238	0.0324	0.0443
Medium	0.1070	0.1627	0.1915	0.2241
High	0.1629	0.2406	0.2782	0.3193

centile and 0.184 is near the 7.6th percentile. This selection allows for a larger range of variances than in the Normal case, but helps the variances of the NWT components to be generally larger than ME variance. For the analysis of the real data set, they are re-set each time the estimate of A is updated.

3.7 Single-Year Results

The quantities of interest for microbiologists are the prevalence π and the WT distribution parameters $\boldsymbol{\theta_{WT}}$. Because the WT distribution is primarily used to calculate the ECOFF, we focus the simulation results on the accuracy and precision of the estimates for prevalence and the observed ECOFF. For the Bayesian methods we use the posterior mean as the estimate. In the interest of brevity, only a small number of results are presented. More results are available in Appendix D.

As the different methods are compared frequently the following names are used as reference. The methods are also color-coded in the figures.

TURN (green): The method of Turnidge et al. [2006]

TURNM (orange): A modification to TURN where the subset selection is always the bins on the left up to the mode plus one

JASP (brown): The method of Jaspers et al. [2014a], but with the modification of using at least four bins or more to ensure the WT mode is included in the subset

JASPB (pink): The method of Jaspers et al. [2016a] where the priors for the WT parameters are determined using estimates from JASP

BayesACME (blue): Our proposed method

The results are presented using side-by-side modified boxplots and numerical summaries. In some cases, the results are highly skewed. For that reason, we use numerical summaries that are resistant to outliers. They are

- Median
- Median Absolute Deviation (MAD)
- Median Absolute Deviation from the True Value (MADT)
- Interquartile Range (IQR)

The median is an outlier resistant measure of central tendency. We can compare it to the true value to obtain a measure of bias. MAD, MADT, and IQR are resistant measures of precision. MADT is akin to the mean squared error (MSE), which is a popular quantity to describe the quality of an estimator. In the numerical summaries, the best two results are in boldface for each main factor level.

3.7.1 Normal Case

Figure 3.4 summarizes the observed ECOFF results for $N_{tot} = 600$ and $\sigma_\delta = 0.5$. The three panels represent the three levels of contamination. At low contamination

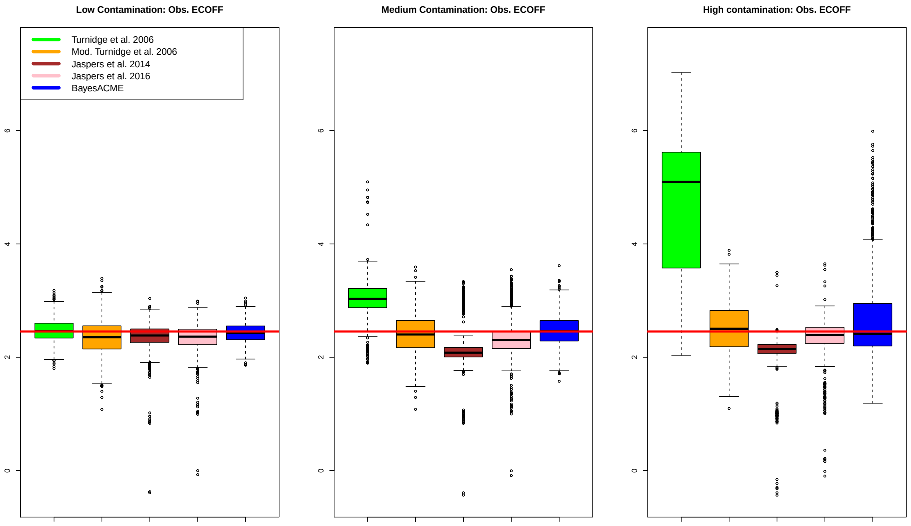


Figure 3.4. This set of boxplots contrasts the results of the ECOFF values of the discussed methods at size 600 and $\sigma_\delta = 0.5$.

the different methods are very comparable. This is anticipated as all methods are basically using the same data to fit the Normal WT distribution. The two subset methods produce nearly unbiased estimates, but TURN is a bit more precise. This is due to JASP at times selecting too few bins and thus under-estimating the variance and mean. The precision of BayesACME is slightly better than TURN, most likely due to the use of a little more data on the right side of the Normal distribution. Consequently, BayesACME has a lower MADT (Table 3.5). Because the TURNM approach does not use as much data as TURN in this setting, the precision is poorer. JASPB incorporates the estimates from JASP into their priors for $\boldsymbol{\theta_{WT}}$ and π so the results roughly mimic that method at this sample size.

In the second and third panels of Figure 3.4, we see that TURN is very sensitive to contamination. This is due to the choice of subset selection heuristic. Notice that TURNM handles contamination much better. TURNM's considerably smaller values of MADT in Table 3.5 than TURN at medium and high contamination reflect this improvement. In using the AIC to select the number of bins, JASP tends to

Table 3.5.
Numerical Summaries for Figure 3.4

Method	**Cont.**	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	Low	**0.0074**	0.1309	0.1323	0.2614
TURNM	Low	−0.1020	0.2031	0.2187	0.4079
JASP	Low	−0.0704	**0.1173**	**0.1166**	**0.2360**
JASPB	Low	−0.0904	0.1349	0.1509	0.2728
BayesACME	Low	**−0.0291**	**0.1204**	**0.1248**	**0.2423**
TURN	Medium	0.5778	0.1685	0.5778	0.3389
TURNM	Medium	**−0.0527**	0.2400	0.2477	0.4488
JASP	Medium	**−0.3734**	**0.0824**	0.3988	**0.1652**
JASPB	Medium	−0.1480	**0.1487**	**0.2094**	**0.2964**
BayesACME	Medium	0.0072	0.1803	**0.1801**	0.3600
TURN	High	2.6422	1.1503	2.6422	2.0438
TURNM	High	**0.0495**	0.3200	0.3215	0.6395
JASP	High	−0.3059	**0.0784**	**0.3067**	**0.1578**
JASPB	High	−0.0590	**0.1436**	**0.1449**	**0.2831**
BayesACME	High	**−0.0395**	0.2890	0.3113	0.7488

have “packs” of estimates that are associated with a particular bin choice. JASPB produces very similar results to JASP, but it does a better job handling contamination by modelling the NWT distribution. BayesACME is quite good at avoiding bias, but a long right tail emerges when there is high contamination. This tail, however, does dissipate with increased sample size. At high contamination and low sample size, the estimate of A is often estimated to be too large and thus the estimated WT distribution absorbed some of the NWT distribution. The informative priors used in JASPB produce smaller variability in the estimates of σ_{obs}. This avoids the long right tail of BayesACME, but it still struggles sometimes on the left with poor prior distributions.

While TURN’s relative performance to BayesACME worsens with increases in contamination and TURNM’s relative performance improves, JASP and JASPB follow a different pattern. BayesACME has a very strong relative performance over the two methods at medium contamination, but high contamination JASPB dominates. JASPB restricts the WT distribution with “relatively” informative priors and BayesACME is the opposite restricting the NWT distribution. Based on these results, BayesACME needs restrictions on both the NWT and WT distributions to be more competitive with JASPB at high contamination for the sample sizes considered.

In terms of prevalence (Figure 3.5), the results from the different methods are more comparable. BayesACME tends to be the least biased between data sets as the level of contamination increases. TURN struggles the most with contamination. However, TURNM is the least impacted by high contamination in terms of bias (Table 3.6). At high contamination, note the long tails of JASPB and BayesACME. Recall similar tails developed for estimates of the observed ECOFF in Figure 3.4. The side of the tail indicates limitations in the way these two approaches handle contamination. The left tail of JASP indicates it is, at times, taking on too few bins. At this sample size, JASPB mimics JASP and has a long left tail. The long right tail of BayesACME indicates part of the NWT distribution is spuriously treated as the WT distribution. In terms of within-data set comparison, the relative accuracy of both JASP and

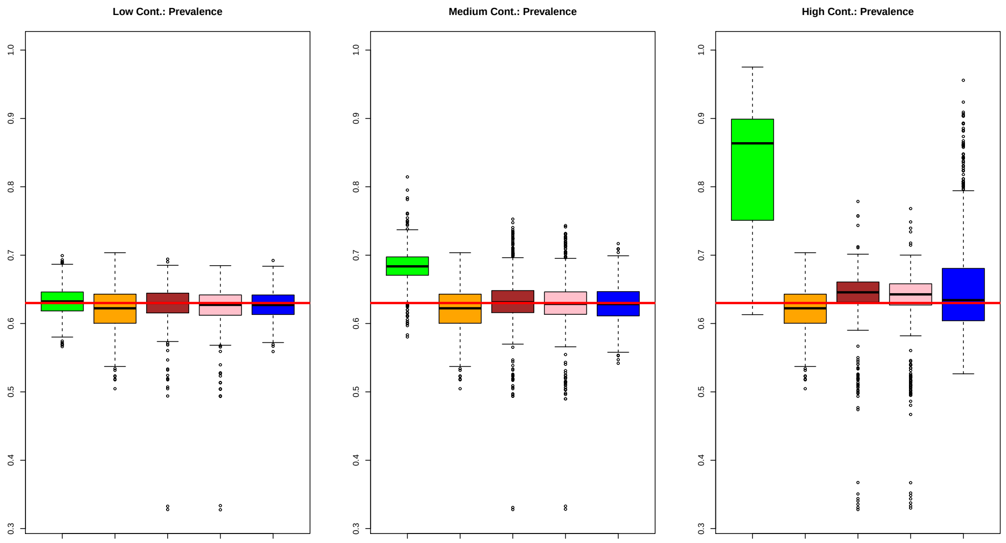


Figure 3.5. This set of boxplots contrasts the results of the prevalence values of the discussed methods at size 600 and $\sigma_\delta = 0.5$.

Table 3.6.
Numerical Summaries for Figure 3.5

Method	**Cont.**	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	Low	**0.0024**	**0.0138**	0.1400	**0.0277**
TURNM	Low	−0.0077	0.0215	0.0214	0.0425
JASP	Low	**0.0000**	0.0145	**0.0145**	0.0289
JASPB	Low	−0.0027	0.0148	0.0149	0.0297
BayesACME	Low	−0.0029	**0.0141**	**0.0138**	**0.0285**
TURN	Medium	0.0536	**0.0135**	0.0536	**0.0268**
TURNM	Medium	0.0018	0.0250	0.0253	0.0504
JASP	Medium	**0.0010**	**0.0160**	**0.0157**	**0.0323**
JASPB	Medium	−**0.0012**	0.0161	**0.0164**	0.0328
BayesACME	Medium	−0.0016	0.0177	0.0179	0.0356
TURN	High	0.2336	0.0808	0.2336	0.1479
TURNM	High	0.0175	0.0323	0.0359	0.0661
JASP	High	0.0156	**0.0146**	**0.0208**	**0.0294**
JASPB	High	**0.0127**	**0.0156**	**0.0192**	**0.0310**
BayesACME	High	**0.0042**	0.0355	0.0341	0.0764

JASPB over BayesACME improves with increases in contamination. This was not the pattern with the corresponding observed ECOFF estimates. In future work, we will investigate how to best determine and to use informative restrictions on the WT and NWT distributions. Both the general pattern of restricting π and it is likely a quantity where external information is available makes it a starting place for future inquiry in adopting additional restrictions.

Figure 3.6 summarizes the results of increased sample size on the observed ECOFF. This boxplot is for medium contamination and $\sigma_\delta = 0.5$. TURNM is good at producing nearly unbiased estimates, and has the best MADT among the subset methods. At sample sizes 600 and 1200, the mixture models have smaller MADT values. Although BayesACME only beats TURMN in terms of the bias between-data sets at size 600, Table 3.7 shows that BayesACME has better precision in terms of lower MAD, MADT, and IQR at all three sample sizes over TURNM. The other subset methods struggle.

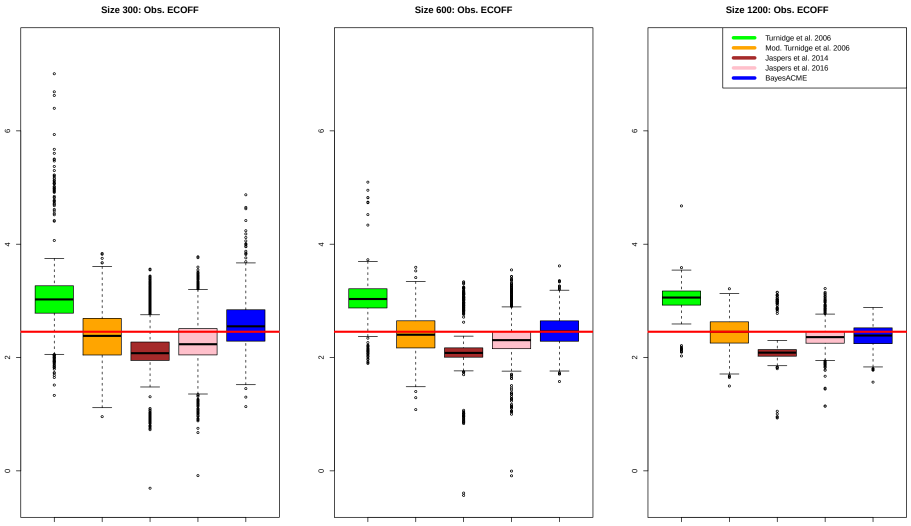


Figure 3.6. The boxplots show the results of the obs. ECOFF estimates of the methods at medium contamination and $\sigma_\delta = 0.5$.

TURN takes on too many bins and JASP takes only too few bins with the hopes of avoiding contamination. The results are then biased. With increases in sample size, JASPB is better able to overcome the priors centered around the biased estimates of JASP. Here BayesACME is able to best this method in accuracy and precision between data sets. For within-data set comparisons, JASPB's relative performance regarding accuracy with BayesACME improves as sample size increases. It is unique among the compared methods suggesting both the benefit of mixture models over subset methods and superiority over BayesACME in terms increases in sample sizes. Although TURNM is the least biased method here, Table 3.7 allows for direct comparison using MADT. The values for MADT from TURNM typically require almost twice as much data in terms of MADT at the examined sample sizes to be comparable to BayesACME. This certainly highlights an advantage of using the full data set with mixture modelling.

Table 3.7.
Numerical Summaries for Figure 3.6

Method	**Sample Size**	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	300	0.5704	0.2416	0.5820	0.4838
TURNM	300	−**0.0730**	0.3238	**0.3272**	0.6435
JASP	300	−0.3783	**0.1438**	0.4472	**0.3239**
JASPB	300	−0.2202	**0.2157**	0.3470	**0.4629**
BayesACME	300	**0.0958**	0.2762	**0.2625**	0.5544
TURN	600	0.5778	0.1685	0.5778	0.3389
TURNM	600	−**0.0527**	0.2400	0.2477	0.4788
JASP	600	−0.3734	**0.0824**	0.3988	**0.1652**
JASPB	600	−0.1480	**0.1487**	**0.2094**	**0.2964**
BayesACME	600	**0.0073**	0.1803	**0.1801**	0.3600
TURN	1200	0.6034	0.1230	0.6034	0.2486
TURNM	1200	−**0.0010**	0.1879	0.1874	0.3747
JASP	1200	−0.3688	**0.0595**	0.3772	**0.1169**
JASPB	1200	−0.0949	**0.1045**	**0.1266**	**0.2074**
BayesACME	1200	−**0.0669**	0.1379	**0.1425**	0.2777

Figure 3.7 summarizes the results of increased sample size on the prevalence at medium contamination and $\sigma_\delta = 0.5$. The increases in sample size reduce the potential outliers for TURN and TURNM. TURNM is far better at avoiding bias from contamination than TURN. Both JASP and JASPB still struggle to some degree with under-estimation in the long left tails even at $N_{tot} = 1200$, but the majority of estimates from both methods are approaching the truth.

Interestingly the bias of both BayesACME and JASPB follow the same pattern as N_{tot} increases. It suggests that as the sample size grows, the NWT left tail becomes over-estimated. JASP is the most successful method at avoiding contamination for estimating π. The majority of the time, it uses a subset to the left of 1.0, which happens to be two standard deviations to the left of the closest component in the NWT distribution, for WT estimation. This demonstrates that subset methods such as JASP may poorly estimate the observed ECOFF, but may produce good estimates of π.

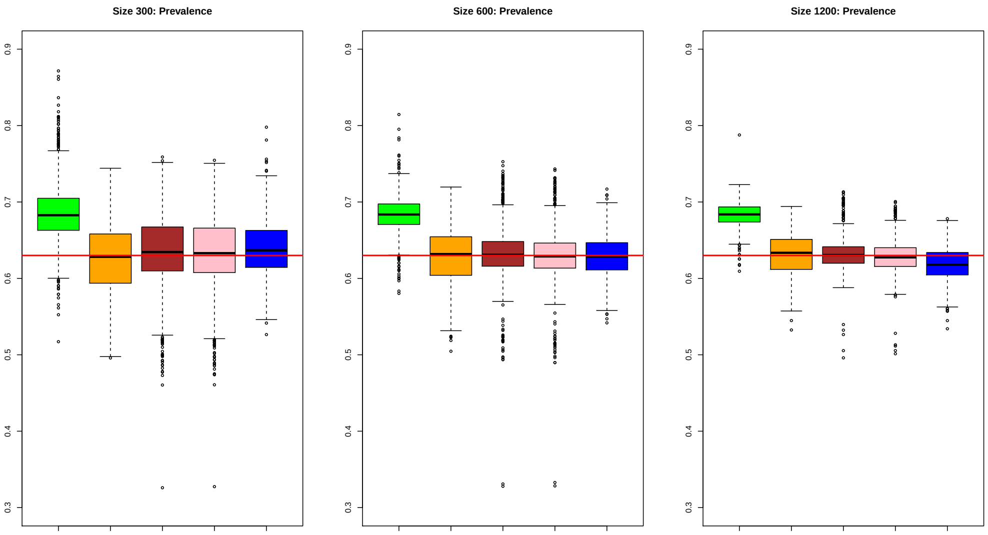


Figure 3.7. These boxplots display the results of the prevalence values of the discussed methods at medium contamination and $\sigma_\delta = 0.5$.

Table 3.8.
Numerical Summaries for Figure 3.7

Method	$\boldsymbol{N_{tot}}$	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	300	0.0527	**0.0212**	0.0531	**0.0419**
TURNM	300	**−0.0018**	0.0321	0.0315	0.0645
JASP	300	0.0045	0.0282	**0.0277**	0.0574
JASPB	300	**0.0030**	0.0284	0.0278	0.0581
BayesACME	300	0.0068	**0.0241**	**0.0246**	**0.0482**
TURN	600	0.0536	**0.0135**	0.0536	**0.0268**
TURNM	600	0.0018	0.0250	0.0253	0.0504
JASP	600	**0.0010**	**0.0160**	**0.0157**	**0.0323**
JASPB	600	**−0.0012**	0.0161	**0.0164**	0.0328
BayesACME	600	−0.0016	0.0177	0.0179	0.0356
TURN	1200	0.0537	**0.0099**	0.0537	**0.0197**
TURNM	1200	0.0035	0.0190	0.0201	0.0392
JASP	1200	**0.0006**	**0.0108**	**0.0109**	**0.0216**
JASPB	1200	**−0.0025**	0.0122	**0.0125**	0.0245
BayesACME	1200	−0.0120	0.0145	0.0168	0.0293

Even in the Normal case, BayesACME demonstrates sensitivity to the prior for σ_δ^2. Figure 3.8 shows the biases for the latent true WT variance, the ME variance, the observed standard deviation, and the observed ECOFF at low contamination for sample size 600. The left panel shows anticipated bias in the estimate of σ_δ^2 decreases as σ_δ^2 increases. The apparent bias at $\sigma_\delta = 0.5$ is from the skewness of the posterior distribution. The posterior mode is nearly unbiased at $\sigma_\delta = 0.5$. Consequently, the next panel on the right shows that the bias increases as σ_δ increases. These results are part of the weakly identifiable nature of the variance components [Gelman, 2014]. The third panel shows the observed standard deviation is under-estimated at three different values of σ_δ. The three side-by-side boxplots should have nearly identical distributions from the vagueness of the Jeffreys prior on σ_{WT}^2. There is a very slight increase in absolute bias from the low contamination becoming "amplified" from the increases in σ_δ. Likely, the culprit for this bias is the location of the mean of the closest NWT component. It contributes to the overlap of the observed distributions far into the WT right tail and far into the NWT left tail. This results in the observed WT distribution slightly "leaching" into the observed NWT distribution.

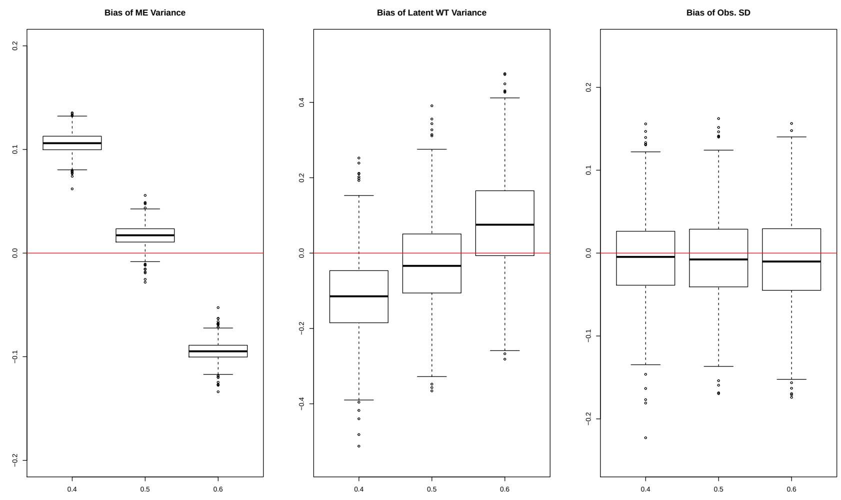


Figure 3.8. The boxplots show the bias of the ME variance, latent WT variance, and the obs. WT standard deviation, each when $\sigma_\delta = 0.4, 0.5, 0.6$, respectively at low contamination and size 600.

3.7.2 Log2gamma Case

As discussed in Chapter 1, the other methods consider the observed WT distribution to be log2gamma rather than a convolution with a log2gamma distribution with a Normal measurement error distribution. Consequently the tails of the observed WT distribution are prone to under-estimation. As a result, the other methods usually provide biased estimates of the observed ECOFF.

Figure 3.9 summarizes the results for medium contamination and $\sigma_\delta = 0.5$. The three panels represent the three sample sizes. BayesACME is unique among the methods in that it accounts for the Normal measurement error. Clearly BayesACME is the only procedure that accurately estimates the observed ECOFF. According to Table 3.9 the absolute bias is smallest and is the most precise as it has the lowest MADT. The other methods all demonstrate considerable bias. In fact, increasing sample sizes makes the discrepancy even more pronounced as the variability of the estimates decreases.

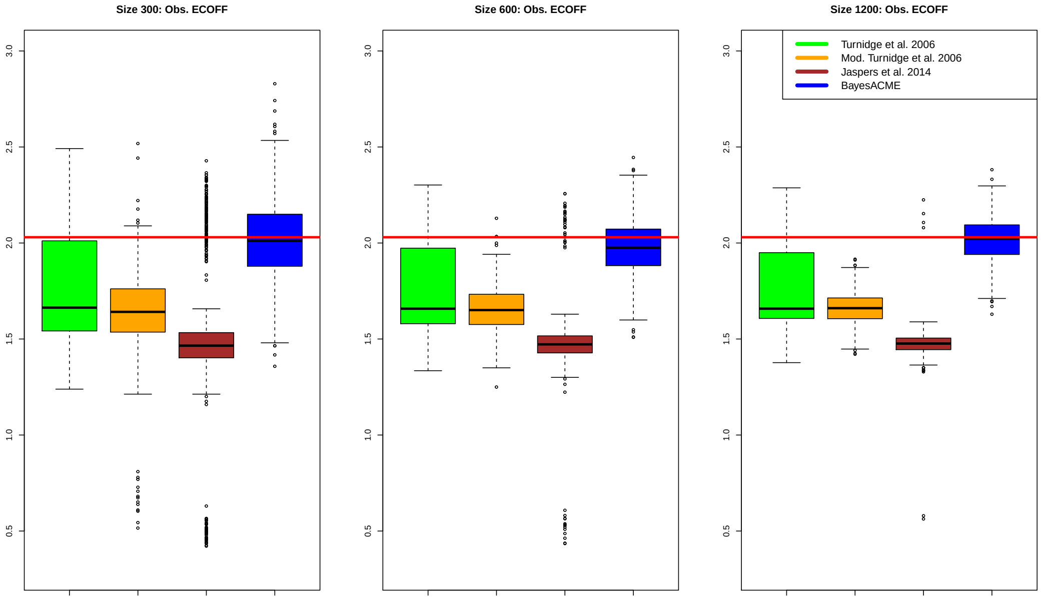


Figure 3.9. The results of the ECOFF values of the compared methods at medium contamination and $\sigma_\delta = 0.5$.

By looking at the results from TURNM, the bias in the TURN estimates is deflated because acquiring additional contaminated bins makes it appear that it compensates for the misspecification. Similarly, for JASP cases where a relatively large number of bins was taken resulted in the "packs" that are close to the truth. In terms of MADT, JASP struggles the most. Like both TURN and TURMN, it has misspecified the observed WT distribution. At low sample sizes, JASP can inadvertently compensate for its misspecification by using more bins, but the probability of this happens tends to 0 as the sample size increases. As the log2gamma distribution is asymmetric, the subset methods struggle to estimate the right tail of the WT distribution. The greater absolute bias from JASP indicates that a majority of the time, it assumes the right tail is decaying at a faster rate than TURN.

Table 3.9.
Numerical Summaries for Figure 3.9

Method	$\boldsymbol{N_{tot}}$	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	300	**−0.3668**	0.1954	**0.3685**	0.4689
TURNM	300	−0.3886	**0.1117**	0.3893	**0.2251**
JASP	300	−0.5646	**0.0667**	0.5646	**0.1311**
BayesACME	300	**−0.0158**	0.1335	**0.1353**	0.2684
TURN	600	**−0.3721**	0.1309	**0.3721**	0.3925
TURNM	600	−0.3794	**0.0782**	0.3794	**0.1571**
JASP	600	−0.5579	**0.0445**	0.5579	**0.0886**
BayesACME	600	**−0.0145**	0.0985	**0.0974**	0.1965
TURN	1200	−0.3720	0.0705	0.3720	0.3423
TURNM	1200	**−0.3694**	**0.0540**	**0.3694**	**0.1079**
JASP	1200	−0.5537	**0.0302**	0.5370	**0.0603**
BayesACME	1200	**−0.0050**	0.0736	**0.0730**	0.1499

Although the other methods misspecify the observed WT distribution, they are capable of producing adequate estimates of prevalence. Figure 3.10 summarizes the different prevalence estimates at different sample sizes at medium contamination where $\sigma_\delta = 0.5$. BayesACME develops some slight bias with increases in sample size. This bias is due to the sensitivity in setting $\hat{\hat{A}}$. As sample sizes increases, the mean value

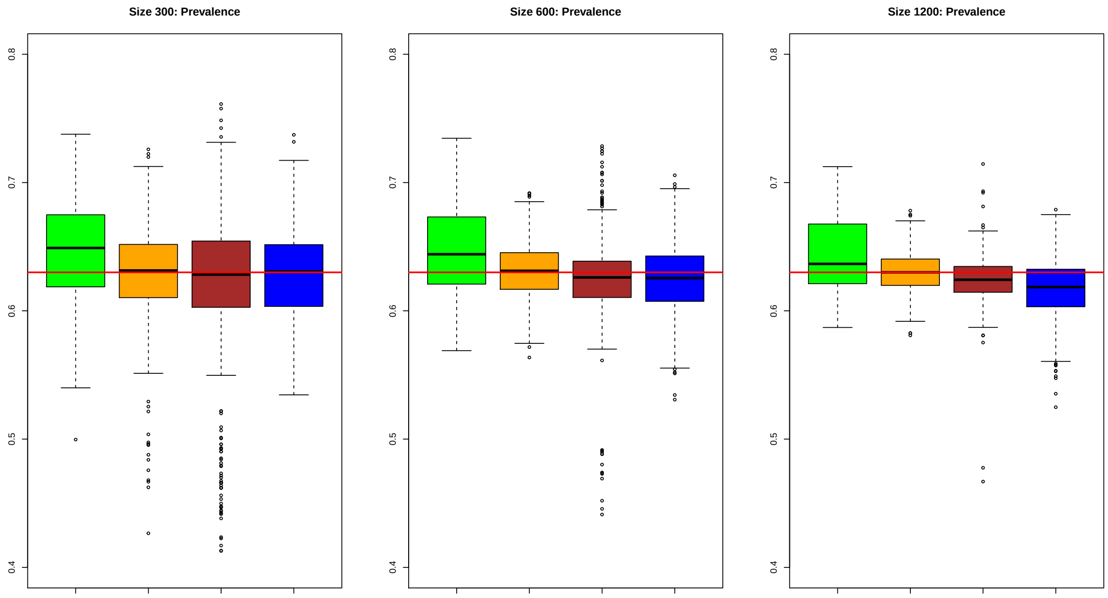


Figure 3.10. Boxplots of results for prevalence of the compared methods at medium contamination and $\sigma_\delta = 0.5$ as sample size varies. Notice that the methods that misspecified the observed WT distribution are able to produce adequate estimates of prevalence.

of $\hat{\hat{A}}$ decreases. This occurrence does not happen at low contamination and is more pronounced at high contamination indicating this is a sensitivity to contamination. The variance based on the fixed prior is likely becoming too big. A smaller value of $\hat{\hat{A}}$ enables a NWT component to absorb more of the WT right tail. As the right tail for a log2gamma distribution decays at relatively faster rate after the mode, this "creeping" effect of the NWT distribution is less of an issue for estimating the observed ECOFF. It may lead to deflated prevalence estimates. Except for the moderate bias in TURN, both TURNM and JASP are adequate at estimating the prevalence despite misspecifying the observed WT distribution. Relative to JASP, TURNM more rapidly loses its long left tail.

In Table 3.10, the bias of TURNM is the most improved by increases in sample size. For within-data set comparisons, TURNM is the only the method to show improvement in accuracy over BayesACME, albeit very slight. Comparative gains in efficiency are experienced by JASP except it is slightly more biased. BayesACME has much better relative performance at low contamination.

Table 3.10.
Numerical Summaries for Figure 3.10

Method	N_{tot}	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	300	0.0190	0.2800	0.0311	0.0561
TURNM	300	**0.0015**	**0.0209**	**0.0208**	**0.0414**
JASP	300	−0.0017	0.0285	0.0262	0.0516
BayesACME	300	**0.0007**	**0.0242**	**0.0246**	**0.0481**
TURN	600	0.0141	0.0258	0.0249	0.0523
TURNM	600	**0.0012**	**0.0143**	**0.0141**	**0.0285**
JASP	600	**0.0040**	**0.0142**	**0.0150**	**0.0283**
BayesACME	600	−0.0045	0.0176	0.0180	0.0353
TURN	1200	0.0066	0.0197	0.0171	0.0464
TURNM	1200	**0.0000**	**0.0103**	**0.0103**	**0.0205**
JASP	1200	**−0.0058**	**0.0101**	**0.0105**	**0.0201**
BayesACME	1200	−0.0114	0.0145	0.0156	0.0292

We can examine the influence of contamination in the log2gamma case. Below in Figure 3.11 at size 1200 and $\sigma_\delta = 0.5$, the results at low, medium, and high contamination from left to right for the observed ECOFF, respectively. Similar to the Normal case, TURN struggles with contamination. By avoiding the bins of the right WT tail, TURNM is robust to contamination. JASP struggles to determine the “correct” number of bins to use in the jump from low to medium contamination. BayesACME produces the least biased of the estimates by properly accounting for ME and handles contamination well by modelling the entire distribution. The precision decreases as contamination increases, but that is anticipated for all methods. There does appear to be slight bias in the observed ECOFF for BayesACME at low and high contamination. At high contamination, BayesACME is anticipated to over-estimate. The slight bias at low contamination does not manifest at lower sample sizes where low contamination is associated with the most accurate estimates of the observed

ECOFF at $\sigma_\delta = 0.5$ as anticipated. Nor does this occur at sample size 1200 and $\sigma_\delta = 0.4$. Likely, there is a NWT mode at this setting that does not manifest at other settings. The prior for σ_δ^2 also plays a role in the magnitude of the bias, where the absolute bias reflects the sensitivity to the prior. This sensitivity is explored later.

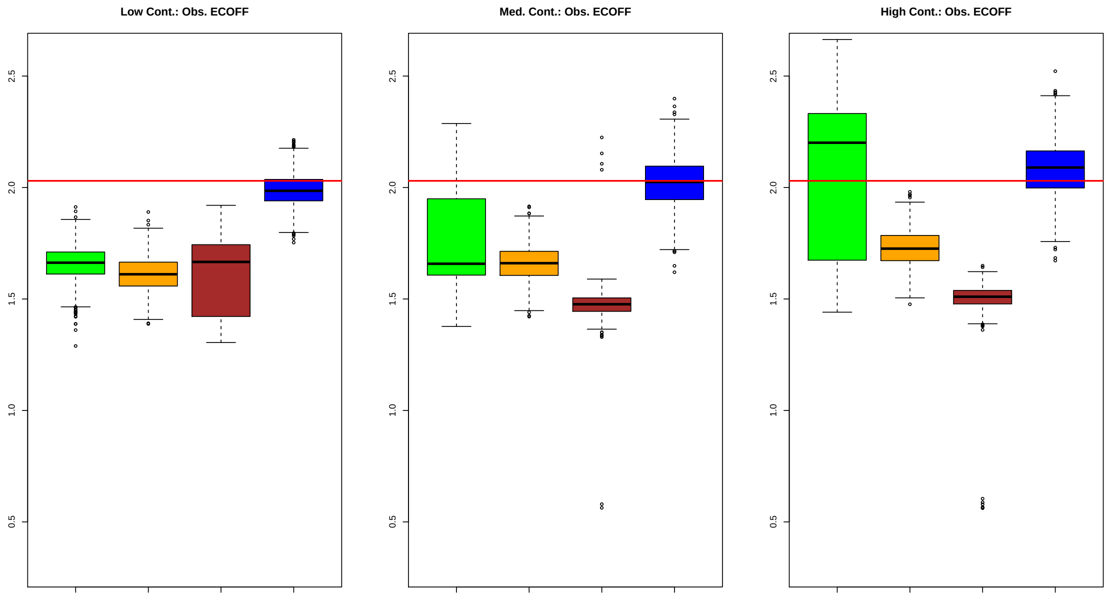


Figure 3.11. Boxplots of estimates of the observed ECOFF at size 1200 with $\sigma_\delta = 0.5$ across contamination levels. Notice the drastic changes in TURNM versus the other methods.

Table 3.11 reveals BayesACME tends to under-estimate. In fact, it only over-estimates the majority of the time at high contamination. This slight bias even at low contamination suggests an area of improvement in the selection of the base distribution or possibly considering an alternative mechanism of managing latent NWT contamination. Interestingly at high contamination, TURN over-estimates not only the other subset methods, but BayesACME as well. Given TURNM drastically under-estimates at low contamination and at medium contamination to a lesser degree, this suggests that managing contamination has become a larger issue than the misspecification of the observed WT distribution.

Table 3.11.
Numerical Summaries for Figure 3.11

Method	Cont.	Median-Truth	MAD	MADT	IQR
TURN	Low	−0.3672	**0.0498**	0.3672	**0.0987**
TURNM	Low	−0.4191	0.0538	0.4191	0.1071
JASP	Low	**−0.3638**	0.1355	**0.3638**	0.3219
BayesACME	Low	**−0.0442**	**0.0485**	**0.0582**	**0.0960**
TURN	Medium	−0.3720	0.0705	0.3720	0.3423
TURNM	Medium	**−0.3694**	0.0540	**0.3694**	**0.1079**
JASP	Medium	−0.5537	**0.0302**	0.5537	**0.0603**
BayesACME	Medium	**−0.0050**	**0.0736**	**0.0730**	0.1499
TURN	High	**0.1710**	0.2424	0.3343	0.6578
TURNM	High	−0.3041	**0.0561**	**0.3041**	**0.1131**
JASP	High	−0.5198	**0.0301**	0.5198	**0.0605**
BayesACME	High	**0.0597**	0.0842	**0.1002**	0.1660

Next we examine prevalence estimation across the three levels of increasing contamination shown from left to right in Figure 3.12. Similar to the Normal case, TURN struggles with contamination, while its modification, TURNM, is relatively

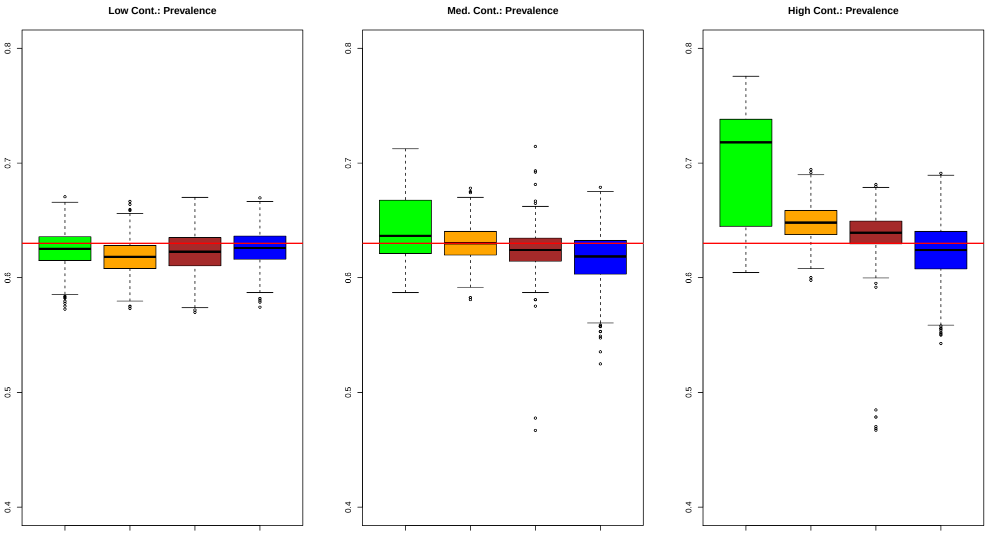


Figure 3.12. Boxplots of estimates of prevalence at size 1200 with $\sigma_\delta = 0.5$.

robust. It is important to note that the prevalence estimates from the subset methods particularly TURNM and also JASP are not biased to the same degree as the observed ECOFF estimates by ignoring ME. There is a bias in BayesACME that is most prominent at medium contamination, but decreases at high contamination. In Table 3.12, BayesACME is the least biased of the compared methods at low and high contamination. At low contamination, BayesACME is the best according to MADT. At high contamination, it has the second best MADT. At medium contamination, the subset methods appear to have smaller bias, with BayesACME having the second best MADT, indicating some sensitivity in the assumptions of the NWT distribution for BayesACME.

Table 3.12.
Numerical Summaries for Figure 3.12

Method	**Cont.**	**Median-Truth**	**MAD**	**MADT**	**IQR**
TURN	Low	**−0.0047**	0.0103	**0.0103**	0.0206
TURNM	Low	−0.0117	**0.0101**	0.0134	**0.0201**
JASP	Low	−0.0072	0.0123	0.0129	0.0246
BayesACME	Low	**−0.0041**	**0.0100**	**0.0101**	**0.0201**
TURN	Medium	0.0066	0.0197	0.0171	0.0464
TURNM	Medium	**0.0000**	**0.0103**	**0.0103**	**0.0205**
JASP	Medium	**−0.0058**	**0.0101**	**0.0105**	**0.0201**
BayesACME	Medium	−0.0114	0.0145	0.0156	0.0292
TURN	High	0.0880	0.0382	0.0880	0.0932
TURNM	High	0.0181	**0.0105**	0.0185	**0.0210**
JASP	High	**0.0093**	**0.0100**	**0.0122**	**0.0200**
BayesACME	High	**−0.0059**	0.0164	**0.0172**	0.0328

The evaluation of prevalence sheds light on the corresponding observed ECOFF results for BayesACME. As the prevalence is relatively unbiased, the method appears to struggle in determining the shape of the log2gamma. On a latent true level, the bias in both α and β is the least at low contamination and the most at medium contamination. The bias in the prevalence estimates reflects this pattern. When this happens at medium contamination, TURNM is able to be more accurate within-data sets than BayesACME.

To show the sensitivity of the prior for σ_δ^2, Figure 3.13 summarizes the changes in estimates bias of the observed ECOFF at low contamination and size 600. Each boxplot from left to right, corresponds with an increase of σ_δ. As the other methods only work on an observed level, increases in σ_δ cause the observed WT distribution (and amplify the low contamination) to become further misspecified. Thus, the increases in σ_δ increase the bias in the observed ECOFF. For BayesACME, changes in σ_δ are indicative of sensitivity to the prior on σ_δ^2. In the left panel, $\sigma_\delta = 0.4$, BayesACME over-estimates the observed ECOFF as the prior places a greater weight on a larger value. At $\sigma_\delta = 0.5$, the observed ECOFF fairly unbiased. At $\sigma_\delta = 0.6$, under-estimation occurs. Generally, these biases decrease as sample size increases.

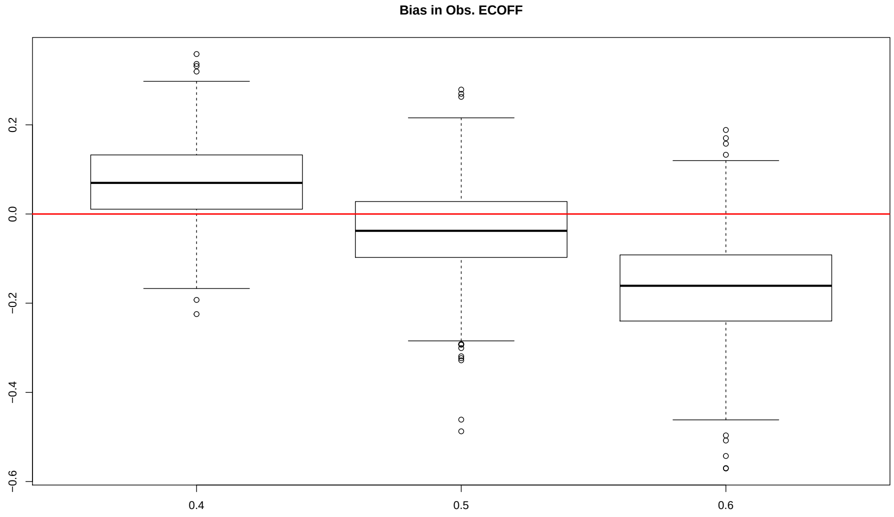


Figure 3.13. Boxplots of bias in the obs. ECOFF at size 600 with low contamination across the three levels of σ_δ.

The magnitude of the bias is greater for BayesACME at $\sigma_\delta = 0.6$ than $\sigma_\delta = 0.4$. Additionally, the precision is influenced as well as MADT is higher for $\sigma_\delta = 0.6$ than $\sigma_\delta = 0.4$. These facts demonstrate the sensitivity to the prior for σ_δ^2. As shown in

Figure 3.1, a value of $\sigma_\delta = 0.6$ is further in the right tail than $\sigma_\delta = 0.4$ in the left tail. Thus, the prior selection σ^2_δ, at least at this sample size, impacts both the magnitude of the bias as well as whether BayesACME over- or under-estimates the truth.

The question emerges about the estimation of prevalence with regards to changes in σ_δ. While the observed ECOFF is clearly impacted, the prevalence estimates tend to be adequate as shown in Figure 3.12 for a sample size of 1200.

3.8 Real Data Set Application

Professor John Turnidge provided data sets from a single Dutch lab with MIC results collected annually from 2011 to 2014. The species is *Neisseria gonorrhoeae* and the drug is Penicillin. The benefit of data collected from a single labs means that there is no between-lab variability provided there is no temporal variability.

Figure 3.14 displays the distribution of the $\log_2(MIC)$ results for 2013. The WT distribution appears to be approximately symmetric. Relying on ocular methods, the Normal distribution seems an appropriate choice.

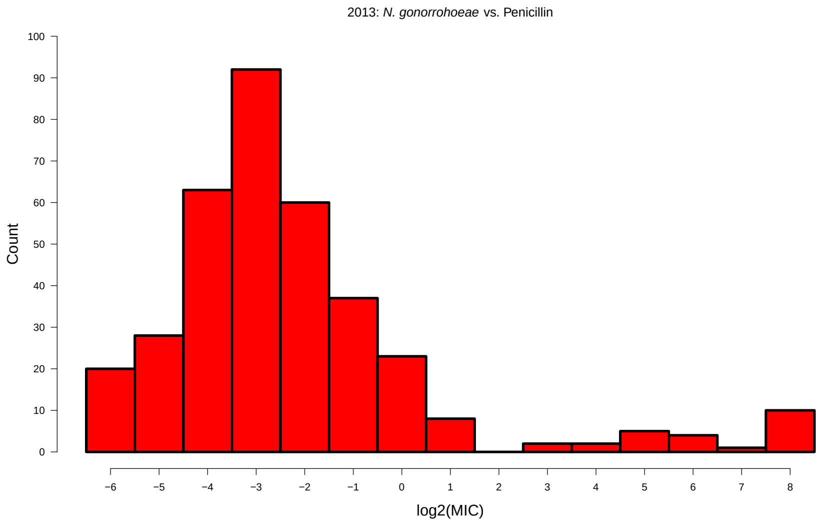


Figure 3.14. This histogram visualizes the 2013 collection of *N. gonorrhoeae* treated with Penicillin.

By working with both distributions, it is apparent in Table 3.13 that the estimate of prevalence relates to the specification of the WT distribution. In the Normal case, BayesACME produces estimates that are closest to JASP. The second closest method is JASPB. For estimation of both μ_{WT} and π, BayesACME is between JASP and JASPB. Interestingly, BayesACME produces the largest estimate of the observed standard deviation, but it is comparable to JASP and JASPB. Consequently BayesACME has the largest estimate of the observed ECOFF. The comparable performance of JASP and BayesACME is anticipated from the use of non-informative priors for $\boldsymbol{\theta_{WT}}$ and π. For large samples, it is similar to maximum likelihood estimation, especially as JASP uses eight bins encompassing the full mode.

Table 3.13.
Results from the *N. gonorrhoeae* vs. Penicillin from a single Dutch lab in 2013

Normal	$\boldsymbol{\mu_{WT}}$	$\boldsymbol{\sigma_{obs}}$	$\boldsymbol{\pi}$	**ECOFF**	**Bins**
TURN	-3.419	1.559	0.913	1.398	7
TURNM	-3.512	1.472	0.879	1.036	5
JASP	-3.491	1.656	0.936	1.768	8
JASPB	-3.163	1.571	0.933	1.692	14
BayesACME	-3.349	1.668	0.930	1.806	All

log2gamma	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\pi}$	**ECOFF**	**Bins**
TURN	1.423	12.871	0.835	-0.695	5
TURNM	1.657	17.451	0.744	-1.040	6
JASP	1.926	22.996	0.755	-1.341	5
BayesACME	1.403	11.521	0.865	-0.190	All

For the purposes of the calculating the χ^2 test statistic, the bins for $\log_2(MIC)$ values greater than one are binned together effectively decreasing the number of bins to nine. The Normal distribution has a better fit in this case. The median of the Normal test statistic posterior distribution is 9.940 while the the median of the posterior distribution for the test in log2gamma distribution is 14.003. One readily available goodness-of-fit metric is estimating the proportion that the test statistic is greater than the 95th percentile of the corresponding asymptotic $\chi^2(J-1)$ distribution

[Johnson et al., 2004]. For the Normal case, this proportion is 0.064 and for the log2gamma the proportion is 0.354. The closer this proportion is to 0.05, the better the approximation of the asymptotic distribution is for the posterior distribution. Rather than just using a particular percentile for comparison, stronger conclusions can be made by viewing the entire distribution. In fact, a visual comparison is displayed using the kernel estimates of the posterior densities of the test statistic in Figure 3.15. The Normal density is to the left of the log2gamma density.

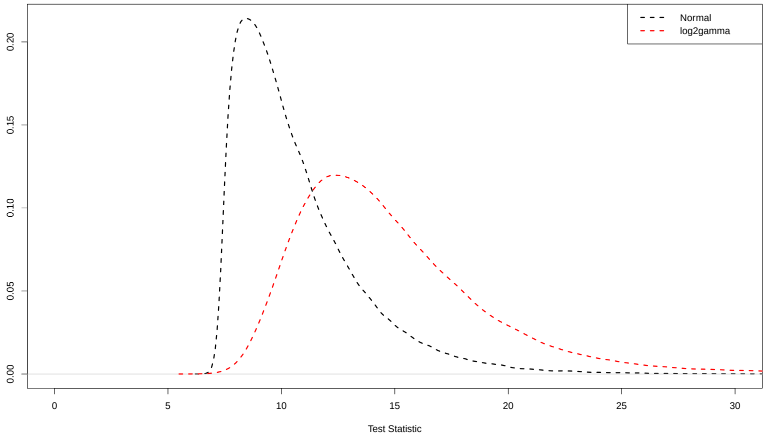


Figure 3.15. These are kernel densities of the resulting posterior distributions of the test statistic. The Normal case is the black dashed curve and the log2gamma distribution is the red-dashed curve.

In Chapter 1, a model-based alternative to the ECOFF was provided to classify a single MIC value as WT or not. Figure 3.16 produces the model-based alternative to the ECOFF for the Normal case. The figure displays the estimated probability that a random isolate with the reported $\log_2(MIC)$ is WT. For each bin, the probability of WT is determined using the equation from Chapter 1. The error bars denote the 95^{th} percent credible intervals. For a single observed MIC observation, there is a posterior probability whether the isolate is WT or not. This figure presents a clear way of assessing a probability that an isolate with a record MIC is WT or not with uncertainty. As the ECOFF value presents a particular point of distinction between the WT and NWT distribution, it is more natural of working for concentrations on a continuous level. Yet, the discrete nature of the MIC assay forces microbiologists to work with $C_1, ..., C_J$.

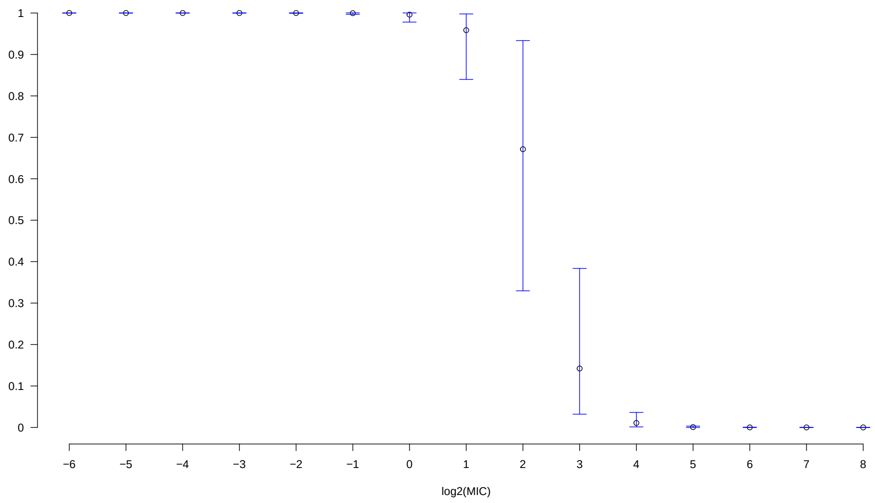


Figure 3.16. For the Normal case case, this plot illustrates the probability of an isolated with the recorded MIC labelled a WT. The blue error bars are the 95% credible intervals.

3.8.1 Conclusion

The proposed method, BayesACME, fills a void in the literature by properly addressing measurement error. In general, it shows improvement in terms of accuracy and precision. In the Normal case, the other Bayesian semiparametric method does show better behavior at high contamination. Certainly, it warrants future investigation in improving BayesACME whether that involves introducing informative priors for the WT distribution or improving the nonparametric estimation for the NWT distribution.

The benefit of modelling measurement error is demonstrated in the observed ECOFF estimates in the log2gamma case. BayesACME is the only method that properly specifies the observed distribution. Consequently, BayesACME is the most accurate method for estimating the observed ECOFF in the log2gamma case. Interestingly, the advantages of modelling measurement error do not necessarily carry over to prevalence estimation.

There is the question of what happens if π approaches 1.0? At $\pi = 1.0$, the observed MIC values are no longer a mixture so BayesACME as a mixture model is inappropriate. The subset methods still work, but at the cost of under-utilizing the data. Now if π is large but less than 1.0, then there may not be too much gained by modelling the NWT distribution. Thus, BayesACME is likely to have a competitive edge when the NWT distribution has some substantial weight. The trade-off is how well can BayesACME estimate the left NWT tail? Little to no assumptions are made by the subset methods and P-splines. The selection of the base distribution for the DPMM is presented as a heuristic. In the concluding chapter, we discuss an alternative way to control the contamination in the base distribution. Given the good performance from TURNM and JASP at estimating prevalence and the need for BayesACME to further restrict the WT distribution in high contamination, incorporating informative priors on π may be a starting place for future work in improving model robustness to contamination.

For both the Normal and even more so for the log2gamma there is sensitivity to the prior for σ_δ^2. So far the prior σ_δ^2 has been investigated by considering unlikely, but plausible, values in its tails. From the results of the simulation studies, we are aware of the impact the prior selection for σ_δ^2 has. Yet, it may be worthwhile to examine results that are further in the tails of the prior distribution.

In the real data analysis, BayesACME has similar performance to JASPB in the Normal case. Interestingly, it produces the greatest estimate for the observed ECOFF in both the Normal and log2gamma case. It is possible for this particular Dutch lab that the measurement error variance may be smaller than is implied with the chosen prior on σ_δ^2.

Given that labs produce annual data, a natural next step is extending the proposed methodology for the purpose of estimating WT prevalence over time. By doing so, BayesACME can serve as a key tool for AMR surveillance.

4. MULTIYEAR AST MONITORING

While single-year estimation is important, it is usually the AMR trend that is of interest to microbiologists, clinicians, and policymakers. In fact, Fuhrmeister and Jones [2019] named longitudinality a key element of AMR surveillance. This chapter extends our single-year approach to handle multiyear data. While one could analyze each year separately and piece together a trend, we think that more can be gained from a joint analysis.

Underlying our proposed method is the premise that the WT distribution does not change over time. It is the natural distribution of bacterial strains that have not yet exhibited resistance mechanisms. This premise is supported on the EUCAST website as their definition of the WT distribution includes the statement that it is the same in space, time, and source (i.e., animal or environment) [EUCAST, 2017].

Given that strains are leaving the WT distribution, this means they are being added to the NWT distribution. However, the MICs of these mutated strains are unknown. Furthermore, additional mutations of the current NWT strains may alter their MIC. This means that the NWT distribution is not static. The changes are not anticipated to be dramatic year to year, but there would be changes nonetheless.

4.1 Previous Literature

In the literature, there is only one method designed to detect temporal changes in AMR. As was done in Chapter 3, Zhang et al. [2020] propose a Bayesian latent class mixture model to describe the observed MIC distribution. However, their single-year mixture model and how they allow the WT and NWT subpopulations to change over time are both very different. First, they assume the observed continuous MIC distribution is simply a mixture of two Normals. Second, they only allow the WT distribution to change over time while the NWT distribution is static. Specifically, for $t = 1, ..., T$, their model for year t is

$$f_t(\mathbf{X}^*) = \pi_t f_{t,WT}(\mathbf{X}^*) + (1 - \pi_t) f_{t,NWT}(\mathbf{X}^*)$$

where

- $f_{t,WT} \sim N(\gamma_0 + \gamma_1 t, \sigma_{obs})$
- $f_{t,NWT} \sim N(\mu_{NWT}, \sigma_{NWT})$
- $\ln(\frac{\pi_t}{1-\pi_t}) \sim N(\theta, \nu)$

Thus, they allow the WT distribution's mean to drift linearly with time. A positive value for the slope, γ_1, implies that AMR is increasing with time. A negative value of γ_1 implies that AMR is decreasing with time. The NWT distribution, on the other hand, remains static. They argue this is reasonable because censoring prevents this trend from being observed. The standard deviations for the WT and NWT distributions are also assumed invariant over time. There is no assumption or anticipation regarding how prevalence changes over time. Thus, the researchers' goals based on fitting this model are fundamentally different from the goals in this chapter.

There is very little discussion in the paper regarding their choice of model. The authors state their assumed linear trend in the WT mean is built on a "naive analysis" of *Salmonella enteric* and the antibiotic CHL in the CDC NARMS data set. We, however, are a bit perplexed over their choice of MIC evolution. We are also concerned about the restrictive assumption that the distribution for each subpopulation is Normal.

4.2 Multiyear Extensions of the Subset Methods

Under our assumption that the WT distribution is static over time, both subset methods can easily be extended for multiyear analysis. We will briefly describe these extensions here. Full details are provided in Appendix E.

Given that TURNM handled contamination better in the single-year analysis, we extend it to multiyear data. This extension means that there is a different prevalence π_t for each of the T years, but a common WT distribution (i.e., single set of parameters $\boldsymbol{\theta_{WT}}$). Using the cumulative counts $B_{t,j}$ for the left-most bins up to the WT mode plus one bin in year t, the following objective function is minimized with respect to $\boldsymbol{\theta_{WT}}$ and $\pi_1, ..., \pi_T$:

$$\sum_{t=1}^{T}\sum_{j=1}^{K_t}[B_{t,j} - N_{t,tot}\boldsymbol{\pi}_t \cdot F_{WT}(C_{t,j}; \boldsymbol{\theta_{WT}})]^2$$

The cumulative distribution function F_{WT} can either be Normal or log2gamma.

For the multiyear extension of JASP, the challenge is determining the number of bins to use in each year. Given the set $(K_1, ..., K_T)$, each year t represents a new independent draw from a multinomial where the bin probabilities for the first K_t bins are based on the same WT distribution and the remaining $J_t - K_t$ bins are free to vary under the restriction that the sum of the bin probabilities in each year must equal one. To avoid the search for the best set $(K_1, ..., K_T)$, we propose applying JASP to each year, and then computing the weighted average of each year's estimate $\boldsymbol{\theta_{t,WT}}$ (using the the observed Hessians for weights). We then update each year's prevalence estimate using the overall estimate of $\boldsymbol{\theta_{WT}}$. It is important to note that $\boldsymbol{\theta_{WT}}$ is not necessarily the MLE for the T years, but rather it is a weighted average of T years using approximate standard errors.

4.3 Multiyear Extension of BayesACME

Similar to the two subset methods, we consider a static WT distribution over T years, but allow the WT prevalence and the NWT distribution to vary. Thus, $\boldsymbol{\theta_{WT}}$ and the $\boldsymbol{\theta_{t,NWT}}$ represent the parameters of the true MIC assay distributions for a given year t. We again deconvolve the continuous observed MIC distribution into the true MIC distribution and measurement error distribution and consider both the Normal and log2gamma distributions for g_{WT}. Likewise, we describe each NWT distribution $g_{t,NWT}$ using a mixture of Normals. For the last 50 years, the MIC assay has, for the purposes of this dissertation, remained unchanged. For that reason, there are no temporal assumptions made regarding measurement error.

Analogous to Chapter 3, we introduce latent values $X_{t,i}$ and $\delta_{t,i}$ to correspond with each $Y_{t,i}$. We also consider an indicator $c_{t,i}$, which denotes if whether isolate i in year t is WT or not. This allows us to specify the model as :

$$c_{t,i} \sim Bernoulli(\pi_t)$$

$$X_{t,i} \sim (g_{WT}(X_{t,i}))^{c_{t,i}}(g_{t,NWT}(X_{t,i}))^{1-c_{t,i}}$$

$$\delta_{t,i} \sim N(0, \sigma_\delta)$$

$$Y_{t,i} = \lceil X_{t,i} + \delta_{t,i} \rceil$$

The likelihood can similarly be expressed as

$$\prod_{t=1}^{T} \prod_{i=1}^{N_{t,tot}} h^*(Y_{t,i} \mid X_{t,i}, \sigma_\delta) g(X_{t,i} \mid \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{t,NWT}}, c_{t,i}) p(c_{t,i} \mid \pi_t)$$

where p is the pmf of the Bernoulli distribution, g is either the Normal or log2gamma pdf, and h^* is a truncated Normal density. The set of unknowns:

$$\{\mathbf{X_1}..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}, \boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}, \pi_1, ..., \pi_T, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{1,NWT}}, ..., \boldsymbol{\theta_{T,NWT}}, \sigma^2_\delta\}$$

has increased due to the introduction of latent vectors, but the benefit of this model specification is that it again makes working with non-Normal WT distributions far more straightforward.

4.3.1 Prior Selection

The priors selected for $\boldsymbol{\theta_{WT}}$ and σ_δ^2 are the same as in the single-year case. Similarly, the priors for $\pi_t \overset{i.i.d.}{\sim} Unif(0.5, 1.0)$ for $t = 1, ..., T$.

We again use a DPMM to model each year's NWT distribution. While we make no assumptions about the shape across years, realistically the changes should be subtle. How subtle, however, is unclear [Mouton et al., 2018]. For now, the same base distribution is assigned to each year's NWT distribution. As each year has the same base distribution, the same general strategy applies as in the single-year case. The modification to estimating A using multiple years is described alongside setting the initial values. The same process to select the priors for σ_B^2 is used as well. Each year's concentration parameter $\alpha_{t,conc}$ has the same prior. Specifically, priors for σ_B^2 from the simulation studies Chapter 3 are used again. In the future, if more information is available about the evolution of the NWT over time, at least for a specific "drug/bug" combination, it can be incorporated into the model.

4.4 Computation

Given that the extension to multiple years is simply another level in our hierarchical model, we again take a Bayesian approach. The general computational approach is Metropolis-within-Gibbs.

4.4.1 Setting Initial Values

Given that we now have multiple years of data describing the WT, we now implement the multiyear TURNM method to generate initial estimates of $\boldsymbol{\theta_{WT}}^{(0)}$ and $\pi_1{}^{(0)}, ..., \pi_T{}^{(0)}$. The mode for the prior of σ_δ^2 is the initial value $\sigma_\delta^{2(0)}$. Like in the single-year scenario, we determine the initial value for the latent true WT variance by subtracting the initial value for the ME variance from the observed WT distribution variance in the Normal case. In the log2gamma case, measurement error is ignored because both α and β require modification. For each year $t = 1, ..., T$, we then proceed to generate the latent vectors using the same strategy in Chapter 3.

In Chapter 3 the value A is determined using a function of $\boldsymbol{\theta_{WT}}$. The process to determine $\hat{A}$ is the same here except $\boldsymbol{\theta_{WT}}$ is estimated initially with the multiyear extension of TURNM. A burn-in of 5000 iterations is used to determine $\hat{\hat{A}}$. The process for selecting the prior on the variances is the same as the single-year case detailed in Chapter 3.

4.4.2 Updates to the Model Parameters

In the Normal case, μ_{WT} is updated in a Gibbs step. Both σ^2_{WT} and σ^2_δ are jointly updated with a Metropolis step. This is different than the single-year case because it handles the initial values $\boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}$ and consequential correlation between the two variance components better than two separate Gibbs steps. In the log2gamma case, α is updated in a Metropolis step and β is updated in a Gibbs step. Also, σ^2_δ is updated in a Gibbs step. Beyond those updates the algorithms for the Normal and log2gamma case are very similar: π_1,...,π_T, $\mathbf{X_1}, ..., \mathbf{X_T}$, $\mathbf{c_1}, ..., \mathbf{c_T}$, and $\boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}$ are updated in Gibbs steps. Each year's DPMM is updated using Algorithm 8 from Neal [2000]. The concentration parameters $\alpha_{1,conc}, ..., \alpha_{T,conc}$ are each updated in separate Gibbs steps using West [1992]. Appendix F contains the details for both the Normal and log2gamma distributions.

Outline of Computation

The outline of the algorithms is very similar. In the Normal case, a non-trivial difference is σ^2_{WT} and σ^2_δ are jointly updated in a Metropolis step.

1. Set initial values for parameters

2. Determine base distribution of DPMM

3. Update Model Parameters for a Set Number of Iterations

- Update $\boldsymbol{\theta_{WT}} \mid \mathbf{X_1}, ..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}$ and $\sigma^2_\delta \mid \boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}$
  - Normal
    - Update $\mu_{WT} \mid \sigma^2_{WT}, \mathbf{X_1}, ..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}$
    - Update $(\sigma^2_{WT}, \sigma^2_\delta) \mid \mu_{WT}, \mathbf{X_1}, ..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}, \boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}$
  - log2gamma
    - Update $\alpha \mid \beta, \mathbf{X_1}, ..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}$
    - Update $\beta \mid \alpha, \mathbf{X_1}, ..., \mathbf{X_T}, \mathbf{c_1}, ..., \mathbf{c_T}$
    - Update $\sigma^2_\delta \mid \boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T}$
- For $t = 1, ..., T$, update $\pi_t \mid \mathbf{c_t}$
- For $t = 1, ..., T$, update $\boldsymbol{\theta_{t,NWT}} \mid \mathbf{X_t}, \mathbf{c_t}$
- For $t = 1, ..., T$, update $\mathbf{X_t} \mid \mathbf{Y_t}, \mathbf{c_t}, \boldsymbol{\delta_t}$
- For $t = 1, ..., T$, update $\boldsymbol{\delta_t} \mid \mathbf{X_t}, \mathbf{c_t}, \mathbf{Y_t}, \sigma^2_\delta$
- For $t = 1, ..., T$, update $\mathbf{c_t} \mid \mathbf{X_t}, \pi_t, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{t,NWT}}$

4. Discard burn-in

4.5 Multiyear Simulation Study

In this section, we perform a simulation study to compare the two extended subset methods with our BayesACME extension. Because our focus is on the change in prevalence, we make the simplifying assumption that only the WT prevalence changes over time and the WT and NWT distributions remain static. In reality, the NWT distribution would also change, but because none of these methods make any assumptions regarding the similarity of the NWT distribution across time, we do not consider it to be a drawback. This also makes it much easier to compare these results to the single-year results under low, medium, and high contamination.

For our simulations, we consider a study over four years (i.e., $T = 4$) under different levels of contamination and sample sizes. For measurement error, we just use $\sigma_\delta = 0.4$ as it is a "low" level of ME. The implication is the log2gamma is the TURNM and JASP produce biased ECOFF estimates, but that is not the focus in this chapter. Because the true value of σ_δ is not centered with the prior, BayesACME produces biased observed ECOFF estimates in the log2gamma case as shown in Figure 3.13, but likely less biased estimates than those from the subset methods. The bias of these methods produce regarding the observed ECOFF is not the focus in this chapter, rather the focus is on the prevalence estimation as it pertains to capturing AMR trend.

In terms of a decline in WT prevalence over time, we consider two cases: a slow rate where the prevalence remains relatively constant over time and a fast rate. Because the prevalence of AMR grows logistically [John Turnidge, personal communication, January 27, 2020], we consider two rates of logit decline. These true WT prevalence values are shown in the Table 4.1 and displayed in Figure 4.1.

Table 4.1.
Changes in Wild-Type Prevalence over Four Years

Rate	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Slow	0.7600	0.7545	0.7489	0.7433
Fast	0.7600	0.6963	0.6180	0.5267

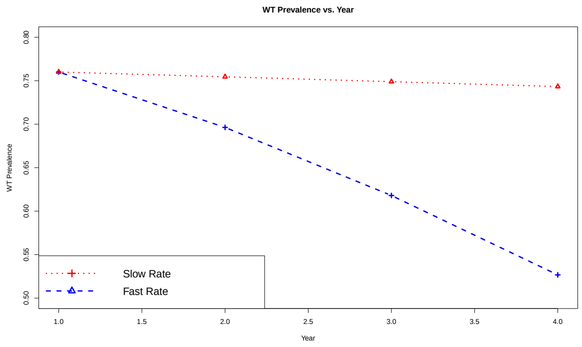


Figure 4.1. This plot shows the trend for the fast (in blue) and slow (in red) rates of decline for the WT prevalence.

We perform a simulation study to demonstrate its application for both the Normal and log2gamma cases. We use the same latent WT distribution as in Chapter 3. For the NWT distribution, we again consider a mixture of two Normals, with components that vary by contamination level. The medium and high contamination are different from the single-year levels because the fast rate decreases π to a value where the overall MIC distribution appears to be unimodal. To maintain biological plausibility, different NWT distributions for the medium and high case are assigned.

- Low Contamination:

$$\mathbf{w_{NWT}} = (\tfrac{5}{37}, \tfrac{32}{37}) \qquad \boldsymbol{\mu_{NWT}} = (4.60, 5.10) \qquad \boldsymbol{\sigma_{NWT}} = (0.95, 1.10)$$

- Medium Contamination:

$$\mathbf{w_{NWT}} = (\tfrac{15}{37}, \tfrac{22}{37}) \qquad \boldsymbol{\mu_{NWT}} = (3.50, 5.00) \qquad \boldsymbol{\sigma_{NWT}} = (1.10, 1.30)$$

- High Contamination:

$$\mathbf{w_{NWT}} = (\tfrac{22}{37}, \tfrac{15}{37}) \qquad \boldsymbol{\mu_{NWT}} = (2.50, 5.00) \qquad \boldsymbol{\sigma_{NWT}} = (0.80, 1.30)$$

The endpoints for censoring are -4 and 7. Each year has the following sample sizes: 300, 600, and 1200. We perform 1000 simulations for each scenario. Table 4.2 reports the Bhattacharyya coefficients for the observed distribution (both continuous and censored) in the Normal and log2gamma cases, respectively. As discussed in

Table 4.2.
Contamination for the Multiyear Simulation Study with ME

Normal	**Low**	**Medium**	**High**
Continuous	0.0259	0.0945	0.1640
Censored	0.0323	0.1073	0.1854

Log2gamma	**Low**	**Medium**	**High**
Continuous	0.0171	0.0836	0.1517
Censored	0.0237	0.0993	0.1801

Chapter 3, the Bhattacharyya coefficient does not take into account prevalence π. Ignoring prevalence becomes more of an issue in the multiyear case because it changes each year. In future work, a scalar quantity to describe contamination that takes into account π is desired.

Similar to the single-year analysis, boxplots and numerical summaries are displayed. As the focus is on the trend, numerical summaries are only presented for the trend analysis. There are two ways of conducting trend analysis. The first is applying the L_2 loss function to the set of prevalence estimates with the respective true values. The other way to quantify the trend is exploiting the pattern of WT decline used in the simulation study. Recall that in the simulation study assumed the logit of the true prevalence values decreases linearly. Then properties comparing the slope of the fitted least squares line are compared. The prevalence estimates for each year are compared using side-by-side boxplots. The trend results are presented using numerical summaries. The numerical summaries reported are

- Median $\ln(L_2)$ (Med. $\ln(L_2)$)
- Bias
- Variance (Var.)
- Mean Squared Error (MSE)

The first quantity, Median $\ln(L_2)$, is the median of the natural logarithm of the L_2 loss of each year's prevalence estimates with their respective true values. This distribution tends to be skewed so the median is selected. Unlike Chapter 3, the distributions tend to be reasonably symmetric so robust statistics are not necessary. Bias is defined here as the arithmetic average minus the true value. We use the sample variance (Var.) and the MSE (the bias squared plus the variance). The estimated MSE serves an indicator of the best overall estimation method. As only three methods are compared, only the best quantity is in boldface. As the results are rounded, ties are settled by looking at the next decimal place. For brevity only some of the results are discussed in this chapter. More are contained in Appendix G.

4.5.1 Normal Results

Slow Rate

Figure 4.2 shows the bias of prevalence. Each panel from left to right represents year 1 to year 4. Figure 4.2 shows that JASP is the least biased across the four years, while BayesACME is generally the most precise. Utilizing the least number of bins, TURNM is the least precise.

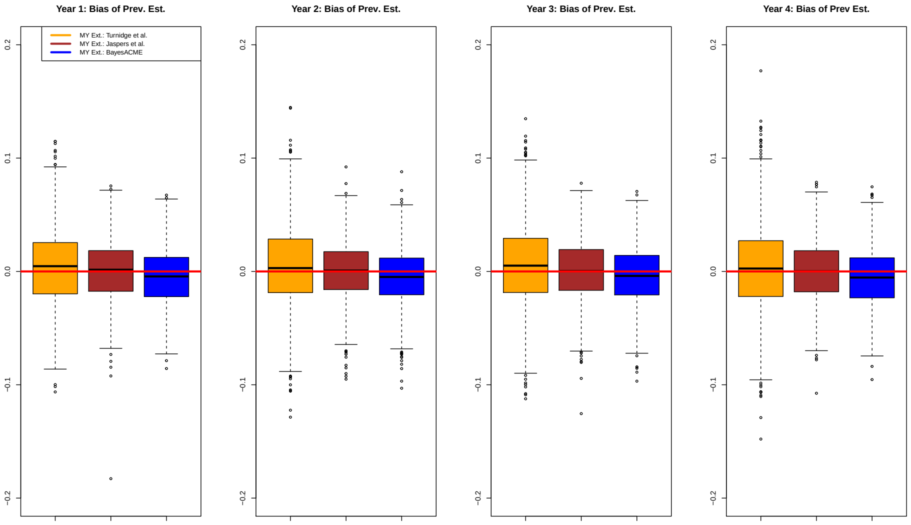


Figure 4.2. Normal distribution at low contamination at Size 300 at the slow rate of WT prevalence decline

Figure 4.3 shows the results with high contamination. As the high contamination violates the "clear separation" assumption in each individual year, JASP handles the contamination the worst. It is the most biased and the least precise in each of the four years. As TURNM excludes bins far on the right tail of the WT distribution, the contamination biases the estimates of TURNM to a lesser extent than JASP.

BayesACME is relatively robust to the contamination; it is the most accurate and most precise in each year.

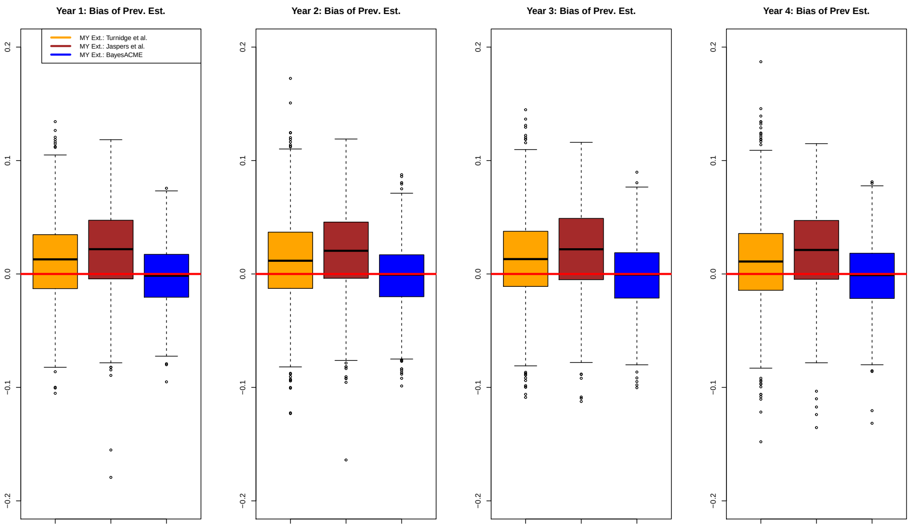


Figure 4.3. Normal distribution at high contamination at size 300 at the slow rate of WT prevalence decline

The previous figures provide only a marginal analysis of each year's prevalence estimate. As the goal is multiyear monitoring of WT prevalence, understanding the trend is key. Table 4.3 displays two different forms of trend analysis. The column Median $\ln(L_2)$ is using the L_2 loss of the estimates from the four years using the true values. To assist comparison, the natural logarithm is taken (i.e., the more negative the better). Generally, BayesACME is the best by having the smallest median value at each contamination level. JASP is the second best in terms of median at low and medium contamination. At high contamination, TURMN outperforms JASP in terms of the median L_2 loss.

Exploiting the simulation study design, the logit of the true prevalence values decreases linearly. Then the bias of the estimated slope coefficient from the estimates is displayed. Table 4.3 reveals that at each contamination level BayesACME has the

Table 4.3.
Numerical Summaries for Normal Distribution at Slow Rate

Method	**Cont.**	**Med.** $\ln(L_2)$	**Bias**	**Var.**	**MSE**
TURNM	Low	−5.4240	−0.0005	0.0077	0.0077
JASP	Low	−6.1102	−0.0002	0.0038	0.0038
BayesACME	Low	**−6.1403**	**−0.0002**	**0.0035**	**0.0035**
TURNM	Med.	−5.3995	−0.0005	0.0078	0.0078
JASP	Med.	−5.9099	0.0009	0.0043	0.0043
BayesACME	Med.	**−6.0795**	**0.0004**	**0.0037**	**0.0037**
TURNM	High	−5.3029	−0.0012	0.0082	0.0082
JASP	High	−5.1073	0.0014	0.0075	0.0075
BayesACME	High	**−5.9028**	**−0.0001**	**0.0042**	**0.0042**

lowest absolute bias and variance. Consequently, BayesACME is the best method in terms of MSE at capturing the trend.

Fast Rate

Figure 4.4 has similar results to the slow rate regarding bias. The exception is year 4 for BayesACME because the true value is close to the lower limit on the prior for prevalence. This prior results in a slightly more biased, but more precise marginal posterior for year 4. The boxplots show across the four years, JASP is now the most accurate of the three methods, but BayesACME is still generally the most precise. Similiar to the slow rate, TURNM is the least precise from using the least number of bins.

At high contamination shown in Figure 4.5, the results are similar to the slow case. It is clear that BayesACME performs the best in this setting. Across the four years, BayesACME, like with the slow rate, is the most accurate and most precise. There does appear to be a relatively stronger performance by JASP than in the slow rate. This suggests in most years, it has better subset selection than at the slow rate.

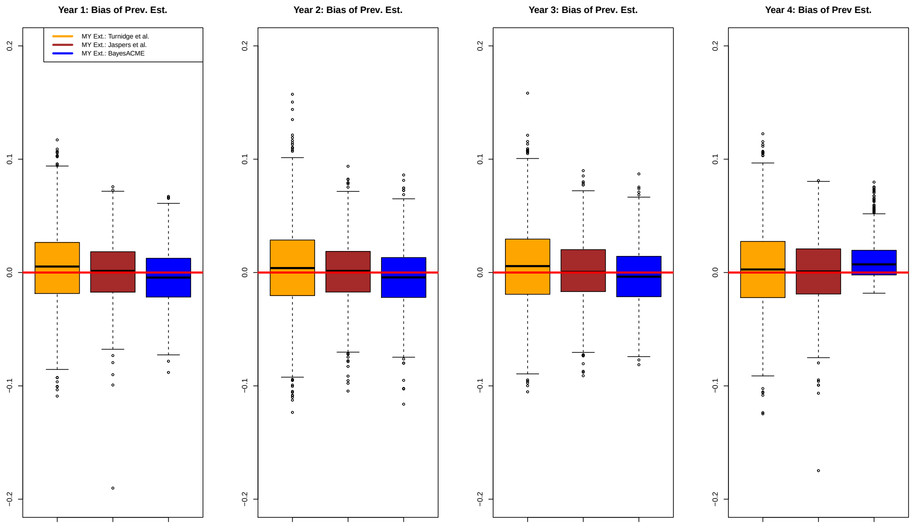


Figure 4.4. Normal distribution at low contamination at Size 300 at the fast rate of WT prevalence decline

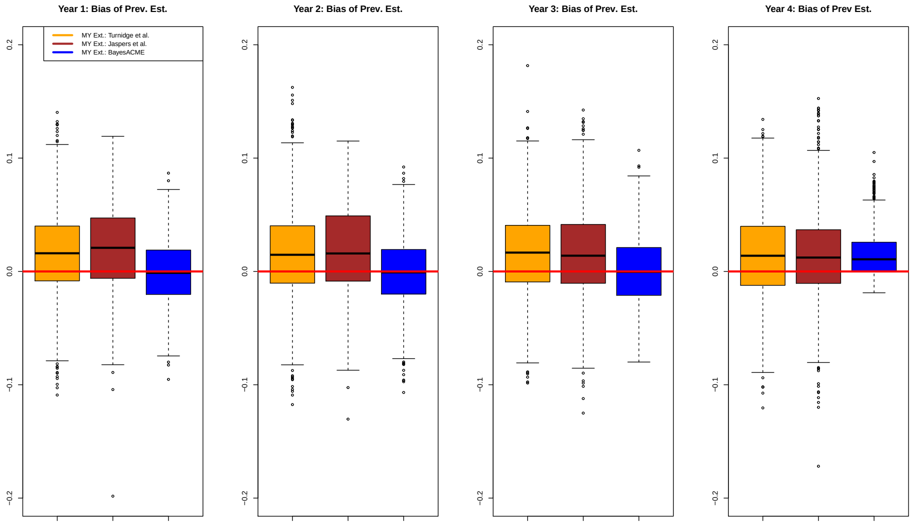


Figure 4.5. Normal distribution at high contamination at size 300 at the fast rate of WT prevalence decline

As reflected in Table 4.4, BayesACME has the smallest median in terms of the natural logarithm of L_2 loss. In the analysis of the slope of the logit of the prevalence estimates, BayesACME is the most precise across contamination levels. It is no longer the case that BayesACME is the most accurate at capturing the trend of prevalence changes in terms of absolute bias. However, it is still the best method at the three contamination levels in terms of MSE. In the future, it is possible to examine the influence the $Unif(0.5, 1.0)$ versus $Unif(0.0, 1.0)$ has on the bias of detecting the trend. At low and medium contamination, the subset methods have smaller bias.

Table 4.4.
Numerical Summaries for Normal Distribution at Fast Rate

Method	**Cont.**	**Med.** $\ln(L_2)$	**Bias**	**Var.**	**MSE**
TURNM	Low	−5.3692	−0.0060	0.0061	0.0062
JASP	Low	−5.9556	**−0.0012**	0.0035	0.0035
BayesACME	Low	**−6.1484**	0.0205	**0.0024**	**0.0028**
TURNM	Med.	−5.3289	−0.0081	0.0063	0.0064
JASP	Med.	−5.6896	**0.0013**	0.0039	0.0039
BayesACME	Med.	**−6.0988**	0.0218	**0.0026**	**0.0031**
TURNM	High	−5.1583	**−0.0152**	0.0069	0.0072
JASP	High	−5.1380	−0.0196	0.0066	0.0069
BayesACME	High	**−5.9041**	0.0184	**0.0030**	**0.0034**

4.5.2 Log2gamma Results

Slow Rate

Figure 4.6 show the results for low contamination at size 300. Except for year 4, BayesACME is the most precise in terms of IQR. The results for best performance in terms of bias are mixed. BayesACME is the most accurate in years 1 and 4. JASP is the most accurate in years 2 and 3.

At high contamination shown in Figure 4.7, the multiyear extensions of TURNM and BayesACME have comparable precision with a very slight edge to TURNM. TURNM is the most accurate across the four years. JASP struggles with contamination and suffers from a loss of precision.

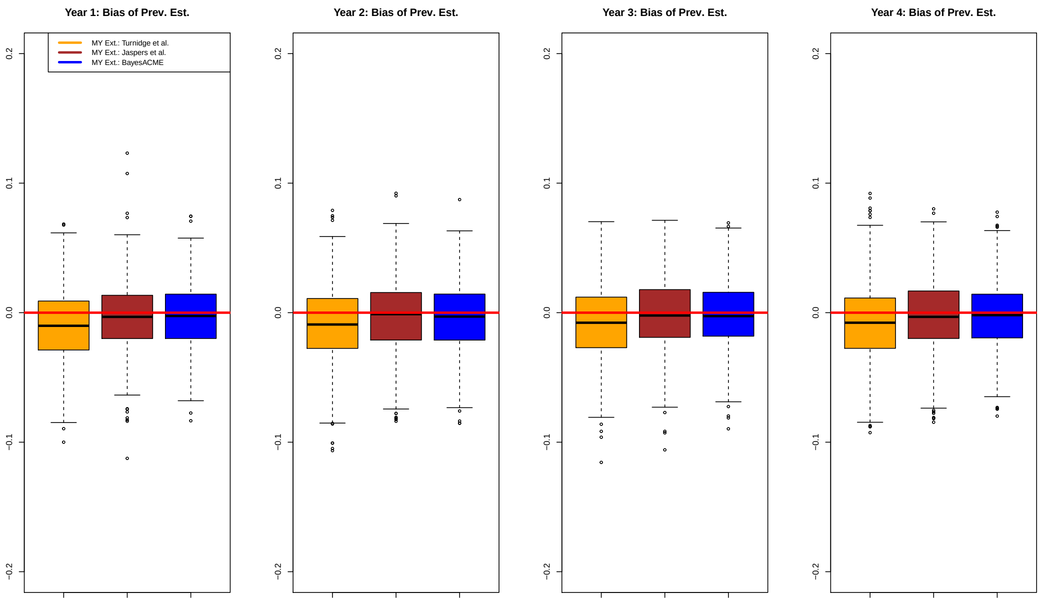


Figure 4.6. Log2gamma distribution at low contamination at size 300 at the slow rate of WT prevalence decline

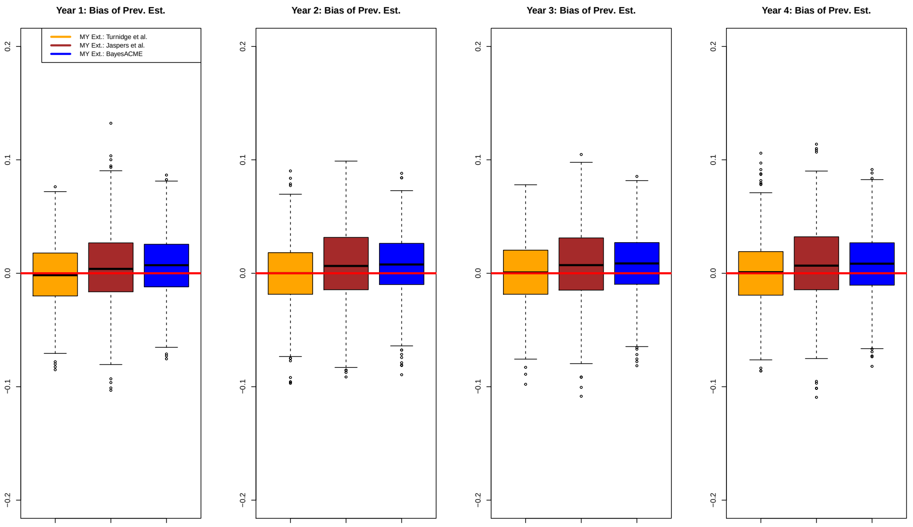


Figure 4.7. Log2gamma distribution at high contamination at size 300 at the slow rate of WT prevalence decline

Of course, the focus is capturing the trend. At the slow rate of decline, it appears at low contamination in terms of natural logarithm of the L_2 loss with the true values the multiyear extension of BayesACME has the best performance. This pattern becomes slightly more pronounced with increases in sample size. Table 4.5 shows the estimated bias of the slope of the logit of the prevalence estimates shows that BayesACME is the most accurate in terms of absolute bias and the most precise at all three contamination levels. In terms of MSE, BayesACME is the best method at these settings.

Table 4.5.
Numerical Summaries for log2gamma Distribution at Slow Rate

Method	**Cont.**	**Med.** $\ln(L_2)$	**Bias**	**Var.**	**MSE**
TURNM	Low	-5.7936	0.0037	0.0047	0.0047
JASP	Low	-6.0334	0.0030	0.0039	0.0040
BayesACME	Low	$\mathbf{-6.1036}$	**0.0026**	**0.0036**	**0.0036**
TURNM	Med.	-5.8479	0.0037	0.0048	0.0048
JASP	Med.	-5.9477	0.0030	0.0043	0.0044
BayesACME	Med.	$\mathbf{-6.0913}$	**0.0023**	**0.0037**	**0.0037**
TURNM	High	-5.8721	0.0034	0.0049	0.0049
JASP	High	-5.5341	0.0025	0.0055	0.0055
BayesACME	High	$\mathbf{-5.8867}$	**0.0023**	**0.0042**	**0.0042**

Fast Rate

At low contamination in Figure 4.8, there is still the slight bias in TURNM. The estimates in the multiyear extension of JASP are nearly unbiased. BayesACME is nearly unbiased except for year 4 where the prior for π_4 becomes informative. For all four years, there is a slight bias in the extension of TURNM. The extension of JASP is quite good at being nearly unbiased. Generally, BayesACME is the most precise.

Figure 4.9 shows the results at high contamination when there is a fast decline in WT prevalence. BayesACME bests JASP in terms of precision in each of the four

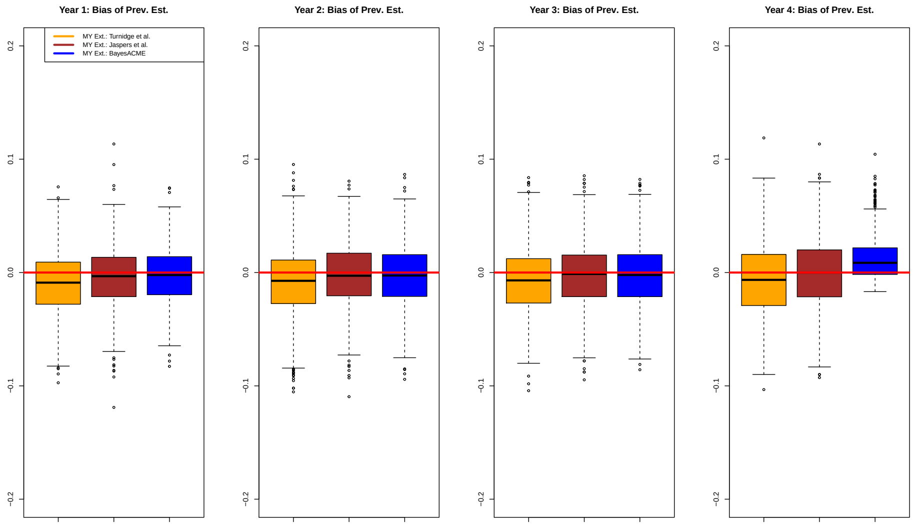


Figure 4.8. Log2gamma distribution at low contamination at size 300 at the fast rate of WT prevalence decline

years of prevalence. It does appear that the estimates from TURNM are the least biased and with the exception of year 4 are the most precise.

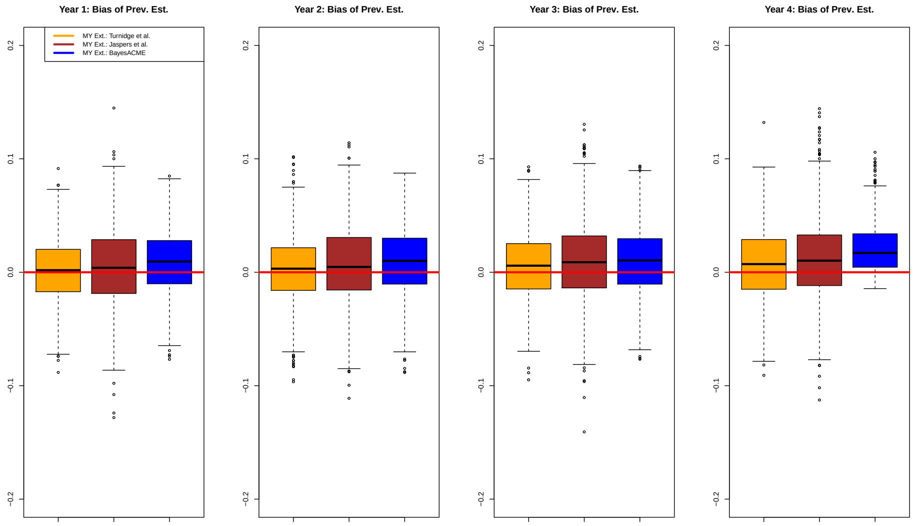


Figure 4.9. Log2gamma distribution at high contamination at size 300 at the fast rate of WT prevalence decline

Table 4.6 displays the natural logarithm of the L_2 loss of the prevalence estimates with the true values. Again BayesACME does the best in terms of median. Similar to the Normal case, at the fast rate, BayesACME is no longer the most accurate in terms of absolute bias. It is still the most precise across contamination levels and is the bias is still mild enough that BayesACME is the best method in terms of estimated MSE. In terms of slope of the logit of the prevalence estimates, BayesACME is the most precise across contamination levels. Except at high contamination, JASP is the most accurate.

Table 4.6.
Numerical Summaries for log2gamma distribution at Fast Rate

Method	**Cont.**	**Med.** $\ln(L_2)$	**Bias**	**Var.**	**MSE**
TURNM	Low	-5.7470	-0.0060	0.0061	0.0062
JASP	Low	-5.9177	$\mathbf{-0.0012}$	0.0035	0.0035
BayesACME	Low	$\mathbf{-6.1574}$	0.0205	**0.0024**	**0.0028**
TURNM	Medium	-5.7966	-0.0081	0.0063	0.0064
JASP	Medium	-5.7499	**0.0013**	0.0039	0.0039
BayesACME	Medium	$\mathbf{-6.0596}$	0.0218	**0.0026**	**0.0031**
TURNM	High	-5.7486	$\mathbf{-0.0152}$	0.0069	0.0072
JASP	High	-5.5100	-0.0196	0.0066	0.0069
BayesACME	High	$\mathbf{-5.7922}$	0.0184	**0.0030**	**0.0034**

4.6 Real Multiyear Data Set Application

Professor John Turnidge provided data sets from a single lab with MIC results collected annually from 2011 to 2014. The advantage of results from a single lab is the elimination of between-lab variability. This advantage enables a longitudinal analysis without that concern. We examine the collection of MIC results of *Neisseria gonorrhoeae* treated with Penicillin. In Chapter 3, the data from 2013 was considered. Now four consecutive years (2011-2014) are considered jointly.

Figure 4.10 shows the MIC distributions for each year. Interestingly, the data from 2011 differs the most on the left. The difference is in terms of the left-most mode and the amount of skewness. The distribution for 2011 has a much rounder mode than the other three years and the distribution is much more right-skewed than the others. Each year's data set has the following sample size: 286, 276, 355, and 327, respectively. Each data set is tested with the same concentrations.

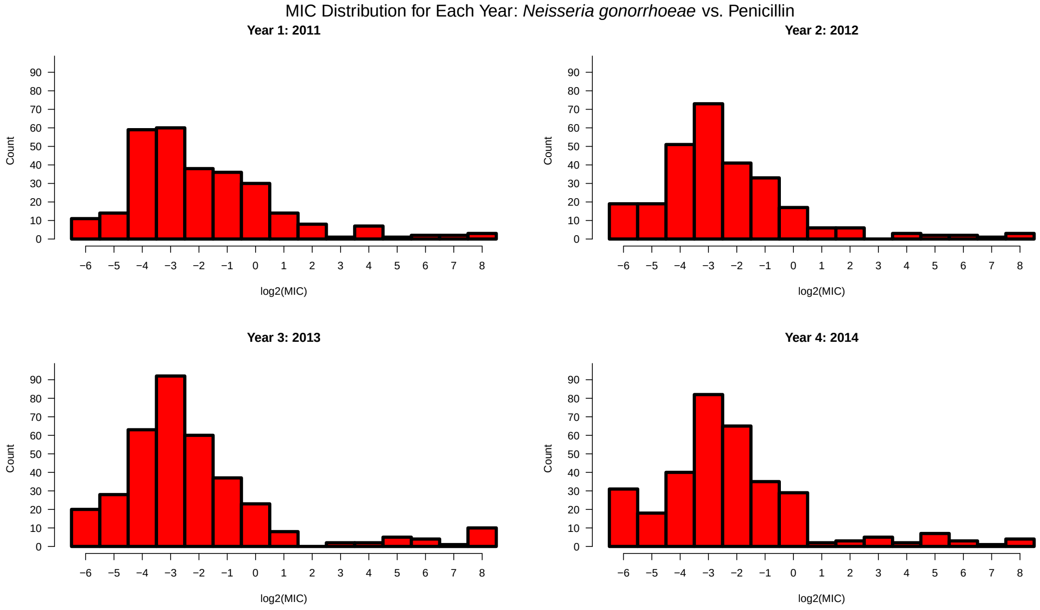


Figure 4.10. Each year's MIC distribution is shown as a histogram.

Figure 4.11 summarizes the resulting fits. The error bars shown for the BayesACME method represent the 95% credible intervals. Using the combined goodness-of-fit statistic for the the four years summed together for the multinomial model, the distribution of the Normal model with a median of 76.250 is smaller than that of the median of 77.399 for test statistic distribution for the log2gamma. The test statistic is large because the data sets for 2011 and 2014 are a bit unusual. The 2011 data set has an unusual mode and 2014 data set has an odd WT left-tail. The estimates from the three methods under the two distributions are displayed in Table 4.7.

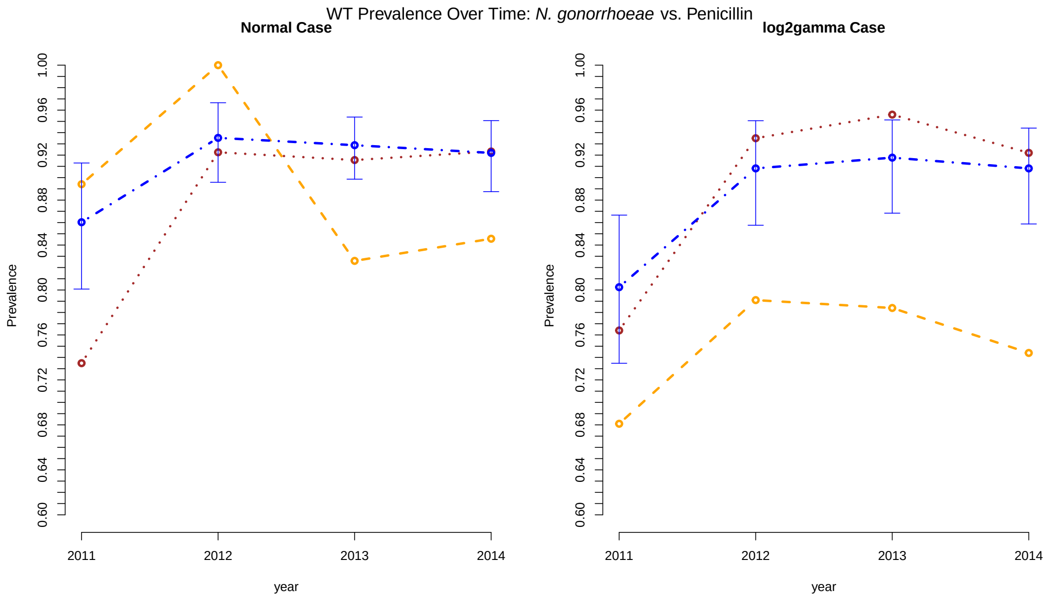


Figure 4.11. This plot is the estimate of prevalence versus time for the three extended methods from 2011 to 2014. The left column is the Normal case and the right column is the log2gamma case.

Because we assume the same WT distribution over time, the model adjusts for the unique pattern in the data of 2011 by lowering its prevalence estimate. Except for that year, the estimates follow a steady trend. The patterns are very consistent in the Normal case across the three methods. The extension of TURNM is more erratic in the Normal case. At this method of estimation uses the mode plus one bin as the subset for estimation it may acquire sensitivity to how bins to the left and right of the left-most mode fluctuate from year to year. In both the Normal and log2gamma cases, BayesACME and JASP tend to agree within the presented credible intervals for the most part. By looking back to Figure 4.10, the WT distribution for year 2011 has a different form. In contrast to the other years, the WT mode is rounder. It may explain why the estimate of WT prevalence is relatively different.

There are results from the state of Maryland from 2009 and 2010, where the WT prevalence is 97% and 94%, respectively [Ghanem and Razeq, 2012]. Table 4.7 below provides the estimates from the two different models. As the geography and time period is slightly different between the state of Maryland estimates and the estimates from BayesACME, they do appear to be comparable except for year 2011.

Table 4.7.
Results from the *N. gonorrhoeae* vs. Penicillin from a Single Dutch Lab from 2011 to 2014

Normal	$\boldsymbol{\mu}$	$\boldsymbol{\sigma_{obs}}$	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	$\boldsymbol{\pi_4}$	**ECOFF**	**Bins**
TURNM	−3.747	1.599	0.894	1.000	0.826	0.846	1.194	(4,4,5,5)
JASP	−3.596	1.441	0.735	0.923	0.916	0.923	0.857	(5,7,7,7)
BayesACME	−3.274	1.749	0.883	0.943	0.932	0.927	2.131	All
log2gamma	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	$\boldsymbol{\pi_4}$	**ECOFF**	**Bins**
TURNM	1.564	15.995	0.681	0.791	0.784	0.744	−0.950	(5,5,6,6)
JASP	1.542	14.000	0.764	0.935	0.956	0.922	0.901	(5,5,5,6)
BayesACME	1.606	12.143	0.803	0.908	0.918	0.908	1.040	All

4.7 Conclusion

This chapter extends the proposed model in Chapter 3 for the purpose of longitudinal surveillance of AMR. Additionally, it extends the two subset methods for comparison. The simulation study shows that the multiyear extension shows the general pattern that BayesACME is better at handling contamination and more precise by modelling the entire mixture. Both the Normal and log2gamma case attest to the improvement in precision. However, in the log2gamma case at the fast rate of decline in WT prevalence and especially at high contamination, the subset methods can become more accurate because BayesACME struggles to estimate the WT right tail. In terms of capturing the trend, BayesACME has a dominant performance with reporting the lowest MSE for the slope at all three contamination levels in the Normal and log2gamma cases.

In future work, a joint prior can be placed on the prevalence. Biologically, the true prevalence values should only decrease when time, geography, and source are the same. Previously, we considered an ordered constraint prior, and found that it resulted in over-estimating the decline in WT prevalence at a slow rate. Perhaps experimenting with a prior that imposes a similar, but a weaker condition such as $\mathbf{E}[\pi_1] > ... > \mathbf{E}[\pi_T]$ is suited for future work.

5. CONCLUDING REMARKS

In this dissertation, we proposed a novel AMR monitoring method and show across most settings that it outperforms current single-year methods in both accuracy and precision. This improvement is due primarily to two reasons. First, we utilize the entire collection of bin counts rather than just using a subset of them to estimate the prevalence and WT parameters. This allows us to incorporate information on the right tail of the WT distribution that the subset methods do not necessarily incorporate. Second, our approach is the only method whose model accounts for measurement error in the assay. While other methods focus on the observed distribution of assay results, we are able to partition the observed distribution into a latent assay distribution and Normal measurement error distribution. This grants us the ability to

1. Describe the observed assay distribution properly when the latent WT distribution is non-Normal
2. Incorporate biological knowledge into the prior of the NWT distribution. Thereby allowing this method to handle cases with various degrees of contamination.

As discussed in Chapter 1, the log2gamma distribution convolved with a Normal measurement error (ME) distribution is no longer a log2gamma. Thus, methods that assume the observed assay distribution is log2gamma are prone to poor estimation of the true underlying log2gamma distribution and thus prone to biases when estimating functions of the distribution such as the observed ECOFF. Outside of assuming the WT distribution is Normal, accounting for ME is essential for proper estimation.

In addition, by partitioning noise from the latent distribution, we incorporate biological information into the relationship between the WT and NWT distribution. In other words, we are able to separate overlap in these two distributions that is truly biological from overlap that is created due to ME. We currently use the concept of a latent ECOFF to set these limits of true biological overlap in the distribution, but our method is flexible enough to consider an alternative use of this information.

Using information to constrain the NWT distribution and imposing little prior information on the WT distribution is the complete opposite of the other Bayesian semiparametric approach in the literature. It takes a more empirical approach using the results from fitting a cruder subset model to set an informative prior for the WT distribution. Our current simulations suggest our approach outperforms this alternative except when the amount of overlap in the observed results is high. The better performance at high contamination may be in part due to the fact their prior is more informative than our current restrictions on the NWT. Further research is needed on this.

In Chapter 4, we considered extending our approach to handle AMR monitoring to across years under the assumption that the WT distribution is static over time. If reasonable, this assumption enables the sharing of information across years in the estimation of the WT distribution and resulting in more precise estimates of prevalence. For comparison purposes, we not only extended our model, but also extended the two subset methods (i.e., TURNM and JASP) to handle multiple years. We again show that our method generally outperforms these other two in terms of precision and accuracy of the trend, especially when the overlap in the WT distribution and NWT distribution is high.

5.1 Limitations

The collections of observed assay values typically come from multiple clinics, performed by various technicians, and across numerous days. Thus, the observed measurement error is not due just to the assay itself but is also confounded with other sources of variation. Our current prior of σ_δ^2 is based on the within-lab variabilities observed in QC studies. As a result, our prior may cover a range of variances that is much smaller than expected in practice.

While properly estimating this variance does not impact our estimation of the observed ECOFF in the Normal case, this variability will impact our estimates of the latent WT and NWT distributions. Not only in terms of partitioning away error, but also because the overlap limits of the two distributions we set. We have yet to do a robust analysis where we consider ME outside the range of the prior. As shown in the simulations, there is considerable sensitivity to the prior σ_δ^2 in the non-Normal WT distributions. If the prior is not properly centered, the estimation of observed ECOFF is biased for non-Normal data.

Along those same lines, our current heuristic for setting a lower limit on the means of the NWT components and an upper limit on the standard deviations currently relies on a decent estimate of the ME variance. While we are confident in the reasoning used to construct these limits knowing the truth, we are unclear if our current estimation approach is the best one for setting these limits.

Although BayesACME typically outperformed the current methods, JASPB performed the best, especially as the degree of contamination increased. While an approach that separates the WT and NWT distributions based on biological arguments is desirable, we have to admit that taking an empirical approach to set a prior for $\boldsymbol{\theta_{WT}}$ also works quite well.

Fortunately, our approach handles the log2gamma distribution and can easily be adapted to incorporate the prior used by JASPB, should its incorporation or a hybrid of our two approaches be viable. This issue only arises in the high contamination setting, but warrants further investigation.

5.2 Future Work

Inspired by the popularity of the Excel Macro ECOFFfinder [EUCAST, 2020], we plan to make our procedure available as an R Shiny app and as an R package. By offering new statistical software, we hope to start a more vibrant discussion over

WT distribution specifications, the inclusion of measurement error, and uncertainty quantification.

In addition, the algorithms proposed are MCMC schemes. There is room to improve the computationally efficiency. For example, the current implementation is written entirely in R, but some of the MCMC steps could easily be sped up using a compiled language such as C or Fortran. It is also possible that by switching from a random-walk MCMC to a more sophisticated mechanism that introduces gradient information may improve things.

5.2.1 Explicitly Including Between-Lab Variability

As mentioned earlier, QC data suggests there is inherent variability in the MIC assay both between and within-labs. Currently, these MIC collections do not specify the lab, but if they did, we could adjust the underlying latent MIC distribution for this source of variability.

Denoting the between-lab error as $\epsilon \sim N(0, \sigma_\epsilon)$, we model the observed MIC using

$$Y = \lceil X + \delta + \epsilon \rceil$$

Using the same QC data as before, we can put a prior on σ_ϵ based on the between-lab variability. This inclusion would allow us to adjust properly for an unequal number of samples from labs that we cannot do now. Of course, this only makes sense if each lab-effect remains constant over the collection time, something that is not yet clear.

5.2.2 Allowable Overlap Between the NWT and WT Distributions

Currently, overlap is managed by controlling the means and standard deviations of the NWT Normal components. Alternative approaches could consider using a single measure to quantify the overlap such as the Bhattarcharyya coefficient. The advantage of this approach is a restriction placed on the entire mixture rather than just the means and standard deviations of the components. As dealing with a latent continuous distribution and using the Bhattarcharyya coefficient to quantify contamination are first proposed here, an acceptable and realistic upper bound placed on the Bhattarcharyya coefficient needs to be determined for real data sets. This modification removes the requirement for pre-specification of parameters and relaxes assumptions about the base distribution in the DPMM.

Another idea is to determine empirical priors for $\boldsymbol{\theta_{WT}}$ and π like JASPB and place little to no restrictions on the DPMM. Likely, there would need to be some requirement on the DPMM to ensure the NWT remains to the right of the WT distribution.

5.2.3 WT Model Comparison

The χ^2 goodness-of-fit test requires either the visual assessment of two different distributions or selecting a heuristic from the χ^2 distribution to determine the better model fit. This may make cases where the two models have very comparable fit difficult to distinguish as the difference between the models is the latent WT distribution. One option is to use an alternative that creates scores for comparison.

The widely applicable information criteria also known as the Watanabe-Akaike Information Criteria (WAIC) addresses this limitation [Watanabe, 2010, 2013]. The WAIC is a generalization of the AIC suited for Bayesian hierarchical models and is asymptotically comparable to Bayesian Leave-One-Out Cross Validation (LOO-CV).

The WAIC is a Bayesian approach for estimating the out-of-sample expectation using a computed log pointwise posterior predictive density [Gelman et al., 2013]. Analogous to information criteria such as the AIC that introduce a penalty for the number of parameters, the WAIC introduces a correction for the effective number of parameters. One must take a sample of n "new" observations denoted $\tilde{Y}_1, ..., \tilde{Y}_n$. Then using S posterior samples, the expected log pointwise predictive density (ellpd) for the “new” data set can be estimated with the log pointwise predictive density (llpd):

$$\sum_{i=1}^{n} \log(\frac{1}{S} \sum_{s=1}^{S} P(\tilde{Y}_i \mid \tilde{\boldsymbol{\theta}}^s))$$

where $\tilde{\boldsymbol{\theta}}^s$ is the s^{th} posterior draw of $\tilde{\boldsymbol{\theta}}$ and P denotes the multinomial probability. To calculate the correction for the effective number of parameters denoted p_{WAIC}, Gelman et al. [2013] recommend summing over the sample variances of the log-likelihood for the S posterior draws for the n “new” data. The WAIC is calculated as $-2llpd + 2p_{WAIC}$. Like other information criteria, the lower the value of the WAIC; the higher the predictive accuracy.

The downside to this method is the selection of the n “new” observations is not clear. In the future, it may be worthwhile exploring if the WAIC can be adapted or modified to focus solely on the latent true WT distribution with the hopes of avoiding issues of low bin counts associated with the multinomial model.

5.2.4 Time Dependent Dirichlet Process Mixture Models

One relatively new development is modelling the NWT distribution jointly with the WT distribution semiparametrically [Jaspers et al., 2016a,b]. Although the NWT may be viewed as a nuisance factor, there is information in this distribution. In the multiyear analysis, we currently assume the WT distribution remains fixed with no restrictions on how the NWT changes over time. The question of how (and even how much) the NWT distribution changes for an arbitrary “drug/bug” combination is an open question [Mouton et al., 2018]. Certainly, labs may want to explore and highlight any changes or trends in this subpopulation when submitting results.

One possibility is to use a Time Dependent Dirichlet Process Mixture Model (DDPMM). The DDPMM generalizes the DPMM model by eliminating the key assumption of exchangeability with the data points and their labeling (a necessary requirement for the Chinese Restaurant Process). The DDPMM generalizes the DPMM by including birth, death, and transition processes into the clusters for the model [Campbell et al., 2013, Lin et al., 2010, MacEachern, 2000]. By doing this, it allows for the parameters of the components to be static over time, but the weights of the mixture to change over time [Fox and Jordan, 2013]. This is a potential generalization to the assumptions of Chapter 4, where the mean and standard deviation of the WT component are relatively static over time, but the mixture weight π changes. Now all the components are relatively fixed, but all the weights are changing over time.

5.2.5 Considering a Hypersusceptible Subpopulation

The methodology discussed in this dissertation is inappropriate for addressing a hypersusceptible subpopulation. It is assumed to not exist. In fairness, its presence is rare. Its inclusion requires adding a component to the mixture model. Denote the density as f_{hyp} with parameter set $\boldsymbol{\theta_{hyp}}$ and weight ξ. Then the overall mixture model becomes:

$$f(\mathbf{X}^*) = \xi f_{hyp}(\mathbf{X}^* \mid \boldsymbol{\theta_{hyp}}) + \pi f_{WT}(\mathbf{X}^* \mid \boldsymbol{\theta_{WT}}) + (1 - \xi - \pi) f_{NWT}(\mathbf{X}^* \mid \boldsymbol{\theta_{NWT}})$$

for the observed data $\mathbf{Y}$.

Currently, not much can be said about the hypersusceptible subpopulation other than it is to the left of the WT population as a distinct population. As it is a likely product of evolution like the NWT distribution, a form of nonparametric estimation is likely required. Currently, it is not clear the best method of estimation and consideration must be made to the possibility that this subpopulation may only span a single bin.

If this subpopulation is present, it creates major issues for WT distribution and prevalence estimation, as the key assumption is that the WT distribution is the left-most subpopulation. It may also contaminate the left WT tail. In fact, it brings up a return to the other value of the ECOFF namely the 0.1^{th} percentile of the WT distribution [Turnidge et al., 2006]. Typically, this value was not of any particular interest as the contamination only occurred from the right, except for excluding possible lab errors. It is possible a left-ward hypersusceptible mode elevates its relevance.

5.2.6 Potential Changes to the MIC Assay

In studying the MIC assay it becomes readily apparent that little has changed with the procedure and its use for over 50 years. As a major challenge with this analysis is the loss of information from censoring, I would like to study the impact of two MIC alterations:

1. What if we consider more than 12 concentrations? Clearly narrower bins in the region of the observed ECOFF would help analysis. If there were fold-level changes in concentration over the same region, what would be the improvement?

2. If more concentrations were not feasible, what would be the gain if each isolate were tested twice? This would not only help in estimating the underling latent MIC, but would help in estimating the within-assay variability.

If the investigation from these questions proves useful, then approaching microbiologists and monitoring agencies to investigate meaningful reforms.

As laboratories collect the data, subsequent methodological considerations should be made. Ideally, laboratories should supply more than the assay results. They should try to provide (or make available) information on the broth, the geographic location where the bacterium was collected, the specific date the experiment was done, and the technician(s) [Mouton et al., 2018]. This list is not exhaustive. It allows for the best use of the results.

REFERENCES

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

D. H. Annis and B. A. Craig. The effect of interlaboratory variability on antimicrobial susceptibility determination. *Diagnostic microbiology and infectious disease*, 53 (1):61–64, 2005a.

D. H. Annis and B. A. Craig. Statistical properties and inference of the antimicrobial mic test. *Statistics in medicine*, 24(23):3631–3644, 2005b.

Y. F. Atchadé, J. S. Rosenthal, et al. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

J. O. Berger and L. R. Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996a.

J. O. Berger and L. R. Pericchi. On the justification of default and intrinsic bayes factors. In *Modelling and Prediction Honoring Seymour Geisser*, pages 276–293. Springer, 1996b.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

T. Campbell, M. Liu, B. Kulis, J. P. How, and L. Carin. Dynamic clustering via asymptotics of the dependent dirichlet process mixture. In *Advances in Neural Information Processing Systems*, pages 449–457, 2013.

B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3): 473–484, 1995.

CDC. Antibiotic resistance laboratory network. `https://www.cdc.gov/drugresistance/solutions-initiative/ar-lab-network.html`, 2020. Last Revised: 2020-02-10.

B. A. Craig. Drug dilution vs drug diffusion: Calibrating the two susceptibility tests. Technical report, Technical Report 99-15, Purdue University, Department of Statistics, 01 2010, 1999.

B. A. Craig. Modeling approach to diameter breakpoint determination. *Diagnostic microbiology and infectious disease*, 36(3):193–202, 2000.

A. Delaigle, P. Hall, et al. Methodology for non-parametric deconvolution when the error distribution is unknown. *JR Stat. Soc. Ser. B. Stat. Methodol*, 78(1):231–252, 2016.

G. DePalma and B. A. Craig. Bayesian monotonic errors-in-variables models with applications to pathogen susceptibility testing. *Statistics in medicine*, 37(3):487–502, 2018.

B. Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

EUCAST. The ecofffinder program. *MIC distributions and epidemiological cut-off value (ECOFF) setting*, SOP 10.0, 2017.

EUCAST. The ecofffinder program. *ECOFFfinder2020*, 2020.

E. B. Fox and M. I. Jordan. Mixed membership models for time series. *arXiv preprint arXiv:1309.3533*, 2013.

D. H. Freedman. The death of antibiotics: We're running out of effective drugs to fight off an army of superbugs. *Newsweek Magazine*, May 2019.

A. S. Fuhrmeister and R. N. Jones. The importance of antimicrobial resistance monitoring worldwide and the origins of sentry antimicrobial surveillance program, 2019.

A. Gelman. How to think about "identifiability" in bayesian inference? Blog: Statistical Modeling, Causal Inference, and Social Science, February 2014. `https://statmodeling.stat.columbia.edu/2014/02/12/think-identifiability-bayesian-inference/`.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

K. Ghanem and J. H. Razeq. Sex and the superbug: Next steps in dealing with multi-drug resistant gonorrhea in maryland, 2012.

D. Görür and C. E. Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

C. Grazian. Estimating mic distributions and cutoffs through mixture models: an application to establish m. tuberculosis resistance. *bioRxiv*, page 643429, 2019.

H. Haario, E. Saksman, J. Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

L. B. Harrison, R. C. Fowler, B. Abdalhamid, A. Selmecki, and N. D. Hanson. lptg contributes to changes in membrane permeability and the emergence of multidrug hypersusceptibility in a cystic fibrosis isolate of pseudomonas aeruginosa. *MicrobiologyOpen*, 8(11):e844, 2019.

S. Jaspers, M. Aerts, G. Verbeke, and P.-A. Beloeil. Estimation of the wild-type minimum inhibitory concentration value distribution. *Statistics in medicine*, 33(2): 289–303, 2014a.

S. Jaspers, M. Aerts, G. Verbeke, and P.-A. Beloeil. A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Computational Statistics & Data Analysis*, 71:30–42, 2014b.

S. Jaspers, P. Lambert, M. Aerts, et al. A bayesian approach to the semiparametric estimation of a minimum inhibitory concentration distribution. *The Annals of Applied Statistics*, 10(2):906–924, 2016a.

S. Jaspers, G. Verbeke, D. Böhning, and M. Aerts. Application of the vertex exchange method to estimate a semi-parametric mixture model for the mic density of escherichia coli isolates tested for susceptibility against ampicillin. *Biostatistics*, 17 (1):94–107, 2016b.

H. Jeffreys. *The theory of probability*. OUP Oxford, 1998.

T. Jinks. Why is it so difficult to develop novel antibiotics. `https://www.bbc.com/news/health-41693229`, 2017. Accessed: 2020-06-05.

V. E. Johnson et al. A bayesian $\chi 2$ test for goodness-of-fit. *The Annals of Statistics*, 32(6):2361–2384, 2004.

A. Jullion and P. Lambert. Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian p-splines models. *Computational statistics & data analysis*, 51(5):2542–2558, 2007.

G. Kahlmeter, D. F. Brown, F. W. Goldstein, A. P. MacGowan, J. W. Mouton, A. Österlund, A. Rodloff, M. Steinbakk, P. Urbaskova, and A. Vatopoulos. European harmonization of mic breakpoints for antimicrobial susceptibility testing of bacteria. *Journal of antimicrobial chemotherapy*, 52(2):145–148, 2003.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

P. Lambert and P. H. Eilers. Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53(4):1388–1399, 2009.

S. Lang and A. Brezger. Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004.

M.-L. T. Lee and G. Whitmore. Statistical inference for serial dilution assay data. *Biometrics*, 55(4):1215–1220, 1999.

F. Liang, C. Liu, and R. Carroll. *Advanced Markov chain Monte Carlo methods: learning from past samples*, volume 714. John Wiley & Sons, 2011.

D. Lin, E. Grimson, and J. W. Fisher. Construction of dependent dirichlet processes based on poisson processes. In *Advances in neural information processing systems*, pages 1396–1404, 2010.

J. Liu, O. Gefen, I. Ronin, M. Bar-Meir, and N. Q. Balaban. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science*, 367(6474): 200–204, 2020.

S. N. MacEachern. Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, pages 1–40, 2000.

J. W. Mouton, A. E. Muller, R. Canton, C. G. Giske, G. Kahlmeter, and J. Turnidge. Mic-based dose adjustment: facts and fables. *Journal of Antimicrobial Chemotherapy*, 73(3):564–568, 2018.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

M. A. Newton, B. Mau, and B. Larget. Markov chain monte carlo for the bayesian analysis of evolutionary trees from aligned molecular sequences. *Lecture Notes-Monograph Series*, pages 143–162, 1999.

D. Nield. Scientists just found 8000 cocktails for fighting antibiotic resistance. www.sciencealert.com/8000-new-drug-combinations-could-fight-antibiotic-resistance, 2018. Accessed: 2020-07-19.

A. O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.

X. Qi. *Nonparametric calibration of two common susceptibility tests using interval-censored data with measurement error.* PhD thesis, Purdue University, 2008.

R. Roemhild, M. Linkevicius, and D. I. Andersson. Molecular mechanisms of collateral sensitivity to the antibiotic nitrofurantoin. *PLoS biology*, 18(1):e3000612, 2020.

G. J. Ross and D. Markwick. dirichletprocess: An r package for fitting complex bayesian nonparametric models, 2018.

G. J. Ross, D. Markwick, K. Mulder, and G. Sighinolfi. dirichletprocess: Build dirichlet process objects for bayesian modelling, version 0.4.0. https://cran.r-project.org/web/packages/dirichletprocess/index.html, 2020.

Seleem. Episode 4: With your health in mind. https://www.purdue.edu/newsroom/podcast/2020/with-your-health-in-mind.html, 2020. Accessed: 2020-03-10.

J. Turnidge, G. Kahlmeter, and G. Kronvall. Statistical characterisation of bacterial wild-type mic value distributions and the determination of epidemiological cut-off values. *Clinical Microbiology and Infection*, 12(5):418–425, 2006.

J. D. Turnidge. Susceptibility test methods: General considerations. In J. H. Jorgensen and M. A. Pfaller, editors, *Manual of Clinical Microbiology, 11th Ed.*, chapter 70, pages 1246–52. ASM Press, Oxford, 2015.

J. van de Kassteele, M. G. van Santen-Verheuvel, F. D. Koedijk, A. P. van Dam, M. A. van der Sande, and A. J. de Neeling. New statistical technique for analyzing mic-based susceptibility data. *Antimicrobial agents and chemotherapy*, 56(3):1557–1563, 2012.

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.

S. Watanabe. Waic and wbic are information criteria for singular statistical model evaluation. In *Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering*, pages 90–94, 2013.

K. Weintraub. Bacteria "tolerant" of one antibiotic are more likely to develop resistance. *Scientific American*, Jan 2020.

M. West. *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03, 1992.

M. Zhang, C. Wang, and A. O'Connor. A hierarchical bayesian latent class mixture model with censorship for detection of linear temporal changes in antibiotic resistance. *PloS one*, 15(1):e0220427, 2020.

X. K. Zhou, M. A. Clyde, J. Garrett, V. Lourdes, M. O'Connell, G. Parmigiani, D. J. Turner, T. Wiles, et al. Statistical methods for automated drug susceptibility testing: Bayesian minimum inhibitory concentration prediction from growth curves. *The Annals of applied statistics*, 3(2):710–730, 2009.

A. LATENT WT DISTRIBUTIONS

Normal Distribution

For a random variable, $X \sim N(\mu, \sigma)$, the mean is $E[X] = \mu$, and the variance is $Var(X) = \sigma^2$ where μ is any real value and $\sigma > 0$. A special case is the standard Normal where $\mu = 0$ and $\sigma = 1$.

The Normal distribution is special in that linear combinations of independent Normal random variables are also Normally distributed. The mean of the linear combination is the linear combination of the means of each of the Normal random variables. Assuming mutual independence among the Normal components, the variance is the linear combination of the Normals random variables. Specifically, for $t = 1, ..., n$ with $X_t \overset{indep.}{\sim} N(\mu_t, \sigma_t)$, then $\sum_{t=1}^{n} a_t X_t \sim N(\sum_{t=1}^{n} a_t \mu_t, \sqrt{\sum_{t=1}^{n} a_t^2 \sigma_t{}^2})$.

We denote the probability density function (pdf) for the standard Normal as $\phi(\cdot)$. In the pdf, π denotes the mathematical constant (i.e., $\pi = 3.14159...$) and not WT prevalence. For an arbitrary random variable X with the Normal distribution, the pdf is

$$f_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

The cumulative distribution function (cdf) is denoted as $\Phi(\frac{x-\mu}{\sigma})$. The cdf does not have a closed-form. The density is bell-shaped and symmetric so the median and mode are also equal to the mean. The parameter μ shifts the density and the parameter σ scales the density.

Figure A.1 contains the Normal density for different values of μ and σ.

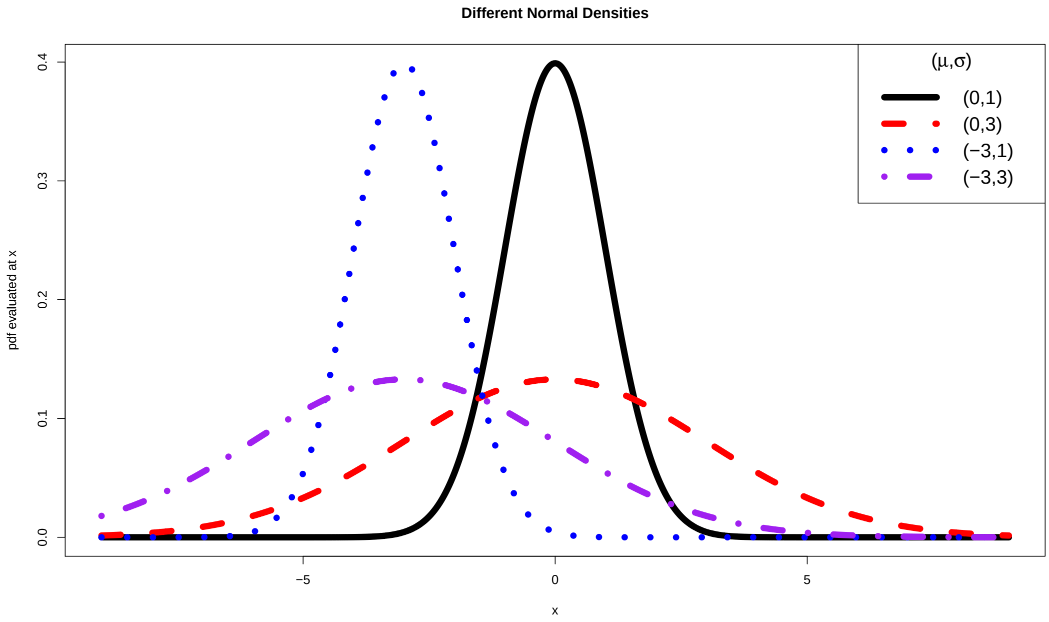


Figure A.1. The pdf of the Normal distribution for different parameters values.

Log2gamma Distribution

The log2gamma distribution is the logarithm base 2 transformation of a gamma random variable with shape parameter α and rate parameter β. Both α and β must be positive. Its probability density function (pdf) is

$$f_X(x) = \ln(2)\frac{\beta^\alpha}{\Gamma(\alpha)}2^{\alpha x}\exp[-\beta 2^x]$$

where $\Gamma(\cdot)$ is the gamma function.

Denoting the digamma function as $\psi(\cdot)$ and the trigamma function as $\psi^{(1)}(\cdot)$, the mean is

$$E[X] = \frac{\psi(\alpha) - \ln(\beta)}{\ln(2)}$$

and the variance is

$$Var[X] = \frac{\psi^{(1)}(\alpha)}{\ln(2)^2}$$

This distribution is skewed to the left with the mode:

$$mode(X) = \log_2(\frac{\alpha}{\beta})$$

Figure A.2 contains several densities of the log2gamma distribution. The parameter α and the spread of the distribution are inversely related. By decreasing β and keeping α fixed, the density shifts to the right.

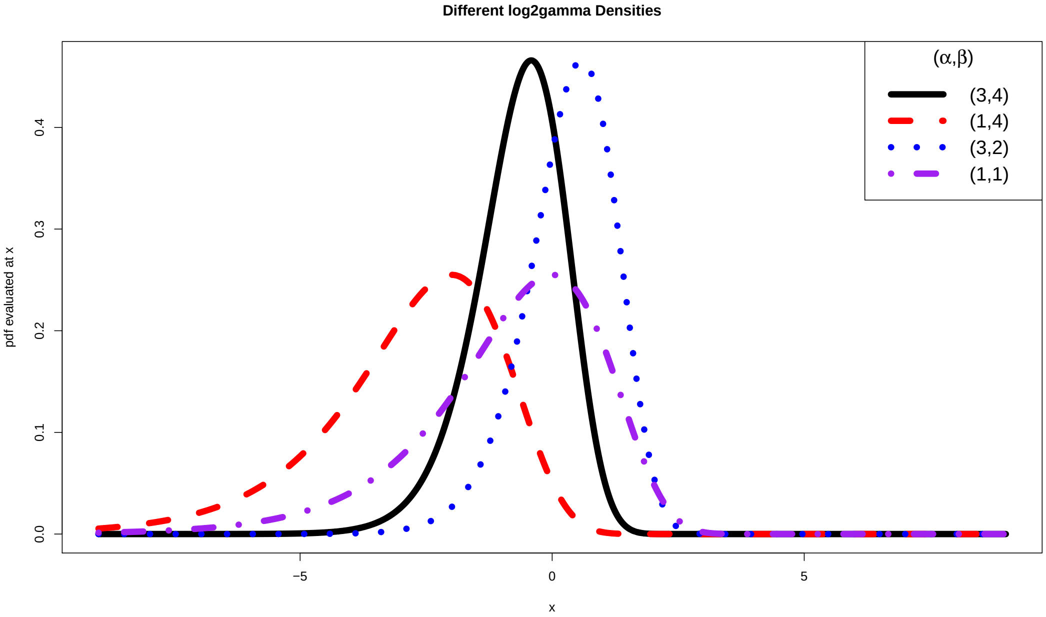


Figure A.2. The pdf of the log2gamma distribution with different parameters values. The parameter α solely determines the spread. Both α and β determine the mean and mode.

B. DERIVATION OF MAXIMUM LIKELIHOOD ESTIMATES FOR JASPERS ET AL. SUBSET METHOD

The method of Jaspers et al. [2014a] assumes the observed results from the MIC assay form a multinomial. For the WT component with bin j with corresponding with concentration C_j. The first K bins on the left are determined parametrically with some distribution F_{WT} with parameter set $\boldsymbol{\theta_{WT}}$

$$\tilde{p_j} = \begin{cases} F_{WT}(C_j; \boldsymbol{\theta_{WT}}) & j = 1 \\ F_{WT}(C_j; \boldsymbol{\theta_{WT}}) - F_{WT}(C_{j-1}; \boldsymbol{\theta_{WT}}) & j = 2, 3, ..., K \end{cases} \tag{B.1}$$

The definition for $\tilde{p_j}$ is different in this appendix than in Chapter 2 by excluding the prevalence, π. This discrepancy is intentional. The definition in Chapter 2 that includes π enables brevity. In this appendix excluding π allows for clarity in the derivation.

We seek to make rigorous the maximum likelihood estimation for the method of Jaspers et al. [2014a]. We have a multinomial where the first K bins on the left are defined parametrically. The remaining $J - K$ are not dependent on $\boldsymbol{\theta_{WT}}$. Using the following log-likelihood, $l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)$ with Lagrange multiplier λ to enforce the constraint that bin probabilities sum to 1. We use $\tilde{p_j}$ to denote the bin probabilities for the WT component for the first K bins and p_j to denote the bin probabilities for the NWT bins. Each bin j has count m_j.

The log-likelihood is

$$l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda) = \sum_{j=1}^{K} m_j \log(\pi \tilde{p_j}) + \sum_{j=K+1}^{J} m_j \log(p_j) + \lambda(1 - \pi \sum_{j=1}^{K} \tilde{p_j} - \sum_{j=K+1}^{J} p_j) \tag{B.2}$$

$$= \log(\pi)\sum_{j=1}^{K} m_j + \sum_{j=1}^{K} m_j \log(\tilde{p_j}) + \sum_{j=K+1}^{J} m_j \log(p_j) + \lambda(1 - \pi\sum_{j=1}^{K}\tilde{p_j} - \sum_{j=K+1}^{J} p_j) \quad \text{(B.3)}$$

In general, when $j \geq 2, l_j = u_j - 1$. Notice, we have a finite telescopic series:

$$\sum_{j=1}^{K}\tilde{p_j} = F_{WT}(C_1;\boldsymbol{\theta_{WT}}) + (F_{WT}(C_2;\boldsymbol{\theta_{WT}}) - F_{WT}(C_1;\boldsymbol{\theta_{WT}})) + ...$$
$$+ (F_{WT}(C_K;\boldsymbol{\theta_{WT}}) - F_{WT}(C_{K-1};\boldsymbol{\theta_{WT}})) = F_{WT}(C_K;\boldsymbol{\theta_{WT}}) \quad \text{(B.4)}$$

Then the log-likelihood can be restated by substituting $F_{WT}(C_K;\boldsymbol{\theta_{WT}}) = \sum_{j=1}^{K}\tilde{p_j}$:

$$l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda) =$$
$$\log(\pi)\sum_{j=1}^{K} m_j + \sum_{j=1}^{K} m_j \log(\tilde{p_j}) + \sum_{j=K+1}^{J} m_j \log(p_j) + \lambda(1 - \pi \cdot F_{WT}(C_K;\boldsymbol{\theta_{WT}}) - \sum_{j=K+1}^{J} p_j) \quad \text{(B.5)}$$

Now we must take the first derivative to each parameter in the log-likelihood and set each each equation to 0. First, we look at the NWT components, p_j, for $j = K+1, ..., J$,

$$\frac{\partial l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)}{\partial p_j} = \frac{m_j}{p_j} - \lambda = 0 \quad \text{(B.6)}$$

Then $p_j = \frac{m_j}{\lambda}$. Next, an equation for prevalence follows

$$\frac{\partial l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)}{\partial \pi} = \frac{1}{\pi}\sum_{j=1}^{K} m_j - \lambda\sum_{j=1}^{K}\tilde{p_j} = \frac{1}{\pi}\sum_{j=1}^{K} m_j - \lambda F_{WT}(C_K, \boldsymbol{\theta_{WT}}) = 0. \quad \text{(B.7)}$$

Then $\pi = \frac{\sum_{j=1}^{K} m_j/\lambda}{F_{WT}(C_K;\boldsymbol{\theta_{WT}})}$. Now an equation for λ is required:

$$\frac{\partial l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)}{\partial \lambda}$$
$$= 1 - \pi \sum_{j=1}^{K} \tilde{p_j} - \sum_{j=K+1}^{J} p_j = 1 - \pi F_{WT}(C_K, \boldsymbol{\theta_{WT}}) - \frac{\sum_{j=K+1}^{J} m_j}{\lambda} = 0 \quad \text{(B.8)}$$

Then

$$0 = 1 - (\frac{\sum_{j=1}^{K} m_j/\lambda}{F_{WT}(C_K;\boldsymbol{\theta_{WT}})}) F_{WT}(C_K, \boldsymbol{\theta_{WT}}) - \frac{\sum_{j=K+1}^{J} m_j}{\lambda} \quad \text{(B.9)}$$

Then

$$0 = 1 - \frac{\sum_{j=1}^{J} m_j}{\lambda} = 1 - \frac{N_{tot}}{\lambda} \quad \text{(B.10)}$$

There is an unique solution: $\lambda = N_{tot}$. Then for $j = K+1, ..., J, p_j = \frac{m_j}{N_{tot}}$. Then

$$\frac{\partial l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)}{\partial \boldsymbol{\theta_{WT}}} =$$
$$\frac{\partial}{\partial \boldsymbol{\theta_{WT}}}[\sum_{j=1}^{K} m_j \log(\pi) + \sum_{j=1}^{K} m_j[\log(F_{WT}(C_j;\boldsymbol{\theta_{WT}})) - F_{WT}(C_{j-1};\boldsymbol{\theta_{WT}}))]$$
$$- \lambda\pi F_{WT}(C_K;\boldsymbol{\theta_{WT}})] = 0 \quad \text{(B.11)}$$

Then

$$\frac{\partial l(\boldsymbol{\theta_{WT}}, \pi, p_{K+1}, ..., p_J, \lambda)}{\partial \boldsymbol{\theta_{WT}}} =$$
$$\frac{\partial}{\partial \boldsymbol{\theta_{WT}}}[\sum_{j=1}^{K} m_j[\log(F_{WT}(C_j;\boldsymbol{\theta_{WT}})) - F_{WT}(C_{j-1};\boldsymbol{\theta_{WT}}))] - \lambda\pi F_{WT}(C_K;\boldsymbol{\theta_{WT}})] = 0$$
$$\text{(B.12)}$$

By substitution, $\pi = \frac{\sum_{j=1}^{K} m_j/\lambda}{F_{WT}(C_K;\boldsymbol{\theta_{WT}})}$.

$$\frac{\partial}{\partial \boldsymbol{\theta_{WT}}}[\sum_{j=1}^{K} m_j[\log(F_{WT}(C_j;\boldsymbol{\theta_{WT}})) - F_{WT}(C_{j-1};\boldsymbol{\theta_{WT}}))] - \sum_{j=1}^{K} m_j = 0 \quad \text{(B.13)}$$

Then define $C_0 = -\infty$,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{WT}} = \arg\max_{\boldsymbol{\theta}_{\boldsymbol{WT}}} \sum_{j=1}^{K} m_j[\log(F_{WT}(C_j;\boldsymbol{\theta}_{\boldsymbol{WT}}) - F_{WT}(C_{j-1};\boldsymbol{\theta}_{\boldsymbol{WT}}))]$$

Notice these solutions do not depend on either π or λ. In contrast, the estimator of π depends on both $\boldsymbol{\theta}_{\boldsymbol{WT}}$ and λ. Using the determined solutions we can state:

$$\hat{\pi} = \frac{\sum_{j=1}^{K} m_j/N_{tot}}{F_{WT}(C_K;\hat{\boldsymbol{\theta}}_{\boldsymbol{WT}})}$$

C. SINGLE-YEAR COMPUTATION

In this appendix, the MCMC schemes for the Normal and log2gamma distributions are each detailed. Each model is treated in a section. The R code for the DPMM is a modification from the *Dirichletprocess* package to accommodate the truncation of the base distribution [Ross et al., 2020].

Computation: Normal WT Distribution for Single-Year Data

Now for computational ease, three latent vectors are introduced. The first is the latent continuous values, $\mathbf{X}$. The second is an indicator for whether an observation is a wild-type or not, $\mathbf{c}$. The third is for values of the measurement error, $\boldsymbol{\delta}$. A Metropolis-within-Gibbs algorithm is used.

For observation i, define

$$\begin{cases} c_i = & 1 \quad if\, WT \\ & 0 \quad else \end{cases}$$

Let $\boldsymbol{\theta_{NWT}} = (\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_k})$. Let $\mathbf{S}$ be a vector denoting the allocations of each NWT observation to cluster $1, ..., k$. Then $\mathbf{S}_{(-i)}$ the vector S without observation i. The vector $\mathbf{n}$ denotes the number of observations in each of the k NWT components.

The superscript (L) denotes the L^{th} iteration. First initial values are determined, the $(0)^{\text{th}}$ values. Note that for the generation of latent vectors only interval censored generation is shown for brevity. As the endpoints of the data set are both known and fixed, the censoring is adjusted accordingly.

Update $\boldsymbol{\theta_{WT}} \mid \mathbf{X}, \mathbf{c}$

- ${\mu_{WT}}^{(L)} \mid {\sigma_{WT}}^{2(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)} \sim N(\bar{X}^{(L-1)}, \sqrt{\frac{{\sigma_{WT}}^{2(L-1)}}{\sum_i {c_i}^{(L-1)}}})$
- ${\sigma_{WT}}^{2(L)} \mid {\mu_{WT}}^{(L)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(t-1)} \sim InvGamma(\sum_i \frac{c_i^{(L-1)}}{2}, \sum_{i:c_i=1} ({X_i}^{(L-1)} - {\mu_{WT}}^{(L)})^2/2)$

Update $\pi \mid \mathbf{c}$

- $\pi^{(L)} \mid \mathbf{c}^{(L-1)} \sim TrBeta(a = .5, b = 1, \alpha' = 1+\sum_i c_i{}^{(L-1)}, \beta' = 1+N_{tot}-\sum_i c_i^{(L-1)})$

Update $\sigma_\delta^2 \mid \boldsymbol{\delta}$

- $\sigma_\delta{}^{2(L)} \mid \boldsymbol{\delta}^{(L-1)} \sim InvGamma(\alpha' = \alpha_\delta + \frac{N_{tot}}{2}, \beta' = \beta_\delta + \frac{\sum_i \delta_i{}^{(L)2}}{2})$

Update $\boldsymbol{\theta_{NWT}} \mid \mathbf{X}, \mathbf{c}$.

- Use Neal [2000] Algorithm 8 with D auxiliary classes. For $i = 1, ..., N_{tot} - \sum_i c_i$, $\mathbf{S}^{(L)} \mid \alpha_{conc}{}^{(L-1)}, \boldsymbol{\theta_{NWT}}{}^{(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)}, \mathbf{S_{-i}}{}^{(L-1)}$

For $h = 1, ...k$,

- $\mu_h{}^{(L)} \mid \mathbf{S}^{(L)}, \boldsymbol{\sigma_h}{}^{(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)} \sim TrN(A, \infty, \mu', \sqrt{\sigma^{2'}})$

where

$$\mu' = \frac{\frac{\mu_{G_0}{}^{(L-1)}}{\sigma_0^2} + \frac{\sum_{i:S^{(L)}{}_i=h} X_i^{(L-1)}}{\sigma_h{}^{2(L-1)}}}{\frac{1}{\sigma_0{}^2} + \frac{n_h{}^{(L)}}{\sigma_h^{2(L-1)}}}$$

and

$$\sigma^{2'} = \frac{1}{\frac{1}{\sigma_0{}^2} + \frac{n_h}{\sigma_h^{2(L-1)}}}$$

- $\sigma_h^{2(L)} \mid \mathbf{S}^{(L)}, \mathbf{X}^{(L)}, \mu_h{}^{(L)}, \mathbf{c}^{(L-1)} \sim InvGamma(\alpha^{'} = \alpha_1 + \frac{n_h}{2}, \beta^{'} = \beta_1 + \frac{\sum_{i:S_i=h}(X_i-\mu_h)^2}{2})$
- $w_h{}^{(L)} = \frac{n_h{}^{(L)}}{N_{tot}(1-c^{(\widehat{L-1})})}$

Update the mean of the base distribution.

- $\mu_{G_0}{}^{(L)} \mid \boldsymbol{\mu_{NWT}}{}^{(L)} \sim TrN(A, \infty, \mu^{'}_{G_0} = \frac{\frac{\mu_{G_B}}{\sigma_{G_B}{}^2} + \frac{\sum_r \mu_r{}^{(L)}}{\sigma_0^2}}{\frac{1}{\sigma_{G_B}{}^2} + \frac{k^{(L)}}{\sigma_0{}^2}}, \sigma^{'}_{G_0} = (\frac{1}{\sigma^2_{G_B}} + \frac{k^{(L)}}{\sigma_0^2})^{-1/2})$

Update α_{conc} using West [1992] where γ is the Euler-Mascheroni constant.

- $\alpha_{conc}{}^{(L)} \mid k^{(L)}, \mathbf{c}^{(L-1)} \overset{asy.}{\sim} Gamma(a + k^{(L)} - 1, b + \gamma + \log(N_{tot} - \sum_i c_i{}^{(L-1)}))$

Update $\mathbf{X} \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\delta}$.

For the observations, $i = 1, ..., N_{tot}$ where $c_i = 1$

- $\mathbf{X}^{(L)} \mid \mathbf{Y}, \mathbf{c}^{(L-1)}, \mu_{WT}{}^{(L)}, \sigma_{WT}{}^{(L)}, \boldsymbol{\delta}^{(L-1)} \overset{ind.}{\sim} TrN(Y-1-\delta^{(L-1)}, Y-\delta^{(L-1)}, \mu_{WT}{}^{(L)}, \sigma_{WT}{}^{(L)})$

For the observations, $i = 1, ..., N_{tot}$ where $c_i = 0$

- $\mathbf{X}^{(L)} \mid \mathbf{Y}, \mathbf{S}^{(L)}, \mathbf{c}^{(L)}, \boldsymbol{\mu_{NWT}}^{(L)}, \boldsymbol{\sigma_{NWT}}^{(L)} \overset{ind.}{\sim}$
  $TrN(\mathbf{Y} - 1 - \boldsymbol{\delta}^{(L-1)}, Y - \boldsymbol{\delta}^{(L-1)}, \boldsymbol{\mu_{NWT}}_{i:S_i{}^{(L)}}{}^{(t)}, \boldsymbol{\sigma_{NWT}}_{i:S_i{}^{(L)}}{}^{(L)})$

Update $\boldsymbol{\delta} \mid \mathbf{Y}, \mathbf{X}, \sigma_\delta^2$.

- $\boldsymbol{\delta}^{(L)} \mid \mathbf{Y}, \mathbf{X}^{(L)}, \mathbf{c}^{(L-1)}, \mathbf{S}^{(L)}, \sigma_\delta{}^{2(L)} \overset{ind.}{\sim}$
  $TrN(Y - 1 - X^{(L-1)}, Y - X^{(L-1)}, \mu = 0, \sigma_\delta{}^{(L)})$

Update $\mathbf{c} \mid \mathbf{X}, \pi, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}$

- $\mathbf{c}^{(L)} \mid \mathbf{X}^{(L)}, \pi^{(L)}, \boldsymbol{\theta_{WT}}^{(L)}, \boldsymbol{\theta_{NWT}}^{(L)} \sim Bernoulli(p')$

where

$$p' = \frac{\frac{\pi^{(L)}}{\sigma_{WT}{}^{(L)}}\phi(\frac{X^{(L)} - \mu_{WT}^{(L)}}{\sigma_{WT}{}^{(L)}})}{\frac{\pi^{(L)}}{\sigma_{WT}{}^{(L)}}\phi(\frac{X^{(L)} - \mu_{WT}{}^{(L)}}{\sigma_{WT}{}^{L}}) + (1 - \pi^{(L)}) \sum_{h=1}^{k^{(L)}} \frac{w_h{}^{(L)}}{\sigma_h{}^{(L)}}\phi(\frac{X^{(L)} - \mu_h{}^{(L)}}{\sigma_h{}^{(L)}})}$$

Then increase $L \leftarrow L + 1$

Computation: log2gamma WT Distribution for Single-Year Data

Update $\boldsymbol{\theta_{WT}} \mid \mathbf{X}, \mathbf{c}$

- $P(\alpha^{(L)} \mid \beta^{(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)}) \propto P(\mathbf{X}^{(L-1)} \mid \mathbf{c}^{(L-1)}, \alpha^{(L-1)}, \beta^{(L-1)}) pr(\alpha^{(L-1)})$

- $\beta^{(L)} \mid \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)}, \alpha^{(L)} \sim Gamma(\alpha' = N_{tot}\bar{c}^{(L-1)}\alpha^{(L)}, \beta' = \sum_{i:c^{(L)}{}_i=1} 2^{X_i^{(L)}})$

Update $\pi \mid \mathbf{c}$

- $\pi^{(L)} \mid \mathbf{c}^{(L-1)} \sim TrBeta(a = .5, b = 1, \alpha' = 1 + \sum_i c_i{}^{(L-1)}, \beta' = 1 + N_{tot} - \sum_i c_i^{(L-1)})$

Update $\sigma_\delta^2 \mid \boldsymbol{\delta}$

- $\sigma_\delta{}^{2(L)} \mid \boldsymbol{\delta}^{(L-1)} \sim InvGamma(\alpha' = \alpha_\delta + \frac{N_{tot}}{2}, \beta' = \beta_\delta + \frac{\sum_i \delta_i{}^{(L)2}}{2})$

Update $\boldsymbol{\theta_{NWT}} \mid \mathbf{X}, \mathbf{c}$.

- Use Neal [2000] Algorithm 8 with D auxiliary classes. For $i = 1, ..., N_{tot} - \sum_i c_i$, $\mathbf{S}^{(L)} \mid \alpha_{conc}{}^{(L-1)}, \boldsymbol{\theta_{NWT}}^{(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)}, \mathbf{S_{-i}}^{(L-1)}$

For $h = 1, ...k$,

- ${\mu_h}^{(L)} \mid \mathbf{S}^{(L)}, {\boldsymbol{\sigma_h}}^{(L-1)}, \mathbf{X}^{(L-1)}, \mathbf{c}^{(L-1)} \sim TrN(A, \infty, \mu', \sqrt{\sigma^{2'}})$

where

$$\mu' = \frac{\frac{{\mu_{G_0}}^{(L-1)}}{\sigma_0^2} + \frac{\sum_{i:{S^{(L)}}_i=h} X_i^{(L-1)}}{{\sigma_h}^{2(L-1)}}}{\frac{1}{{\sigma_0}^2} + \frac{{n_h}^{(L)}}{\sigma_h^{2(L-1)}}}$$

and

$$\sigma^{2'} = \frac{1}{\frac{1}{{\sigma_0}^2} + \frac{n_h}{\sigma_h^{2(L-1)}}}$$

- $\sigma_h^{2(L)} \mid \mathbf{S}^{(L)}, \mathbf{X}^{(L)}, {\mu_h}^{(L)}, \mathbf{c}^{(L-1)} \sim InvGamma(\alpha^{'} = \alpha_1 + \frac{n_h}{2}, \beta^{'} = \beta_1 + \frac{\sum_{i:S_i=h}(X_i - \mu_h)^2}{2})$
- ${w_h}^{(L)} = \frac{{n_h}^{(L)}}{N_{tot}(1 - c^{(\hat{L}-1)})}$

Update the mean of the base distribution.

- ${\mu_{G_0}}^{(L)} \mid {\boldsymbol{\mu_{NWT}}}^{(L)} \sim TrN(A, \infty, \mu^{'}_{G_0} = \frac{\frac{\mu_{G_B}}{{\sigma_{G_B}}^2} + \frac{\sum_r {\mu_r}^{(L)}}{\sigma_0^2}}{\frac{1}{{\sigma_{G_B}}^2} + \frac{k^{(L)}}{{\sigma_0}^2}}, \sigma^{'}_{G_0} = (\frac{1}{\sigma^2_{G_B}} + \frac{k^{(L)}}{\sigma_0^2})^{-1/2})$

Update α_{conc} using West [1992] where γ is the Euler-Mascheroni constant.

- ${\alpha_{conc}}^{(L)} \mid k^{(L)}, \mathbf{c}^{(L-1)} \overset{asy.}{\sim} Gamma(a + k^{(L)} - 1, b + \gamma + \log(N_{tot} - \sum_i {c_i}^{(L-1)}))$

Update $\mathbf{X} \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\delta}$.

For the observations, $i = 1, ..., N_{tot}$ where $c_i = 1$

- $\mathbf{X}^{(L)} \mid \mathbf{Y}, \mathbf{c}^{(L)}, \boldsymbol{\delta}^{(L)}, \alpha^{(L)}, \beta^{(L)} \sim log_2(TrGamma(a = 2^{Y-1-\delta^{(L)}}, b = 2^{Y-\delta^{(L)}}, \alpha^{(L)}, \beta^{(L)}))$

For the observations, $i = 1, ..., N_{tot}$ where $c_i = 0$

- $\mathbf{X}^{(L)} \mid \mathbf{Y}, \mathbf{S}^{(L)}, \mathbf{c}^{(L)}, {\boldsymbol{\mu_{NWT}}}^{(L)}, {\boldsymbol{\sigma_{NWT}}}^{(L)} \overset{ind.}{\sim}$
  $TrN(\mathbf{Y} - 1 - \boldsymbol{\delta}^{(L-1)}, Y - \boldsymbol{\delta}^{(L-1)}, {\boldsymbol{\mu_{NWT}}_{i:{S_i}^{(L)}}}^{(t)}, {\boldsymbol{\sigma_{NWT}}_{i:{S_i}^{(L)}}}^{(L)})$

Update $\boldsymbol{\delta} \mid \mathbf{Y}, \mathbf{X}, \sigma_\delta^2$.

- $\boldsymbol{\delta}^{(L)} \mid \mathbf{Y}, \mathbf{X}^{(L)}, \mathbf{c}^{(L-1)}, \mathbf{S}^{(L)}, {\sigma_\delta}^{2(L)} \overset{ind.}{\sim}$
  $TrN(Y - 1 - X^{(L-1)}, Y - X^{(L-1)}, \mu = 0, {\sigma_\delta}^{(L)})$

Update $\mathbf{c} \mid \mathbf{X}, \pi, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{NWT}}$

- $\mathbf{c}^{(L)} \mid \mathbf{X}^{(L)}, \pi^{(L)}, {\boldsymbol{\theta_{WT}}}^{(L)}, {\boldsymbol{\theta_{NWT}}}^{(L)} \sim Bernoulli(p')$

where

$$p' = \frac{\pi^{(L)} f_{l2g}(X^{(L-1)}; \alpha^{(L)}, \beta^{(L)})}{\pi^{(L)} f_{l2g}(X^{(L-1)}; \alpha^{(L)}, \beta^{(L)}) + (1 - \pi^{(L)}) \sum_{h=1}^{k^{(L)}} \frac{{w_h}^{(L)}}{{\sigma_h}^{(L)}} \phi(\frac{X^{(L-1)} - {\mu_h}^{(L)}}{{\sigma_h}^{(L)})}}$$

Then increase $L \leftarrow L + 1$

D. SINGLE-YEAR SIMULATION RESULTS

Normal Case

For $\sigma_\delta = 0.4, 0.5, 0.6$, we compare the performance of Turnidge et al. [2006] (TURN) in green, in orange is the method Turnidge et al. [2006] but with a subset of the mode plus one bin (TURNM), in brown is the method of Jaspers et al. [2014a] (JASP), and blue is BayesACME. When $\sigma_\delta = 0.5$, the method of Jaspers et al. [2016a] (JASPB) is compared as well in pink. For the Bayesian methods, the posterior mean (following burn-in) is taking as the Bayes estimate. The red horizontal line denotes the truth. Each set of boxplots is for the different level of contamination in the order of low, medium, and high as we move from left to right.

Comparison with Jaspers et al. (2016)

Only for $\sigma_\delta = 0.5$, we compare the the performance of Turnidge et al. [2006], Jaspers et al. [2014a], and Jaspers et al. [2016b] to BayesACME.

ECOFF Values

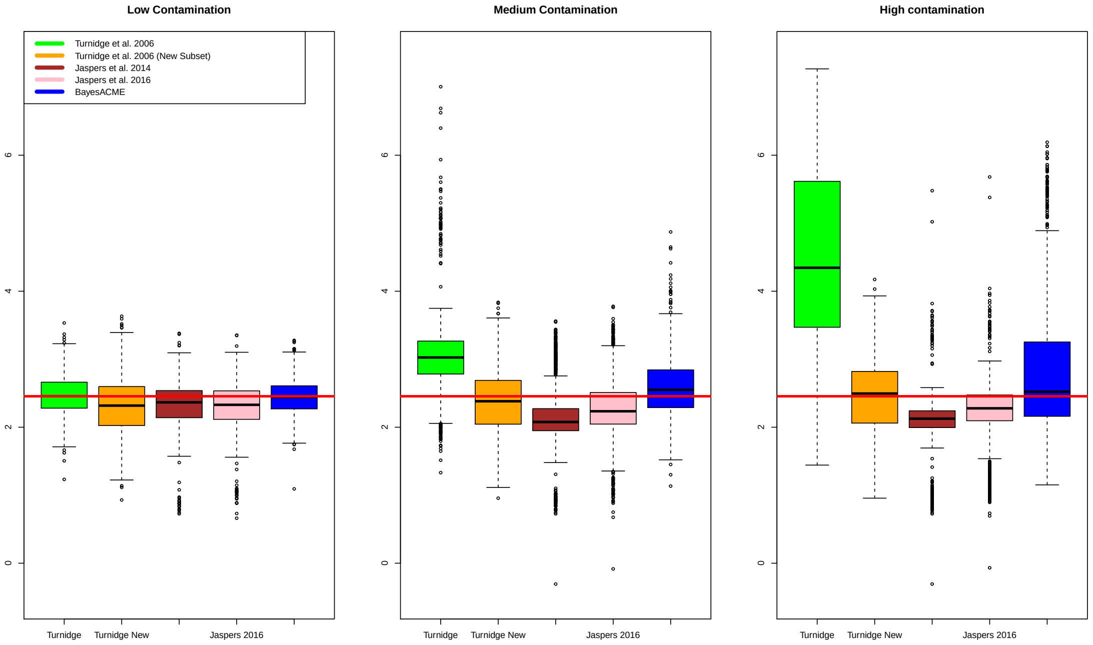


Figure D.1. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.5$.

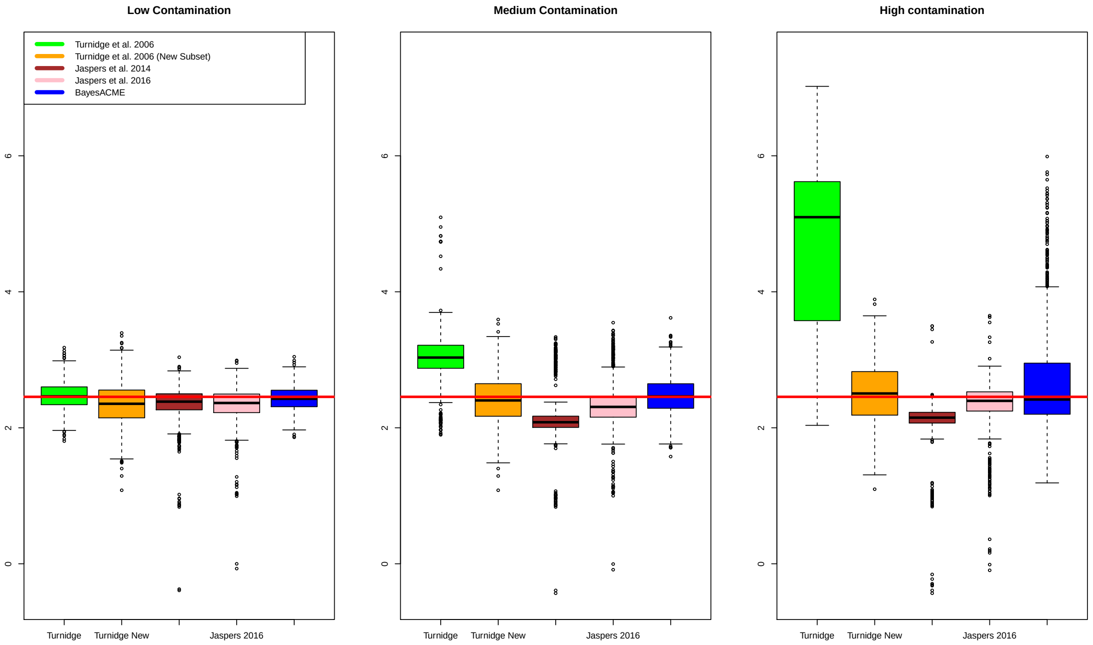


Figure D.2. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.5$.

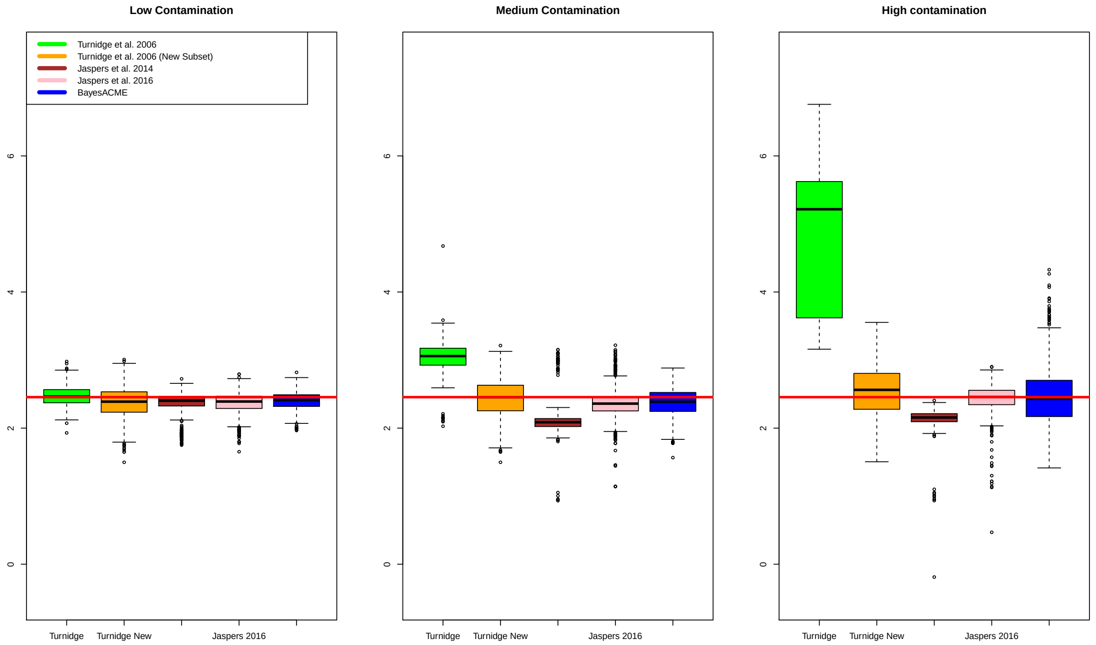


Figure D.3. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.5$.

Prevalence Estimates

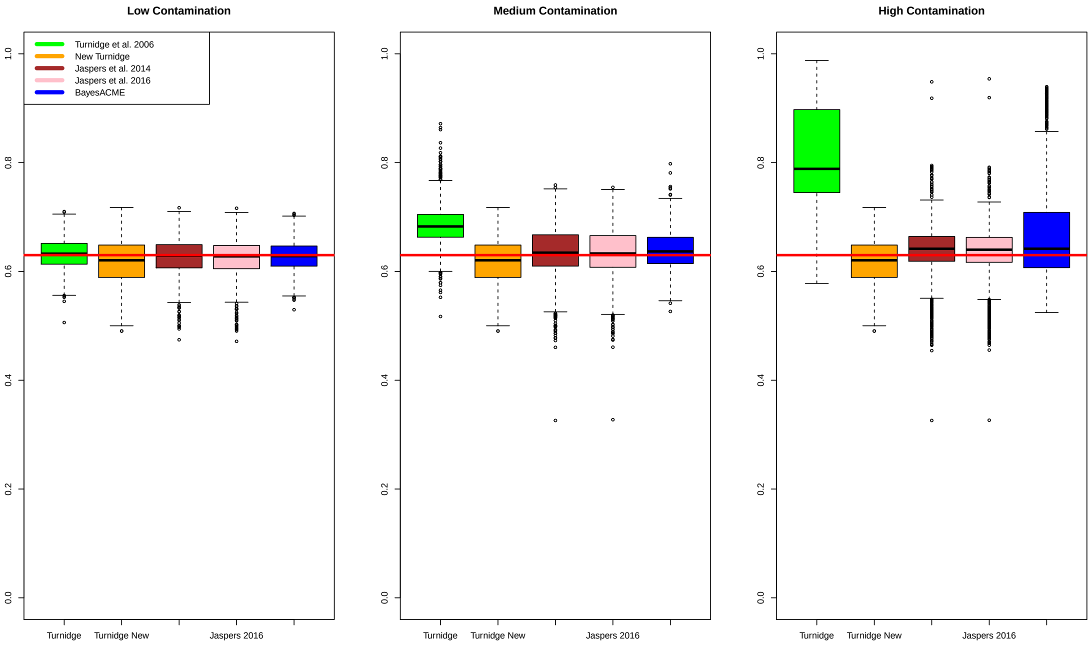


Figure D.4. This set of boxplots contrasts the estimates of the prevalence estimates of the discussed methods at size 300 and $\sigma_\delta = 0.5$.

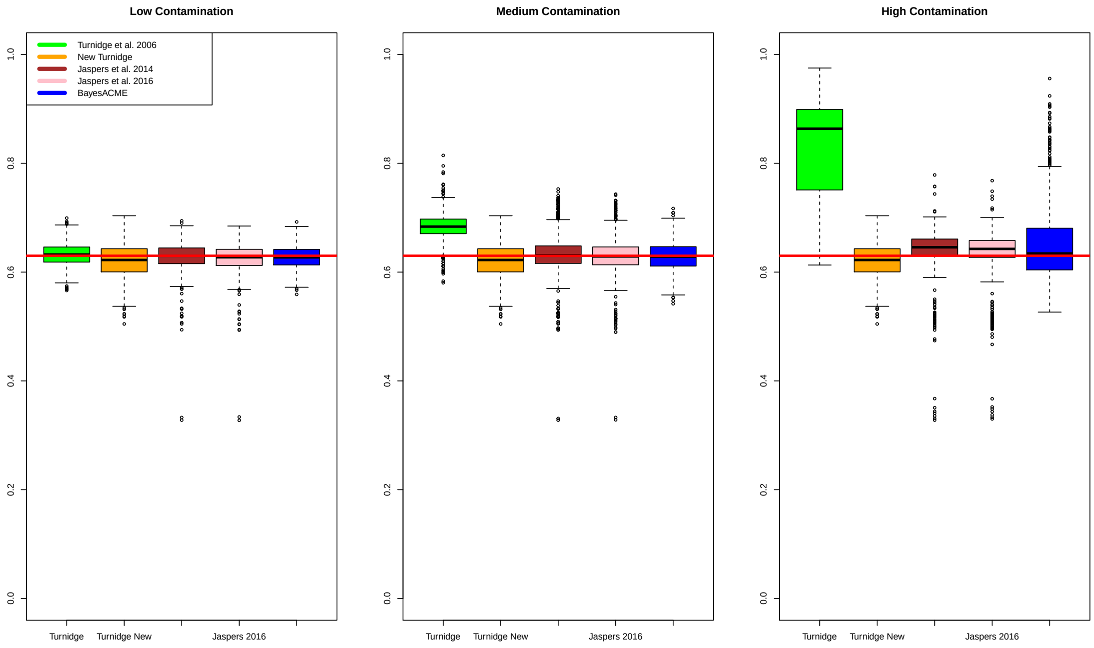


Figure D.5. This set of boxplots contrasts the estimates of the prevalence estimates the discussed methods at size 600 and $\sigma_\delta = 0.5$.

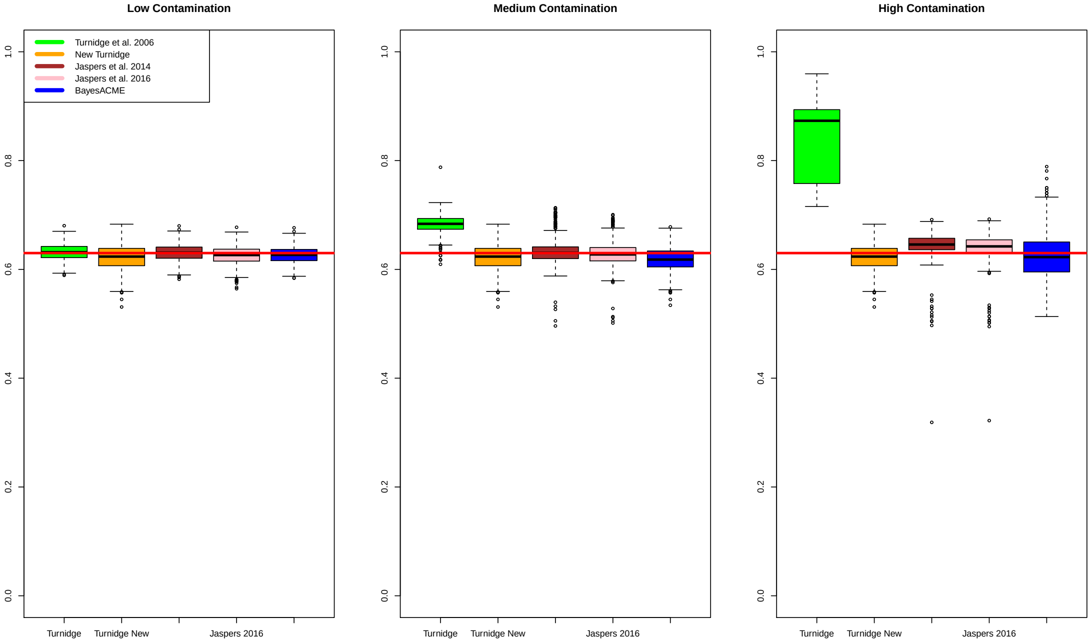


Figure D.6. This set of boxplots contrasts the estimates of the prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.5$.

Prior Sensitivity to σ_δ

ECOFF Values

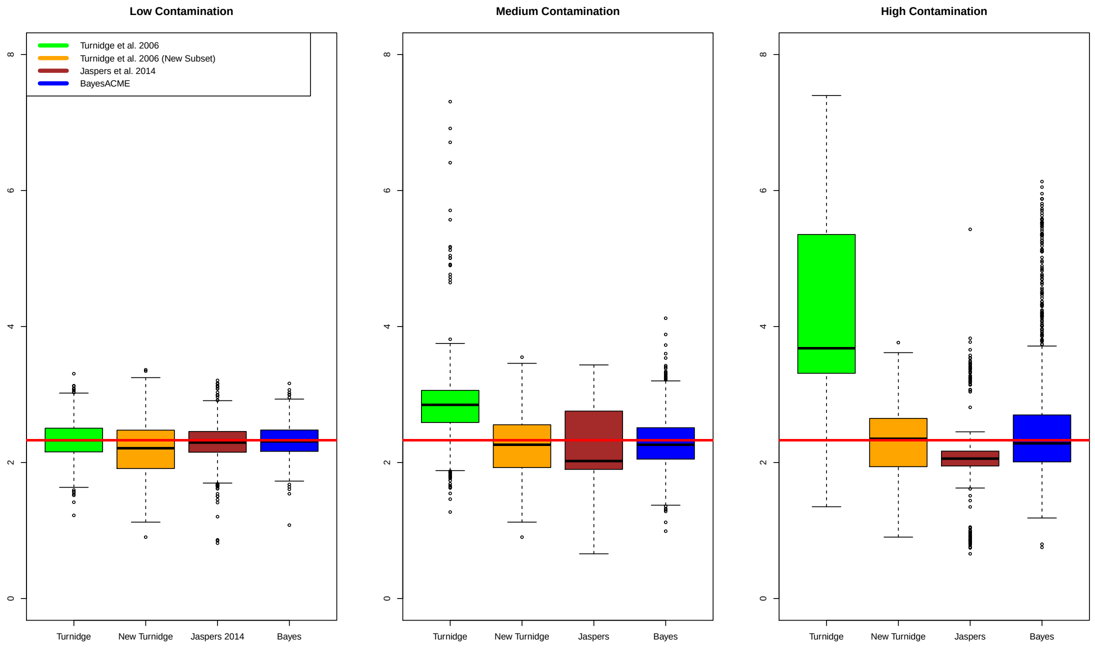


Figure D.7. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$.

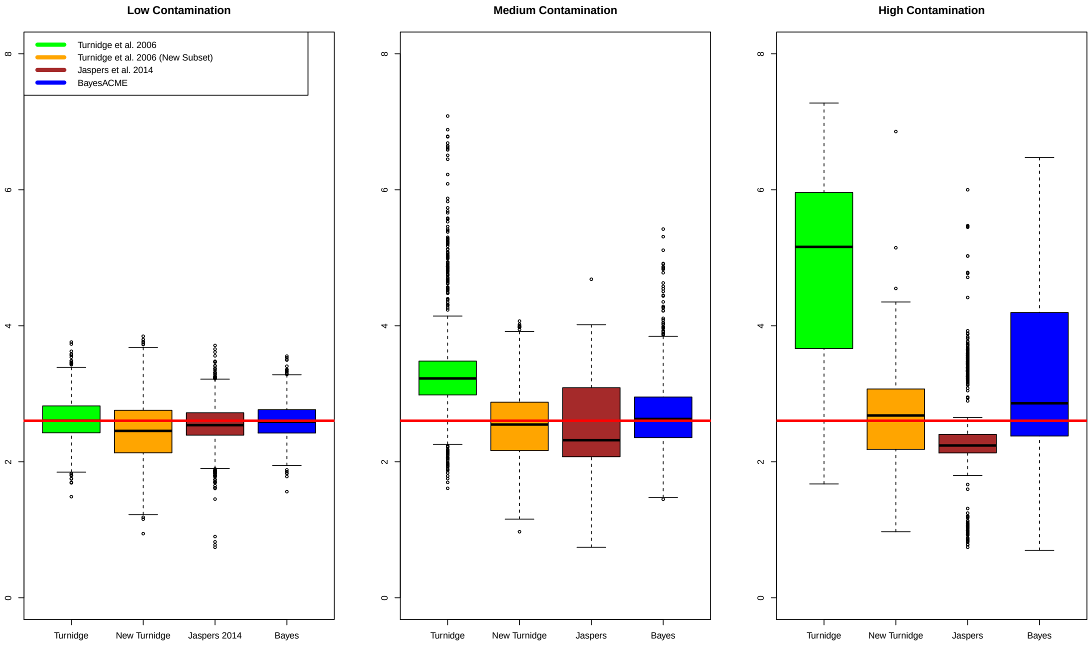


Figure D.8. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$.

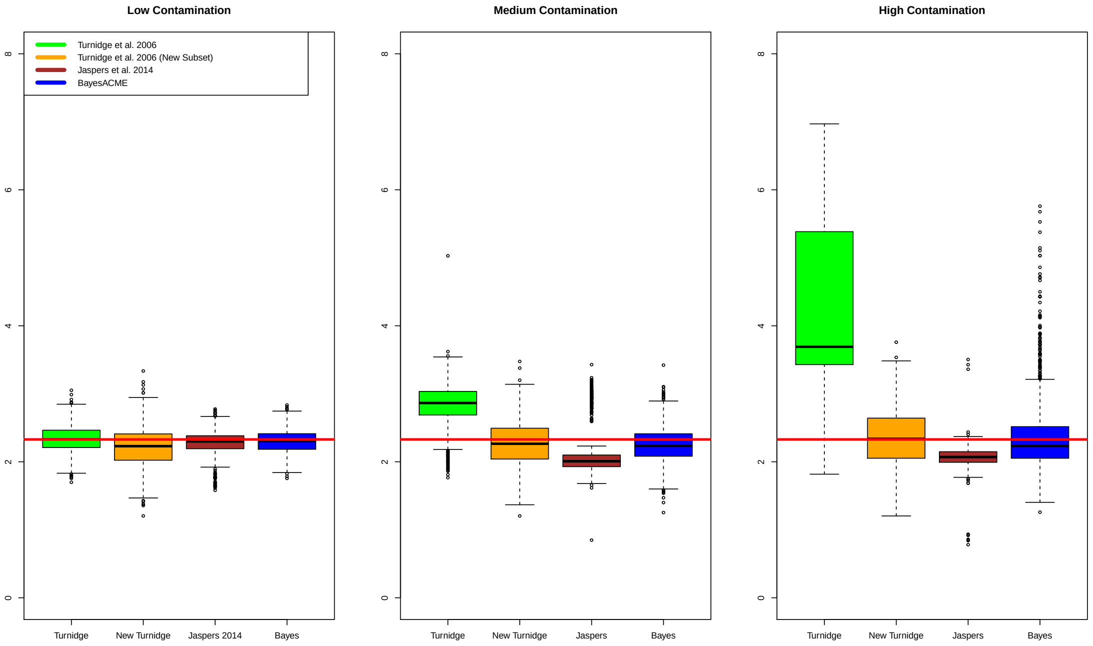


Figure D.9. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$.

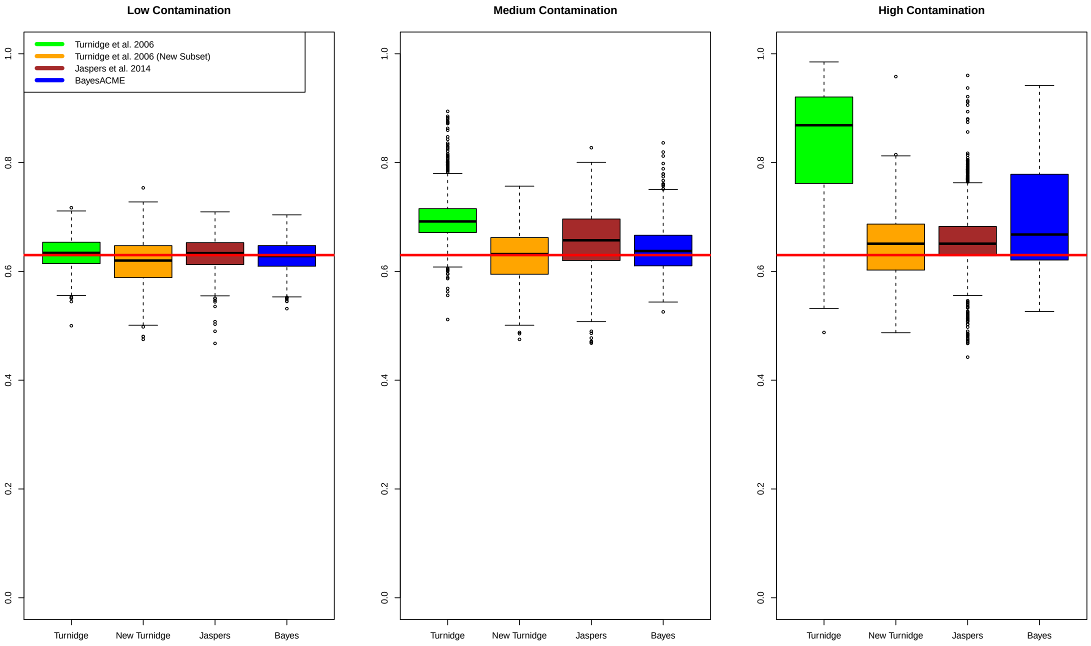


Figure D.10. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$.

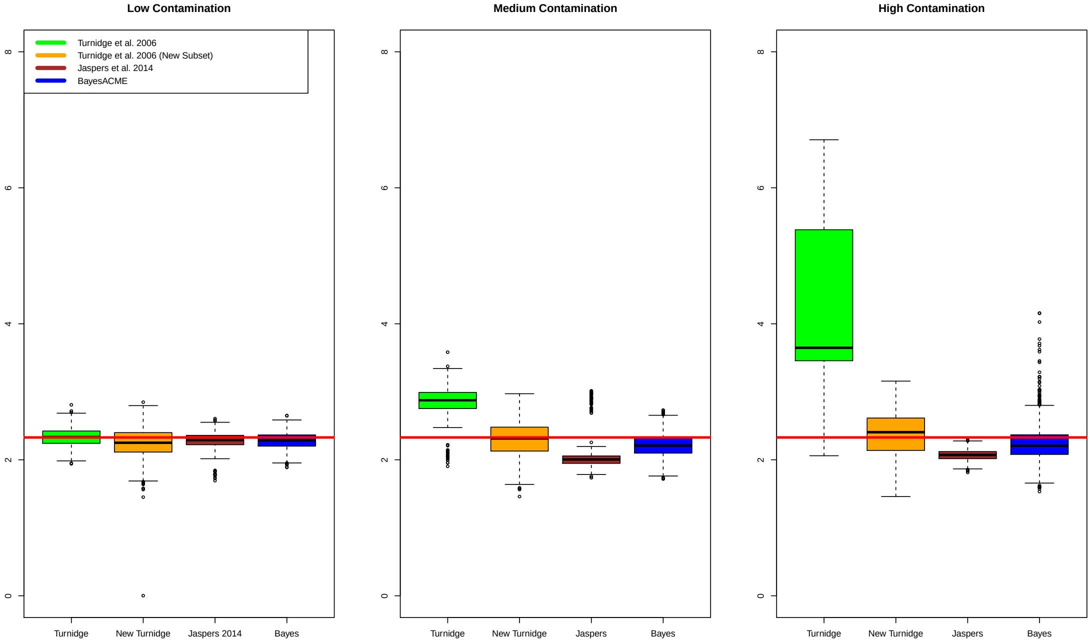


Figure D.11. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$.

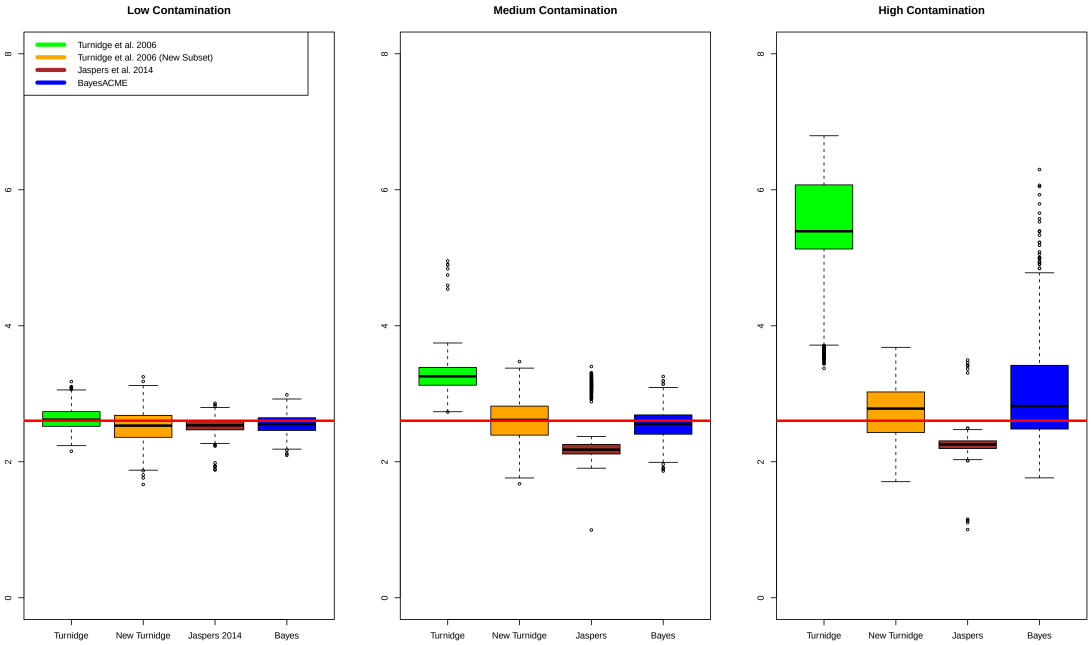


Figure D.12. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$.

Prevalence Estimates

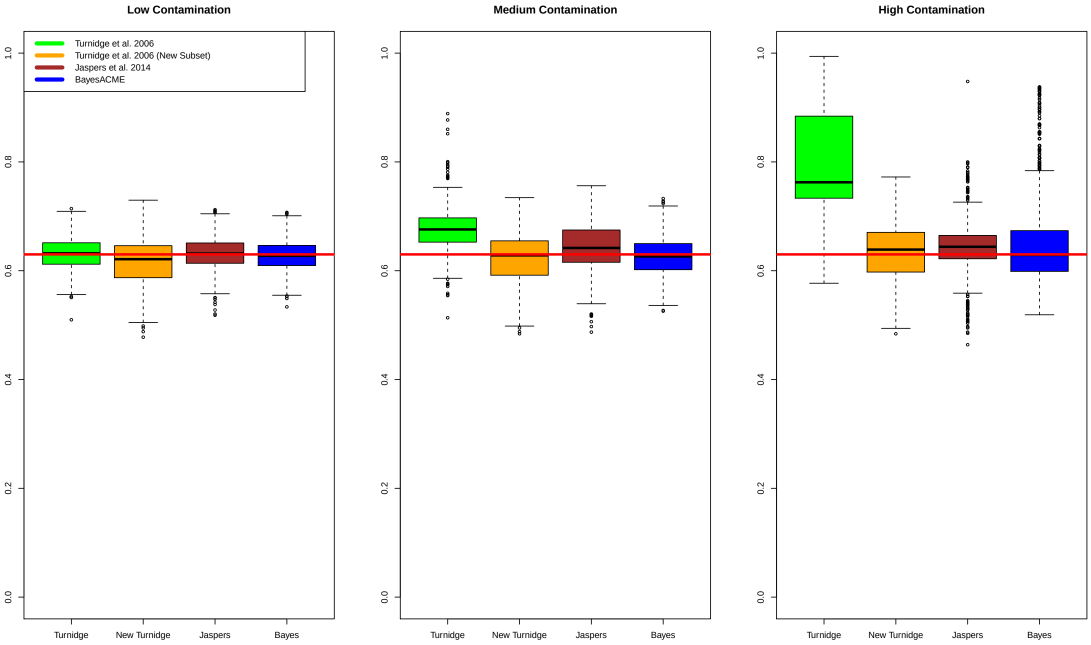


Figure D.13. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$.

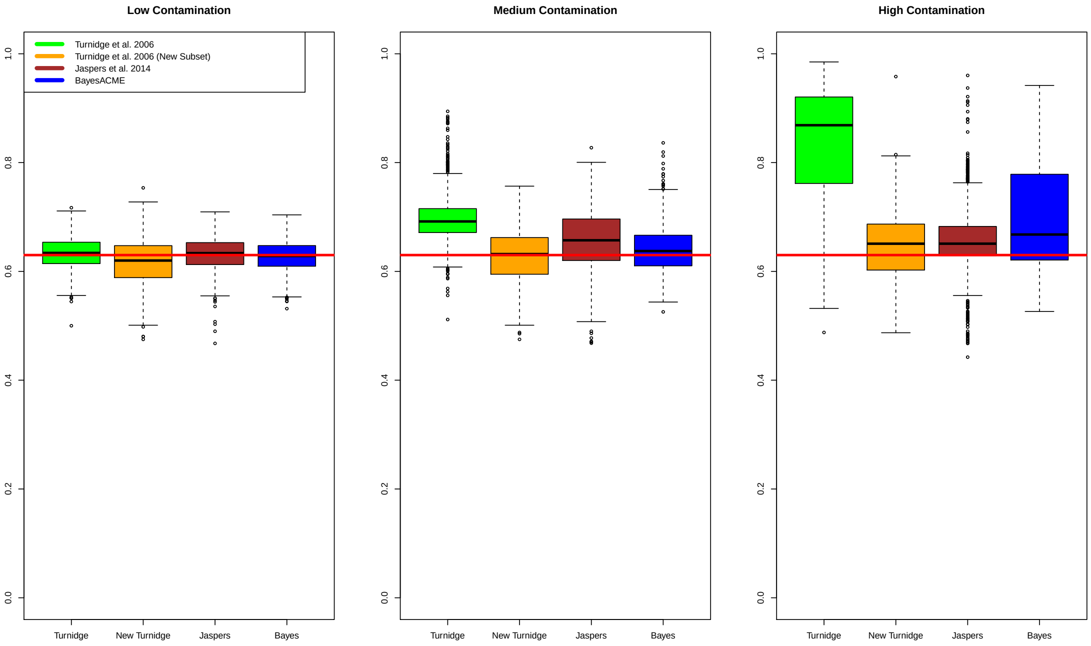


Figure D.14. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$.

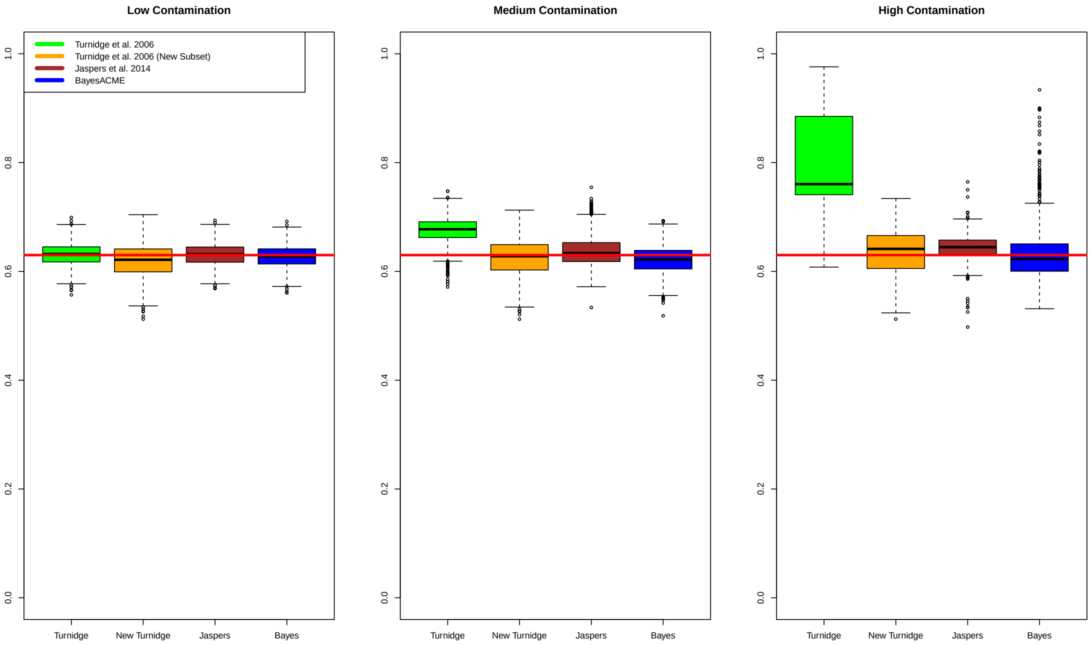


Figure D.15. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$.

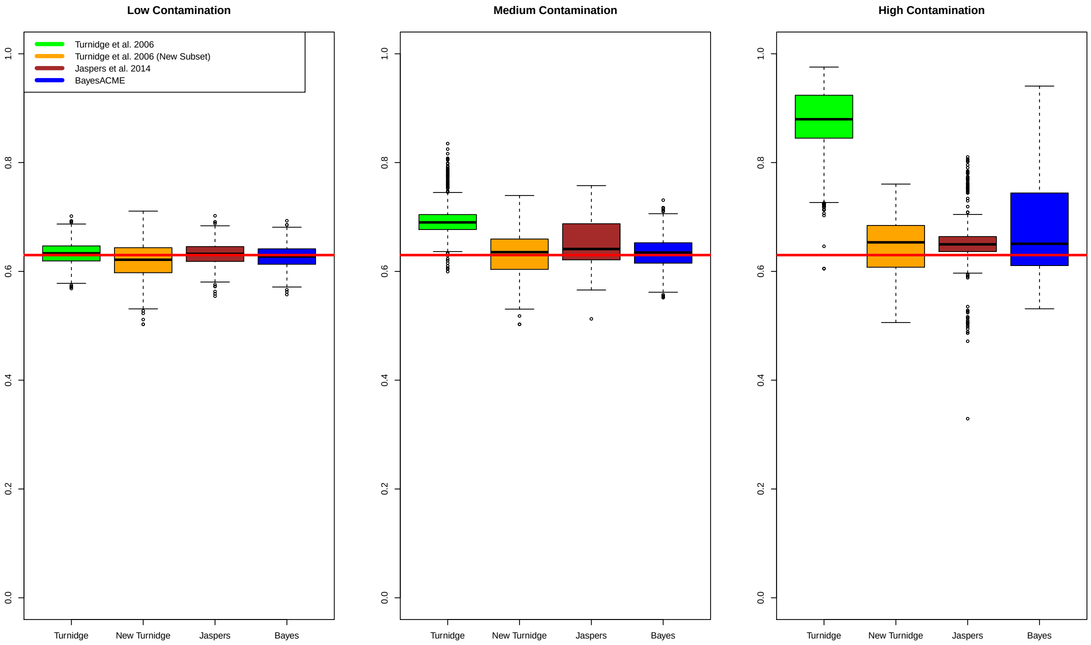


Figure D.16. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$.

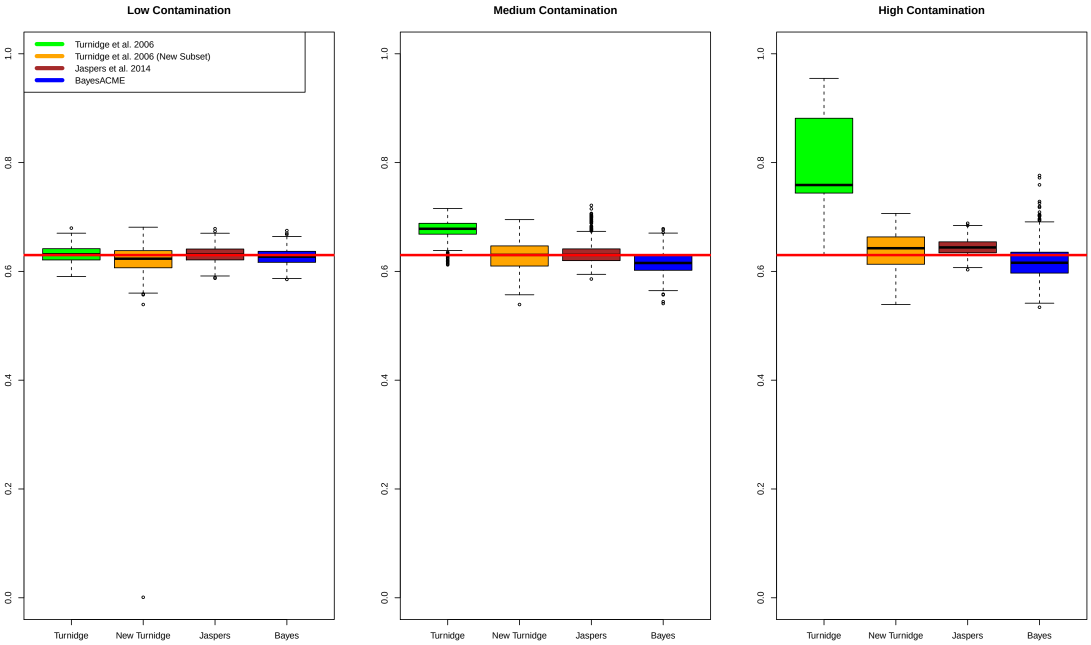


Figure D.17. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$.

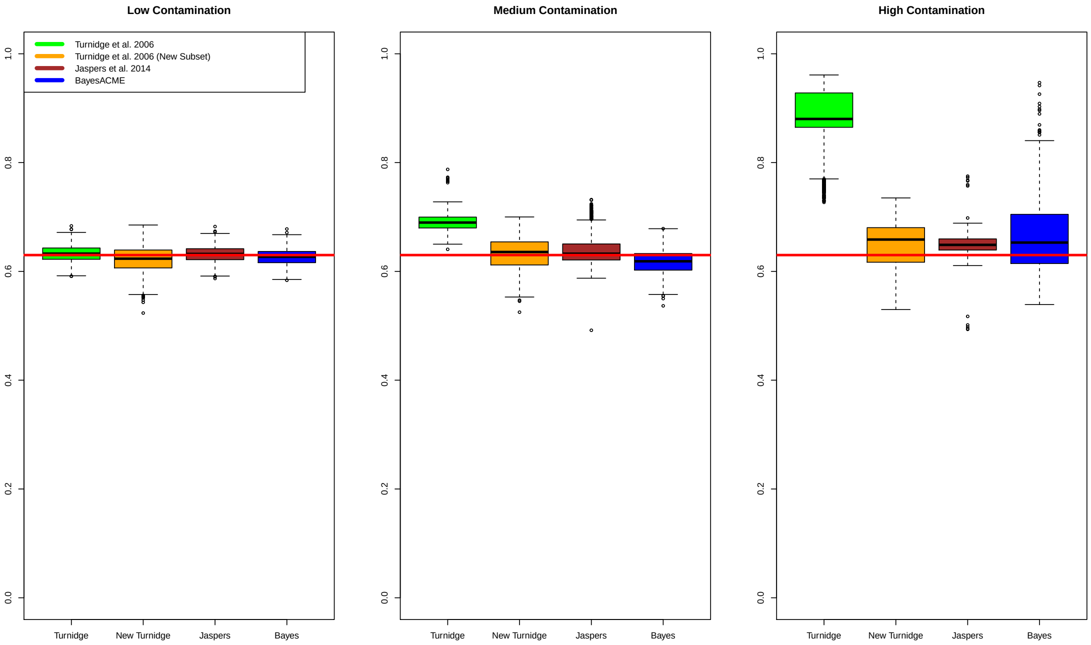


Figure D.18. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$.

log2gamma Case

ECOFF values

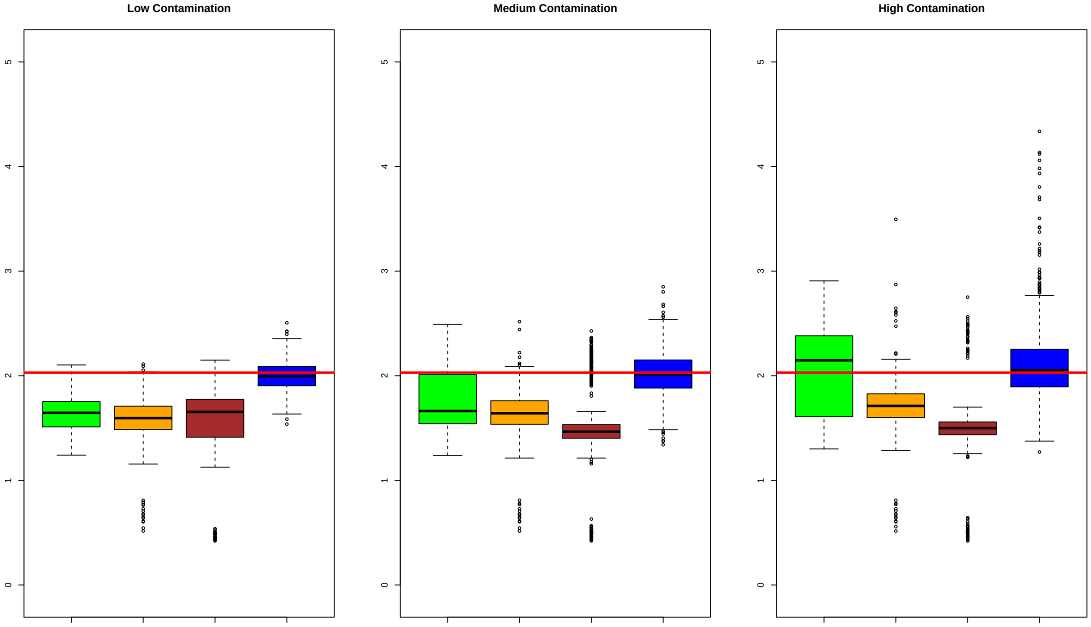


Figure D.19. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.5$.

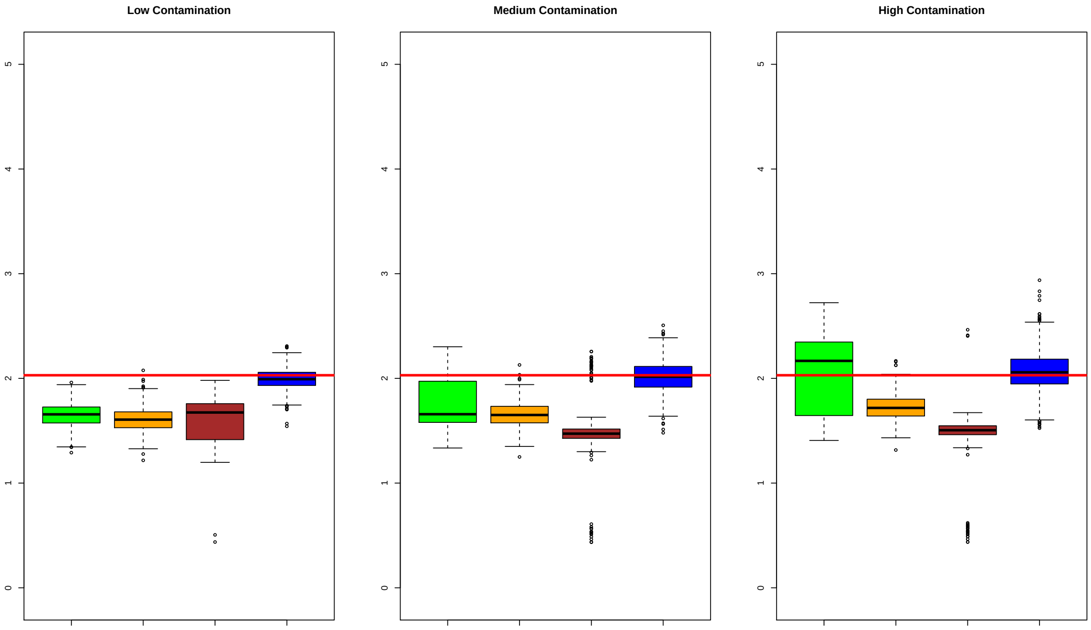


Figure D.20. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.5$.

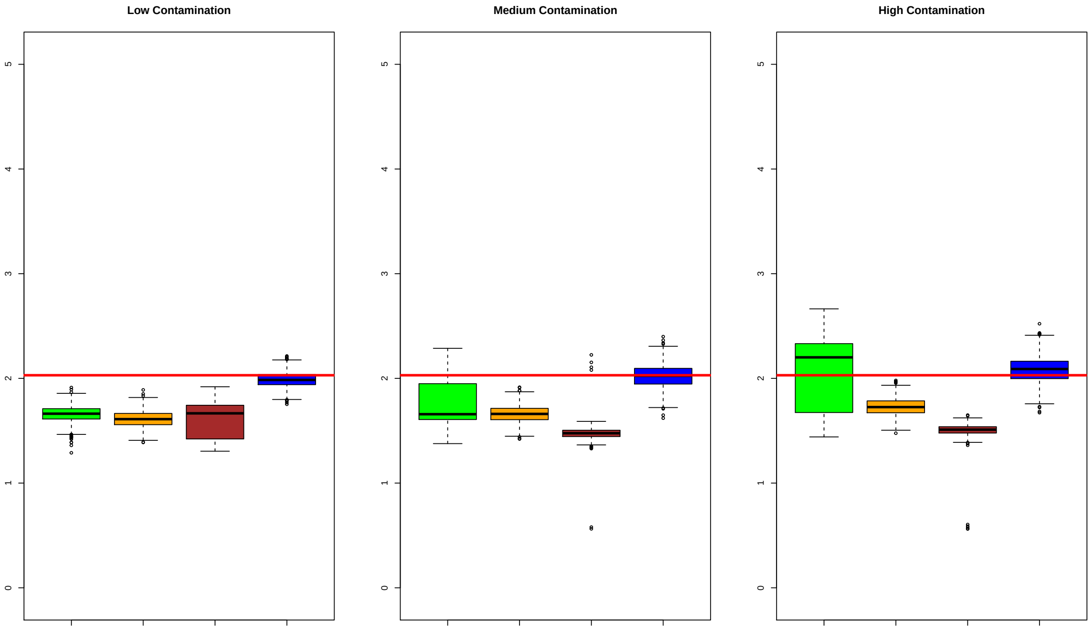


Figure D.21. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.5$.

Prevalence

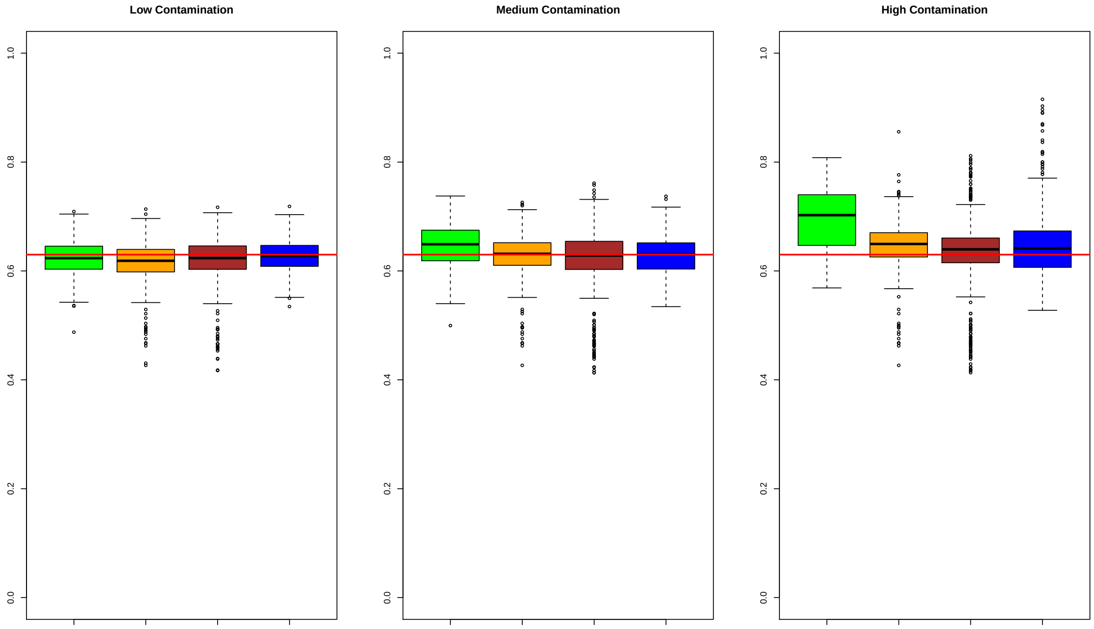


Figure D.22. This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 300 and $\sigma_\delta = 0.5$.

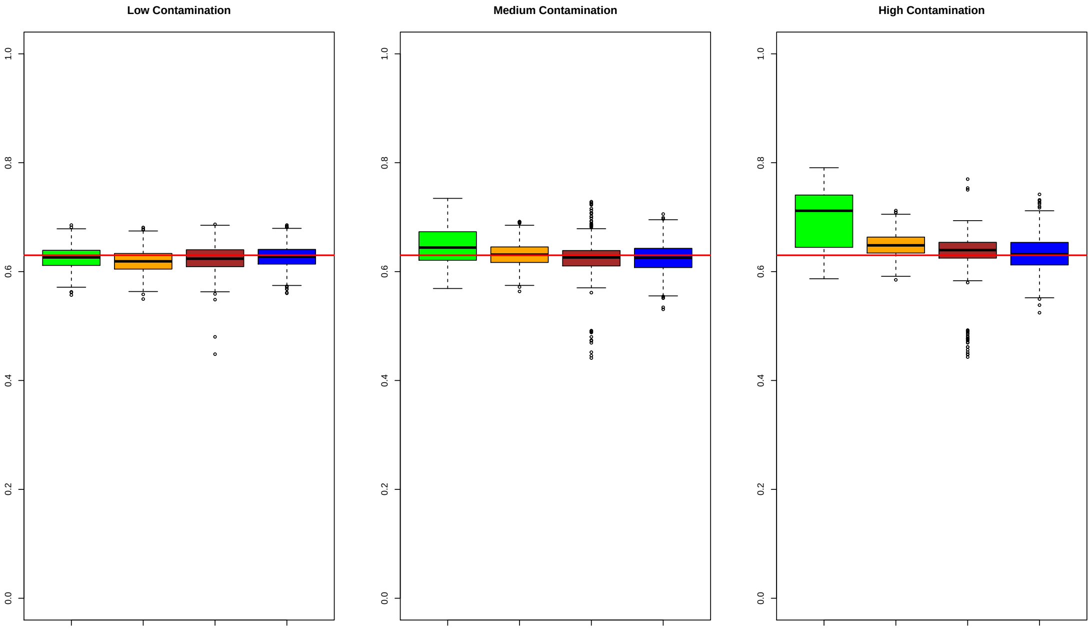


Figure D.23. This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 600 and $\sigma_\delta = 0.5$.

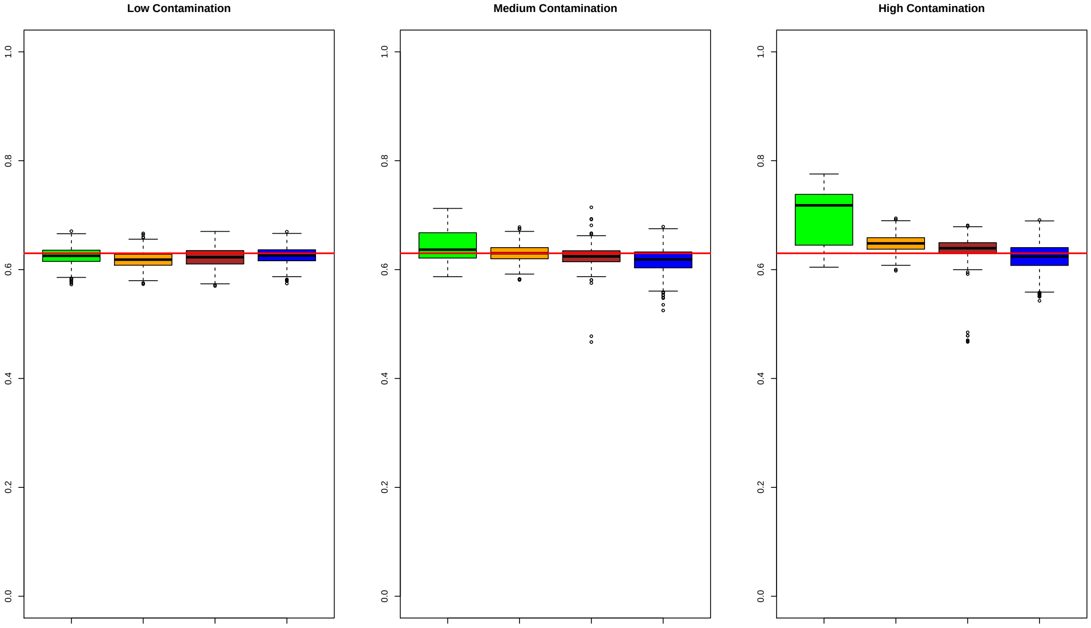


Figure D.24. This set of boxplots contrasts the results of the Prevalence Estimates of the compared methods at size 1200 and $\sigma_\delta = 0.5$.

Prior Sensitivity to σ_δ

ECOFF Values

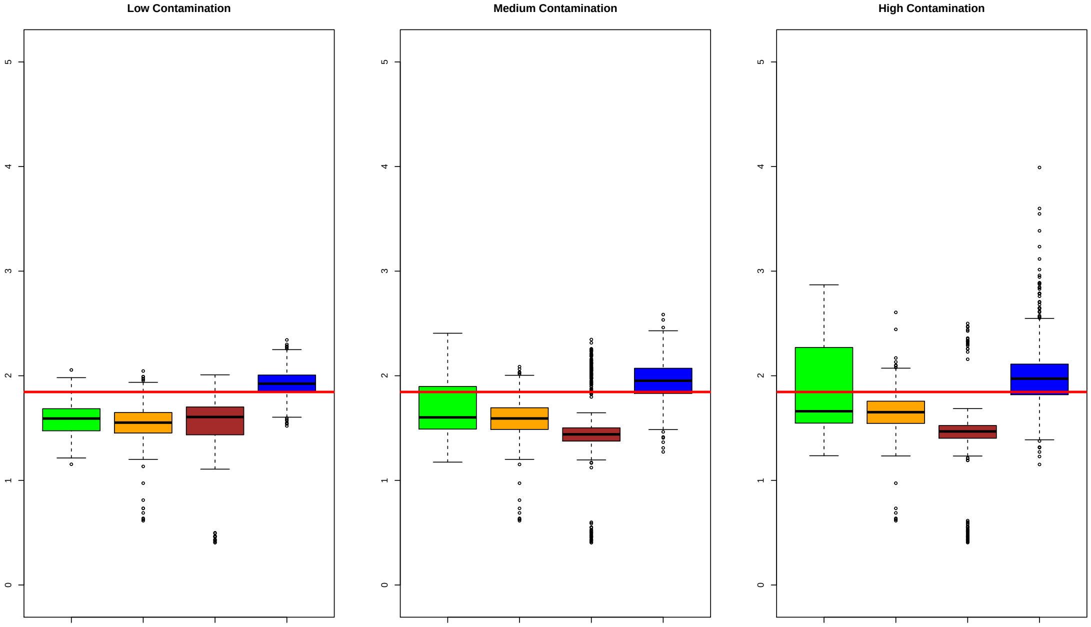


Figure D.25. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.4$.

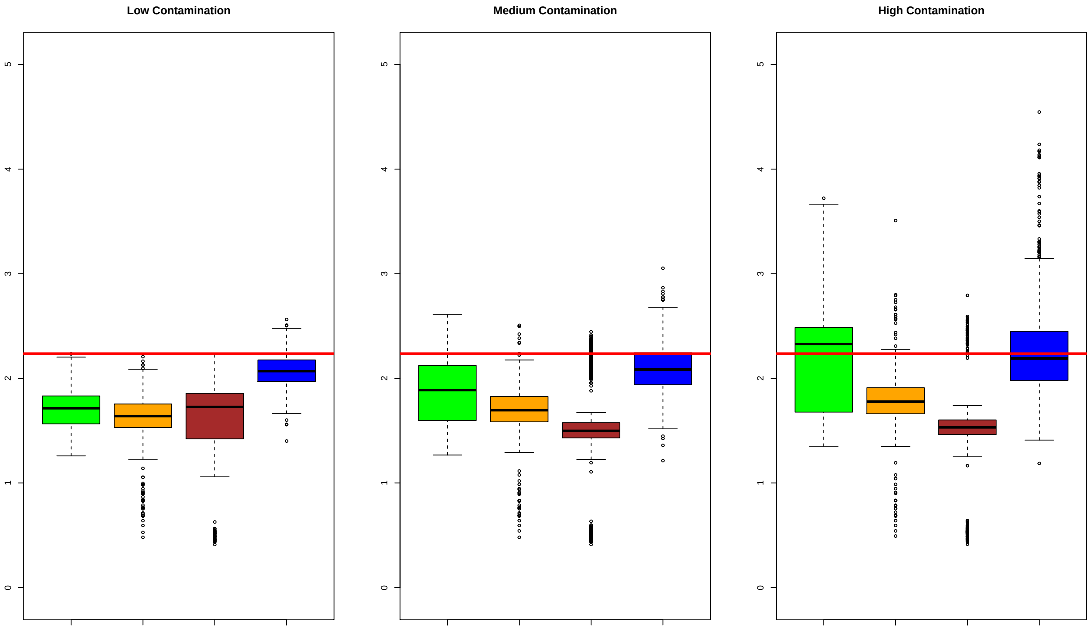


Figure D.26. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 300 and $\sigma_\delta = 0.6$.

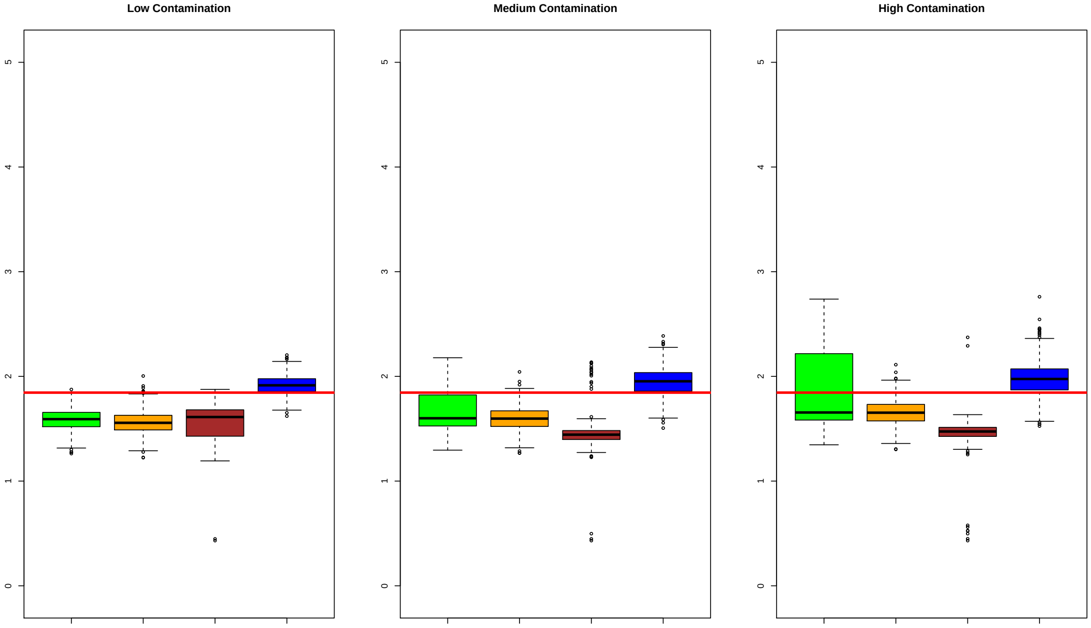


Figure D.27. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.4$.

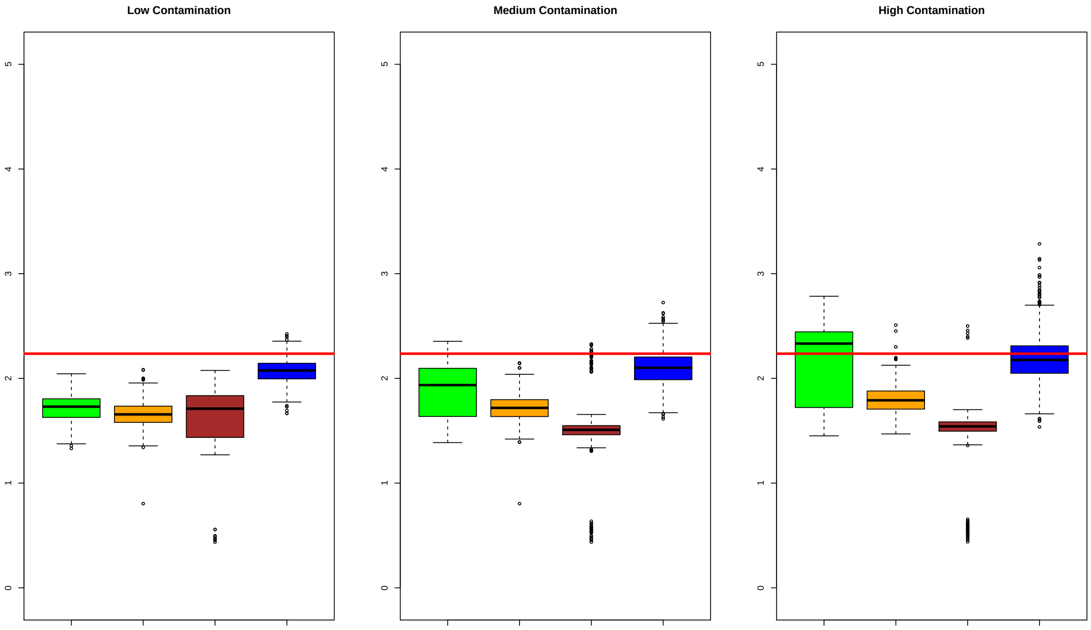


Figure D.28. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 600 and $\sigma_\delta = 0.6$.

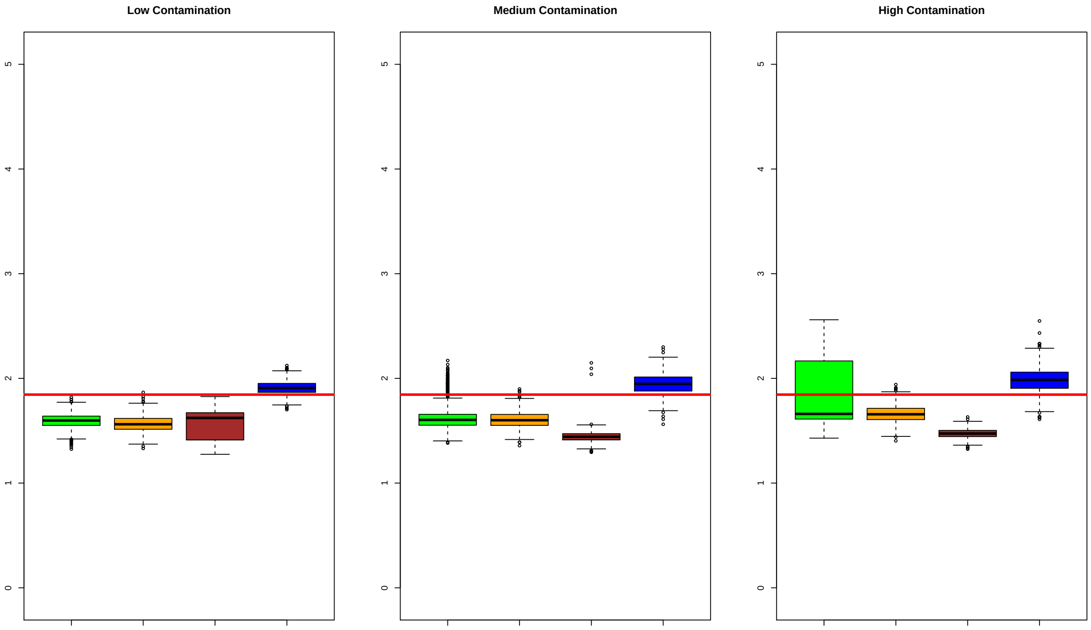


Figure D.29. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.4$.

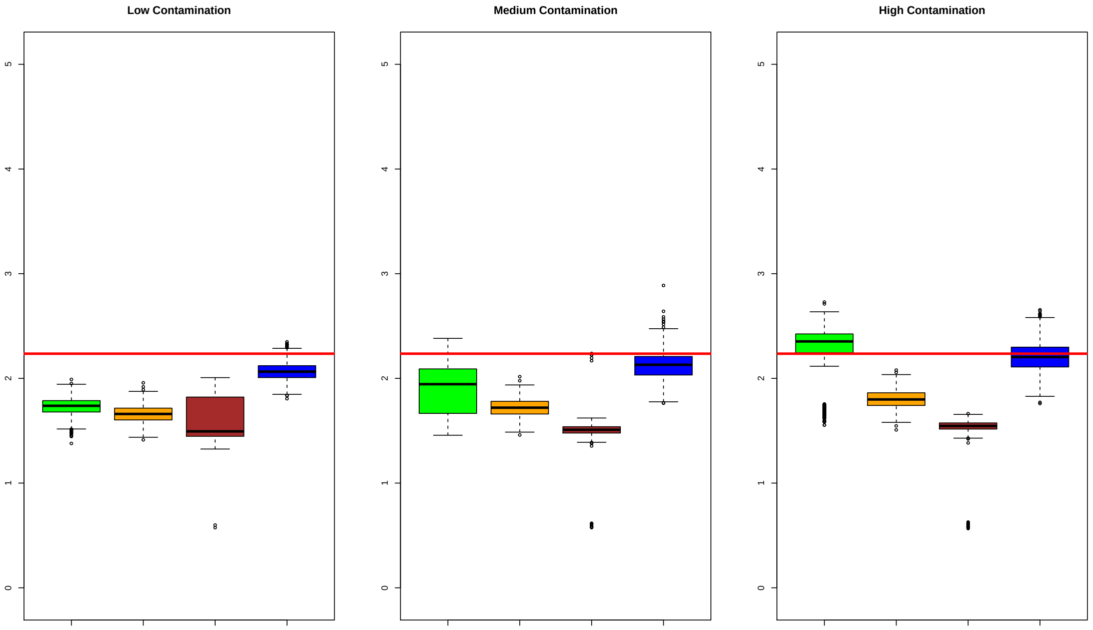


Figure D.30. This set of boxplots contrasts the results of the ECOFF values of the compared methods at size 1200 and $\sigma_\delta = 0.6$.

Prevalence Estimates

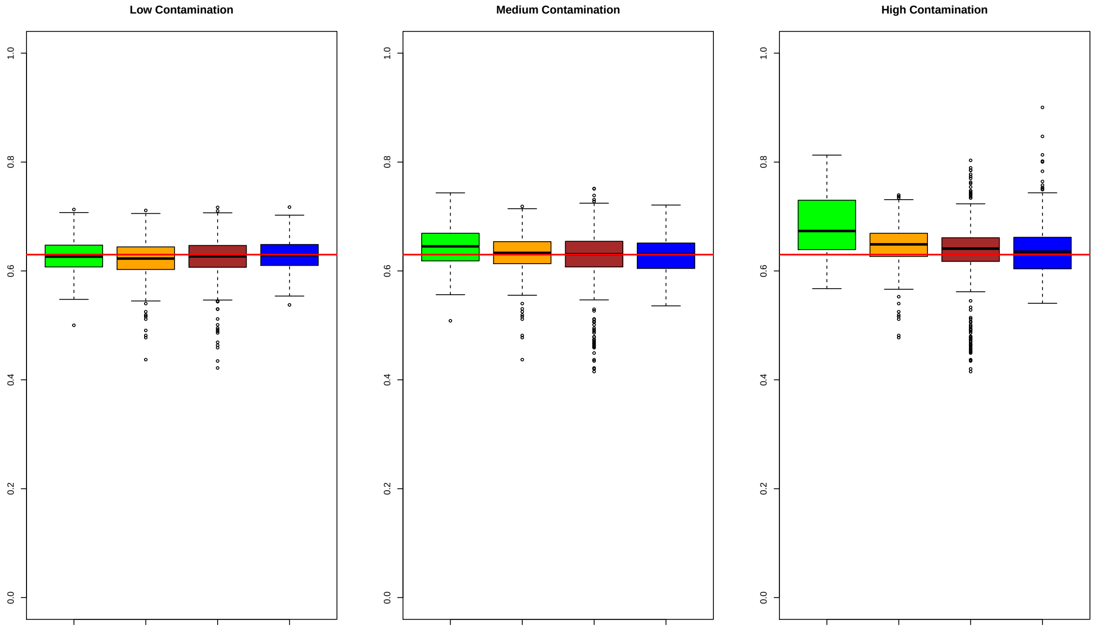


Figure D.31. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 300 and $\sigma_\delta = 0.4$.

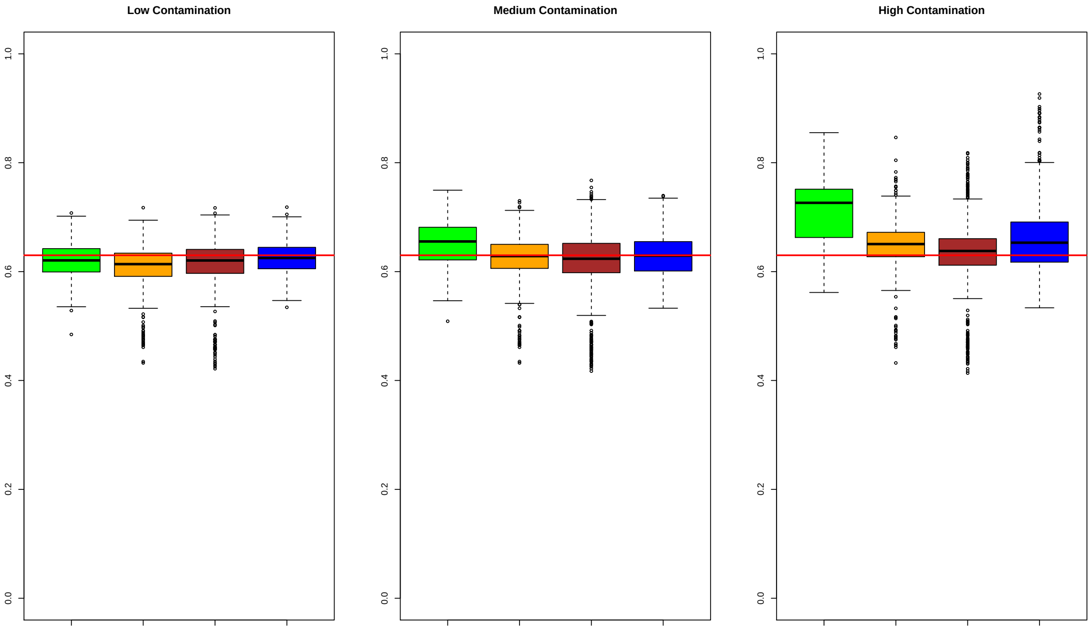


Figure D.32. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 300 and $\sigma_\delta = 0.6$.

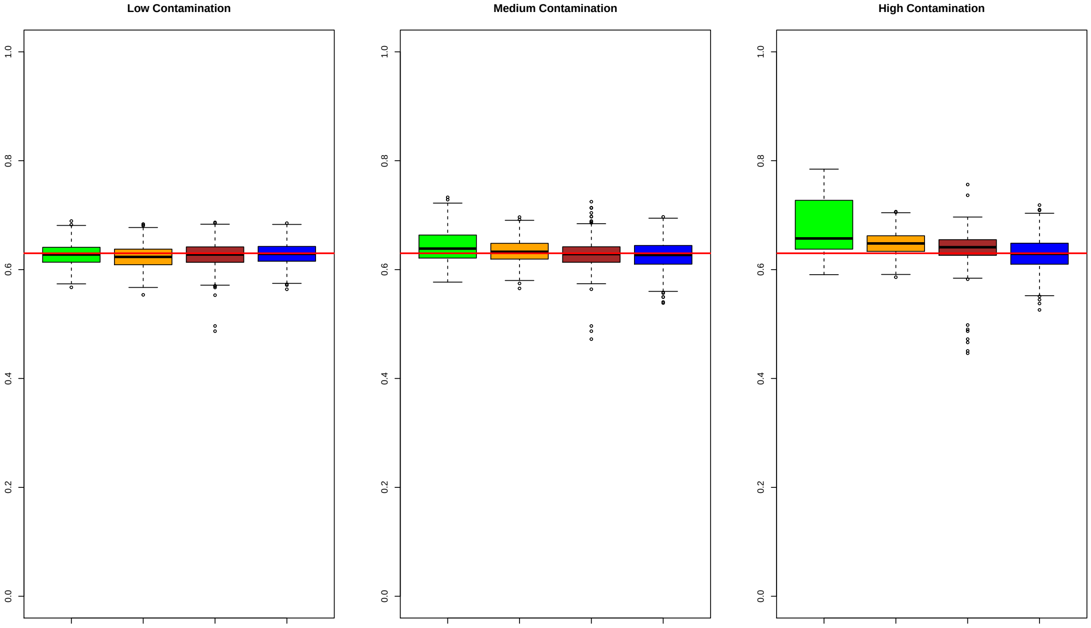


Figure D.33. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 600 and $\sigma_\delta = 0.4$.

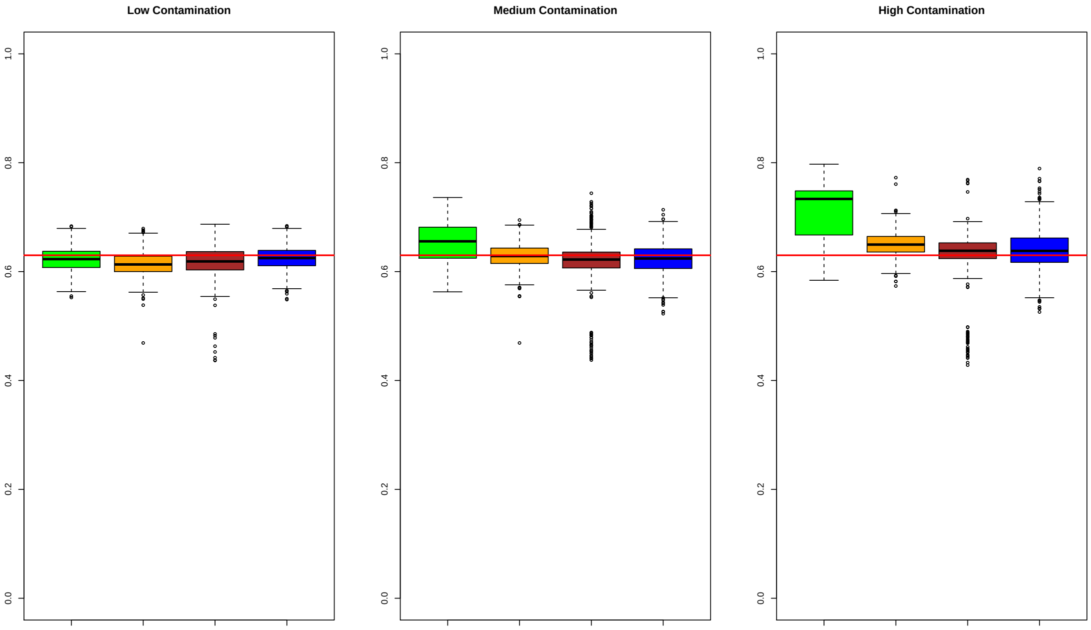


Figure D.34. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 600 and $\sigma_\delta = 0.6$.

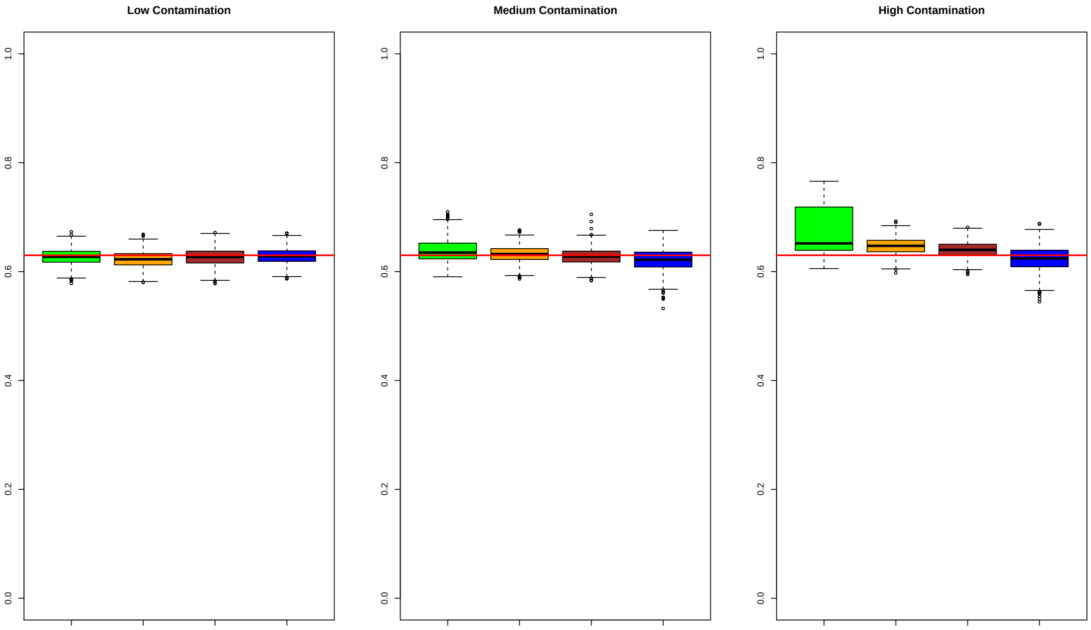


Figure D.35. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.4$.

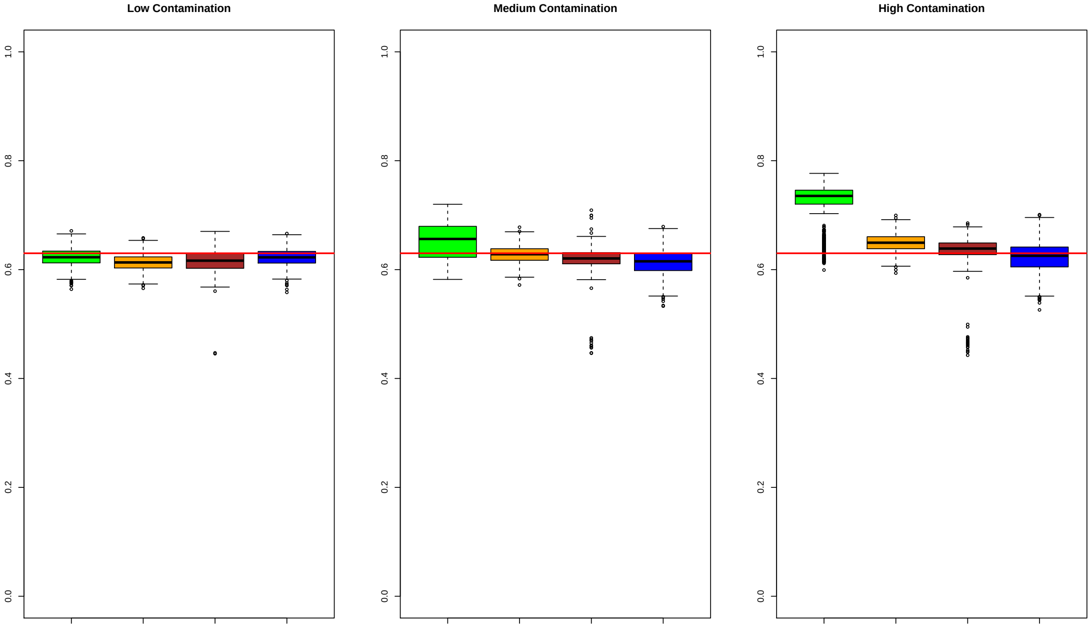


Figure D.36. This set of boxplots contrasts the results of the Prevalence estimates of the compared methods at size 1200 and $\sigma_\delta = 0.6$.

E. MULTIYEAR EXTENSIONS TO SUBSET METHODS

Extension of Turnidge et al. [2006]

Define $\mathbf{N}_{tot}$ as the vector containing the number of observations for each particular year. The cut point in each year can be different. The selection criteria for each year's subset for WT distribution parameter estimation is the same as TURNM; the left-most bins up and to including the mode plus one bin. Denote the the vector of length T of each year's cutpoint as $\mathbf{K}$.

Parameters: $\pi_1, ..., \pi_T, \theta_{\mathbf{WT}}$

1. The presumed wild-type data are pre-selected using the heuristic of choosing one bin to the right of the mode (choose left-most $\mathbf{K}$ bins) for each year.
2. Use non-linear least squares estimation to determine:

$$\arg \min_{\pi_1,...,\pi_T,\theta_{\mathbf{WT}}} \sum_{t=1}^{T} \sum_{j=1}^{K_t} [B_{t,j} - N_{t,tot}\boldsymbol{\pi}_t \cdot F_{WT}(C_{t,j}; \boldsymbol{\theta_{WT}})]^2.$$

Extension of Jaspers et al. [2014a]

Parameters: $\pi_1, ..., \pi_T, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{1,NWT}}, ..., \boldsymbol{\theta_{T,NWT}}$

As both the log2gamma and the Normal distribution with known censoring are members of the regular exponential family. Then they both satisfy regularity conditions. Thus the application of asymptotic theory of maximum likelihood estimates (MLEs) is appropriate. The following procedure is employed:

1. Apply the method of Jaspers et al. [2014a] to each year. This produces a different estimate of $\boldsymbol{\theta_{WT}}$ for each year.
2. For each year's estimate of $\boldsymbol{\theta_{t,WT}}$ to determine the observed Fisher Information using the negative of the Hessian of the log-likelihood. Denote it as $I_t(\hat{\boldsymbol{\theta}}_{\boldsymbol{t,WT}})$

3. For each year's estimate of $\boldsymbol{\theta_{t,WT}}$ use the observed Fisher Information to calculate *approximate* standard errors. Denote this as

$$SE(\hat{\boldsymbol{\theta}}_{t,WT}) = \sqrt{diag(I_t(\hat{\boldsymbol{\theta}}_{t,WT})^{-1})}$$

4. Now calculate the weighted averages of the the estimates for $\hat{\boldsymbol{\theta}}_{overall,WT}$ by using the following formula:

$$\hat{\boldsymbol{\theta}}_{overall,WT} = \frac{\sum_{t=1}^{T} \hat{\boldsymbol{\theta}}_{t,WT} SE(\hat{\boldsymbol{\theta}}_{t,WT})^{-2}}{\sum_{t=1}^{T} SE(\hat{\boldsymbol{\theta}}_{t,WT})^{-2}}.$$

5. Using the estimate, $\hat{\boldsymbol{\theta}}_{overall,WT}$ recalculate the estimate of each year's prevalence.

The method that best allows each year to have a different subset of bins starting from the left is to apply the method of Jaspers et al. [2014a] to each year, estimate the standard errors using the observed Fisher Information, average the estimates.

Alternative Multiyear Extension of Jaspers et al. [2014a]

An alternative is estimating the T multinomials jointly. To satisfy the assumption for changes in the MIC distribution over time, there a single $\boldsymbol{\theta_{WT}}$ among the the T years. Then consider all possible subsets of left-most bins. For each iteration's subsets, the different years WT subset can be effectively pooled together to determine $\hat{\boldsymbol{\theta}}_{\boldsymbol{WT}}$. Then treat the T as scaled components of a "larger" multinomial for the purpose of a single AIC value.

The major drawback to this approach is realistically only (or almost only) models where each year have the same cut-off are considered. For subsets, where that is not the case, the iteration's cut point enforces an inappropriate truncation. This results in misleading counts for each of the considered bins, and in turn poor estimates of $\boldsymbol{\theta_{WT}}$. In contrast the method above is far more accommodating by allowing each year to have a different number of WT bins.

F. MULTIYEAR COMPUTATION

This chapter considers a generalization of the algorithm for the single-year setting. Before there was only one year, now there are T years. The model is detailed in Chapter 4. Like in the single-year case, the R code for the DPMM for each year's NWT distribution is a modification of the *dirichletprocess* package [Ross et al., 2020].

Computation: Normal WT Distribution for Multiple Years

Now for computational ease, three latent vectors are introduced for each year $t = 1, ..., T$. The first is the latent continuous values, $\mathbf{X_t}$. The second is an indicator for whether an observation is a wild-type or not, $\mathbf{c_t}$. The third is for values of the measurement error, $\boldsymbol{\delta_t}$. A Metropolis-within-Gibbs algorithm is used.

For year $t = 1, ..., T$ and for observation i, define

$$\begin{cases} c_{t,i} = & \begin{matrix} 1 & if\, WT \\ 0 & else \end{matrix} \end{cases}$$

Let $\boldsymbol{\theta_{t,NWT}} = (\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_k})$. Let $\mathbf{S_t}$ be a vector denoting the allocations of each NWT observation to cluster $1, ..., k_t$. Then $\mathbf{S}_{t,(-i)}$ the vector S_t without observation i. The vector $\mathbf{n_t}$ denotes the number of observations in each of the k_t NWT components.

The superscript (L) denotes the L^{th} iteration. First initial values are determined, the $(0)^{\text{th}}$ values. Note that for the generation of latent vectors only interval censored generation is shown for brevity. As the endpoints of the data set are both known and fixed, the censoring is adjusted accordingly. Denote $\tilde{\mathbf{X}} = (\mathbf{X_1}, ..., \mathbf{X_T})$, $\tilde{\mathbf{c}} = (\mathbf{c_1}, ..., \mathbf{c_T})$, and $\tilde{\boldsymbol{\delta}} = (\boldsymbol{\delta_1}, ..., \boldsymbol{\delta_T})$

Update $\mu_{WT} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{c}}$

- ${\mu_{WT}}^{(L)} \mid {\sigma_{WT}}^{2(L-1)}, \tilde{\mathbf{X}}^{(L-1)}, \tilde{\mathbf{c}}^{(L-1)} \sim N(\bar{\tilde{X}}^{(L-1)}, \sqrt{\frac{{\sigma_{WT}}^{2(L-1)}}{\sum_i \tilde{c}_i^{(L-1)}}})$

 Jointly update σ^2_{WT} and σ^2_δ

- $P({\sigma^2_{WT}}^{(L)}, \sigma^{2(L)}_\delta \mid {\mu_{WT}}^{(L)}, {\sigma_{WT}}^{2(L-1)}, \sigma^{2(L-1)}_\delta, \tilde{\mathbf{X}}^{(L-1)}, \tilde{\mathbf{c}}^{(L-1)}, \tilde{\boldsymbol{\delta}}^{(L-1)} \propto$
 $P(\tilde{\mathbf{X}}^{(L-1)} \mid \mu^{(L)}_{WT}, {\sigma_{WT}}^{2(L-1)}, \tilde{\mathbf{c}}^{(L-1)}) P(\boldsymbol{\delta}^{(\tilde{\boldsymbol{L}}\mathbf{-1})} \mid \sigma^{2(L-1)}_\delta) pr({\sigma_{WT}}^{2(L-1)}) pr(\sigma^{2(L-1)}_\delta)$

For $t = 1, ..., T$, update $\pi_t \mid \mathbf{c_t}$

- ${\pi_t}^{(L)} \mid {\mathbf{c_t}}^{(L-1)} \sim TrBeta(a = .5, b = 1, \alpha' = 1 + \sum_i {c_{t,i}}^{(L-1)}, \beta' = 1 + N_{t,tot} - \sum_i c^{(L-1)}_{t,i})$

For $t = 1, ..., T$, update $\boldsymbol{\theta_{t,NWT}} \mid \mathbf{X_t}, \mathbf{c_t}$.

- Use Neal [2000] Algorithm 8 with D auxiliary classes. For $i = 1, ..., N_{t,tot} - \sum_i c_{t,i}$, ${\mathbf{S_t}}^{(L)} \mid {\alpha_{t,conc}}^{(L-1)}, {\boldsymbol{\theta_{t,NWT}}}^{(L-1)}, {\mathbf{X_t}}^{(L-1)}, {\mathbf{c_t}}^{(L-1)}, {\mathbf{S_{t,-i}}}^{(L-1)}$

For $h = 1, ...k$,

- ${\mu_{t,h}}^{(L)} \mid {\mathbf{S_t}}^{(L)}, {\boldsymbol{\sigma_{t,h}}}^{(L-1)}, {\mathbf{X_t}}^{(L-1)}, {\mathbf{c_t}}^{(L-1)} \sim TrN(A, \infty, \mu'_t, \sqrt{\sigma^{2'}_t})$

where

$$\mu'_t = \frac{\frac{{\mu_{G_{t,0}}}^{(L-1)}}{\sigma^2_0} + \frac{\sum_{i:{S_t}^{(L)}{}_i = h} {X_t}_i^{(L-1)}}{{\sigma_{t,h}}^{2(L-1)}}}{\frac{1}{{\sigma_0}^2} + \frac{{n_{t,h}}^{(L)}}{{\sigma^2_{t,h}}^{(L-1)}}}$$

and

$$\sigma^{2'} = \frac{1}{\frac{1}{{\sigma_0}^2} + \frac{n_{t,h}}{{\sigma^2_{t,h}}^{(L-1)}}}$$

- ${\sigma^2_{t,h}}^{(L)} \mid {\mathbf{S_t}}^{(L)}, {\mathbf{X_t}}^{(L)}, {\mu_{t,h}}^{(L)}, {\mathbf{c_t}}^{(L-1)} \sim InvGamma(\alpha^{'} = \alpha_1 + \frac{n_{t,h}}{2}, \beta' = \beta_1 + \frac{\sum_{i:S_{t,i}=h} (X_{t,i} - \mu_{t,h})^2}{2})$
- ${w_{t,h}}^{(L)} = \frac{{n_{t,h}}^{(L)}}{N_{t,tot}(1 - {c_t}^{(\hat{L}-1)})}$

For $t = 1, ..., T$, update the mean of the base distribution.

- ${\mu_{G_{t,0}}}^{(L)} \mid {\boldsymbol{\mu_{t,NWT}}}^{(L)} \sim TrN(A, \infty, \mu^{'}_{G_{t,0}} = \frac{\frac{\mu_{G_B}}{{\sigma_{G_B}}^2} + \frac{\sum_r {\mu_{t,r}}^{(L)}}{\sigma^2_0}}{\frac{1}{{\sigma_{G_B}}^2} + \frac{k^{(L)}_t}{{\sigma_0}^2}}, \sigma^{'}_{G_{t,0}} = (\frac{1}{\sigma^2_{G_B}} + \frac{k^{(L)}_t}{\sigma^2_0})^{-1/2})$

For $t = 1, ..., T$, update $\alpha_{t,conc}$ using West [1992] where γ is the Euler-Mascheroni constant.

- ${\alpha_{t,conc}}^{(L)} \mid k_t^{(L)}, {\mathbf{c_t}}^{(L-1)} \overset{asy.}{\sim} Gamma(a + k_t^{(L)} - 1, b + \gamma + \log(N_{t,tot} - \sum_i c_{t,i}^{(L-1)}))$

For $t = 1, ..., T$, update $\mathbf{X_t} \mid \mathbf{Y_t}, \mathbf{c_t}, \boldsymbol{\delta_t}$.

For the observations, $i = 1, ..., N_{t,tot}$ where $c_{t,i} = 1$

- ${\mathbf{X_t}}^{(L)} \mid \mathbf{Y_t}, {\mathbf{c_t}}^{(L-1)}, {\mu_{WT}}^{(L)}, {\sigma_{WT}}^{(L)}, {\boldsymbol{\delta_t}}^{(L-1)} \overset{ind.}{\sim} TrN(Y_t - 1 - {\delta_t}^{(L-1)}, Y - {\delta_t}^{(L-1)}, {\mu_{WT}}^{(L)}, {\sigma_{WT}}^{(L)})$

For $t = 1, ..., T$, for the observations, $i = 1, ..., N_{t,tot}$ where $c_{t,i} = 0$

- ${\mathbf{X_t}}^{(L)} \mid \mathbf{Y_t}, {\mathbf{S_t}}^{(L)}, {\mathbf{c_t}}^{(L)}, {\boldsymbol{\mu_{t,NWT}}}^{(L)}, {\boldsymbol{\sigma_{t,NWT}}}^{(L)} \overset{ind.}{\sim}$
  $TrN(\mathbf{Y_t} - 1 - {\boldsymbol{\delta_t}}^{(L-1)}, Y - {\boldsymbol{\delta_t}}^{(L-1)}, \boldsymbol{\mu_{t,NWT}}_{i:{S_{t,i}}^{(L)}}, {\boldsymbol{\sigma_{t,NWT}}_{i:{S_{t,i}}^{(L)}}}^{(L)})$

For $t = 1, ..., T$, update $\boldsymbol{\delta_t} \mid \mathbf{Y_t}, \mathbf{X_t}, \sigma_\delta^2$.

- ${\boldsymbol{\delta_t}}^{(L)} \mid \mathbf{Y_t}, {\mathbf{X_t}}^{(L)}, {\mathbf{c_t}}^{(L-1)}, {\mathbf{S_t}}^{(L)}, {\sigma_{\delta_t}}^{2(L)} \overset{ind.}{\sim}$
  $TrN(Y_t - 1 - X_t^{(L-1)}, Y_t - X_t^{(L-1)}, \mu = 0, {\sigma_\delta}^{(L)})$

For $t = 1, ..., T$, update $\mathbf{c_t} \mid \mathbf{X_t}, \pi_t, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{t,NWT}}$

- ${\mathbf{c_t}}^{(L)} \mid {\mathbf{X_t}}^{(L)}, \pi_t^{(L)}, {\boldsymbol{\theta_{WT}}}^{(L)}, {\boldsymbol{\theta_{t,NWT}}}^{(L)} \sim Bernoulli(p'_t)$

where

$$p'_t = \frac{\frac{{\pi_t}^{(L)}}{{\sigma_{WT}}^{(L)}}\phi(\frac{X_t^{(L)} - \mu_{WT}^{(L)}}{{\sigma_{WT}}^{(L)}})}{\frac{{\pi_t}^{(L)}}{{\sigma_{WT}}^{(L)}}\phi(\frac{{X_t}^{(L)} - {\mu_{WT}}^{(L)}}{{\sigma_{WT}}^{L}}) + (1 - {\pi_t}^{(L)})\sum_{h=1}^{k_t^{(L)}} \frac{{w_{t,h}}^{(L)}}{{\sigma_{t,h}}^{(L)}}\phi(\frac{{X_t}^{(L)} - {\mu_{t,h}}^{(L)}}{{\sigma_{t,h}}^{(L)}})}$$

Then increase $L \leftarrow L + 1$

Computation: log2gamma WT Distribution for Multiple Years

Update $\boldsymbol{\theta_{WT}} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{c}}$

- $P(\alpha^{(L)} \mid \beta^{(L-1)}, \tilde{\mathbf{X}}^{(L-1)}, \tilde{\mathbf{c}}^{(L-1)}) \propto P(\tilde{\mathbf{X}}^{(L-1)} \mid \tilde{\mathbf{c}}^{(L-1)}, \alpha^{(L-1)}, \beta^{(L-1)})pr(\alpha^{(L-1)})$
- $\beta^{(L)} \mid \tilde{\mathbf{X}}^{(L-1)}, \tilde{\mathbf{c}}^{(L-1)}, \alpha^{(L)} \sim Gamma(\alpha' = \sum_t {N_{t,tot}}^{\tilde{}\,(\bar{L}-1)} c\alpha^{(L)}, \beta' = \sum_{i:\tilde{c}_i^{(L)}=1} 2^{X_i^{(L)}})$

Update $\sigma_\delta^2 \mid \tilde{\boldsymbol{\delta}}$

- ${\sigma_\delta}^{2(L)} \mid \tilde{\delta}^{(L-1)} \sim InvGamma(\alpha' = \alpha_\delta + \frac{\sum_t N_{t,tot}}{2}, \beta' = \beta_\delta + \frac{\sum_{t,i} {{\delta_{t,i}}^{(L)}}^2}{2})$

For $t = 1, ..., T$, update $\pi_t \mid \mathbf{c_t}$

- ${\pi_t}^{(L)} \mid {\mathbf{c_t}}^{(L-1)} \sim TrBeta(a = .5, b = 1, \alpha' = 1 + \sum_i {c_{t,i}}^{(L-1)}, \beta' = 1 + N_{t,tot} - \sum_i c_{t,i}^{(L-1)})$

For $t = 1, ..., T$, update $\boldsymbol{\theta_{t,NWT}} \mid \mathbf{X_t}, \mathbf{c_t}$.

- Use Neal [2000] Algorithm 8 with D auxiliary classes. For $i = 1, ..., N_{t,tot} - \sum_i c_{t,i}$, ${\mathbf{S_t}}^{(L)} \mid {\alpha_{t,conc}}^{(L-1)}, {\boldsymbol{\theta_{t,NWT}}}^{(L-1)}, {\mathbf{X_t}}^{(L-1)}, {\mathbf{c_t}}^{(L-1)}, {\mathbf{S_{t,-i}}}^{(L-1)}$

For $h = 1, ...k$,

- ${\mu_{t,h}}^{(L)} \mid {\mathbf{S_t}}^{(L)}, {\boldsymbol{\sigma_{t,h}}}^{(L-1)}, {\mathbf{X_t}}^{(L-1)}, {\mathbf{c_t}}^{(L-1)} \sim TrN(A, \infty, \mu'_t, \sqrt{\sigma_t^{2'}})$

where

$$\mu'_t = \frac{\frac{{\mu_{G_{t,0}}}^{(L-1)}}{\sigma_0^2} + \frac{\sum_{i:{S_t}^{(L)}{}_i=h} {X_t}_i^{(L-1)}}{{\sigma_{t,h}}^{2(L-1)}}}{\frac{1}{{\sigma_0}^2} + \frac{{n_{t,h}}^{(L)}}{{\sigma_{t,h}^2}^{(L-1)}}}$$

and

$$\sigma^{2'} = \frac{1}{\frac{1}{{\sigma_0}^2} + \frac{n_{t,h}}{{\sigma_{t,h}^2}^{(L-1)}}}$$

- ${\sigma_{t,h}^2}^{(L)} \mid {\mathbf{S_t}}^{(L)}, {\mathbf{X_t}}^{(L)}, {\mu_{t,h}}^{(L)}, {\mathbf{c_t}}^{(L-1)} \sim InvGamma(\alpha^{'} = \alpha_1 + \frac{n_{t,h}}{2}, \beta' = \beta_1 + \frac{\sum_{i:S_{t,i}=h} (X_{t,i} - \mu_{t,h})^2}{2})$
- ${w_{t,h}}^{(L)} = \frac{{n_{t,h}}^{(L)}}{N_{t,tot}(1 - {c_t}^{(\hat{L}-1)})}$

For $t = 1, ..., T$, update the mean of the base distribution.

- ${\mu_{G_{t,0}}}^{(L)} \mid {\boldsymbol{\mu_{t,NWT}}}^{(L)} \sim TrN(A, \infty, \mu^{'}_{G_{t,0}} = \frac{\frac{\mu_{G_B}}{{\sigma_{G_B}}^2} + \frac{\sum_r {\mu_{t,r}}^{(L)}}{\sigma_0^2}}{\frac{1}{{\sigma_{G_B}}^2} + \frac{k_t^{(L)}}{{\sigma_0}^2}}, \sigma^{'}_{G_{t,0}} = (\frac{1}{\sigma_{G_B}^2} + \frac{k_t^{(L)}}{\sigma_0^2})^{-1/2})$

For $t = 1, ..., T$, update $\alpha_{t,conc}$ using West [1992] where γ is the Euler-Mascheroni constant.

- ${\alpha_{t,conc}}^{(L)} \mid k_t^{(L)}, {\mathbf{c_t}}^{(L-1)} \overset{asy.}{\sim} Gamma(a + k_t^{(L)} - 1, b + \gamma + \log(N_{t,tot} - \sum_i {c_{t,i}}^{(L-1)}))$

For $t = 1, ..., T$, update $\mathbf{X_t} \mid \mathbf{Y_t}, \mathbf{c_t}, \boldsymbol{\delta_t}$.

For the observations, $i = 1, ..., N_{t,tot}$ where $c_{t,i} = 1$

- ${\mathbf{X_t}}^{(L)} \mid \mathbf{Y_t}, {\mathbf{c_t}}^{(L)}, {\boldsymbol{\delta_t}}^{(L)}, \alpha^{(L)}, \beta^{(L)} \sim log_2(TrGamma(a = 2^{Y_t - 1 - \delta_t^{(L)}}, b = 2^{Y_t - \delta_t^{(L)}}, \alpha^{(L)}, \beta^{(L)}))$

For $t = 1, ..., T$, for the observations, $i = 1, ..., N_{t,tot}$ where $c_{t,i} = 0$

- $\mathbf{X_t}^{(L)} \mid \mathbf{Y_t}, \mathbf{S_t}^{(L)}, \mathbf{c_t}^{(L)}, \boldsymbol{\mu_{t,NWT}}^{(L)}, \boldsymbol{\sigma_{t,NWT}}^{(L)} \overset{ind.}{\sim}$ $TrN(\mathbf{Y_t} - 1 - \boldsymbol{\delta_t}^{(L-1)}, Y - \boldsymbol{\delta_t}^{(L-1)}, \boldsymbol{\mu_{t,NWT}}_{i:S_{t,i}{}^{(L)}}, \boldsymbol{\sigma_{t,NWT}}_{i:S_{t,i}{}^{(L)}}{}^{(L)})$

For $t = 1, ..., T$, update $\boldsymbol{\delta_t} \mid \mathbf{Y_t}, \mathbf{X_t}, \sigma^2_\delta$.

- $\boldsymbol{\delta_t}^{(L)} \mid \mathbf{Y_t}, \mathbf{X_t}^{(L)}, \mathbf{c_t}^{(L-1)}, \mathbf{S_t}^{(L)}, \sigma_{\delta_t}{}^{2(L)} \overset{ind.}{\sim}$ $TrN(Y_t - 1 - X_t^{(L-1)}, Y_t - X_t^{(L-1)}, \mu = 0, \sigma_\delta{}^{(L)})$

For $t = 1, ..., T$, update $\mathbf{c_t} \mid \mathbf{X_t}, \pi_t, \boldsymbol{\theta_{WT}}, \boldsymbol{\theta_{t,NWT}}$

- $\mathbf{c_t}^{(L)} \mid \mathbf{X_t}^{(L)}, \pi_t^{(L)}, \boldsymbol{\theta_{WT}}^{(L)}, \boldsymbol{\theta_{t,NWT}}^{(L)} \sim Bernoulli(p'_t)$

where

$$p'_t = \frac{\pi_t^{(L)} f_{l2g}(X_t^{(L)}; \alpha^{(L)}, \beta^{(L)})}{\pi_t{}^{(L)} f_{l2g}(X_t^{(L-1)}; \alpha^{(L)}, \beta^{(L)}) + (1 - \pi_t^{(L)}) \sum_{h=1}^{k_t^{(L)}} \frac{w_h{}^{(L)}}{\sigma_h{}^{(L)}} \phi(\frac{X_t^{(L-1)} - \mu_h{}^{(L)}}{\sigma_h{}^{(L)}})}$$

Then increase $L \leftarrow L + 1$

G. MULTIYEAR PREVALENCE ESTIMATION RESULTS

Normal Distribution

Slow Rate

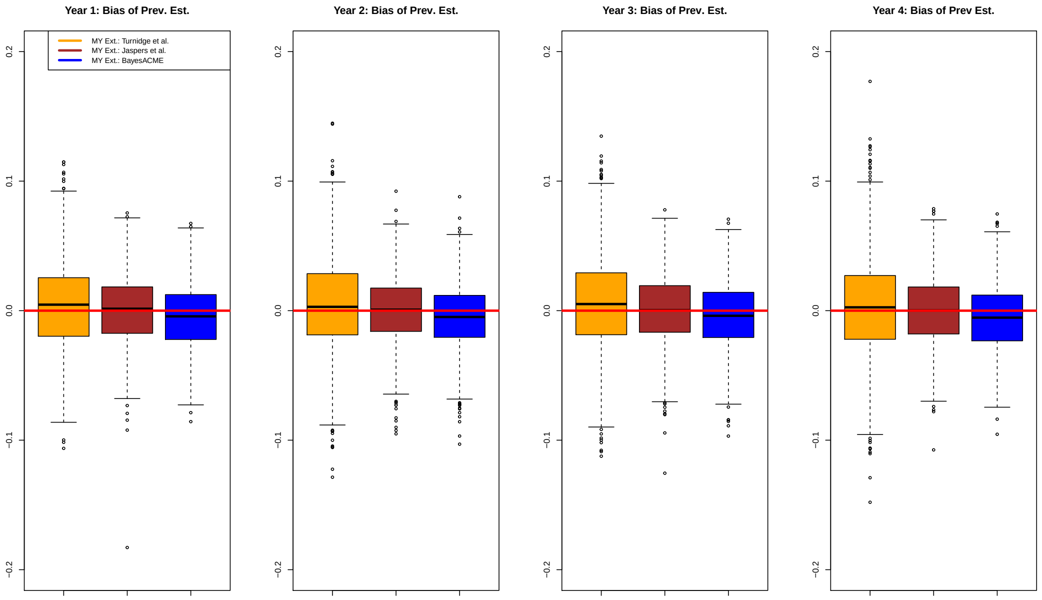


Figure G.1. Normal with Low Contamination at Size 300

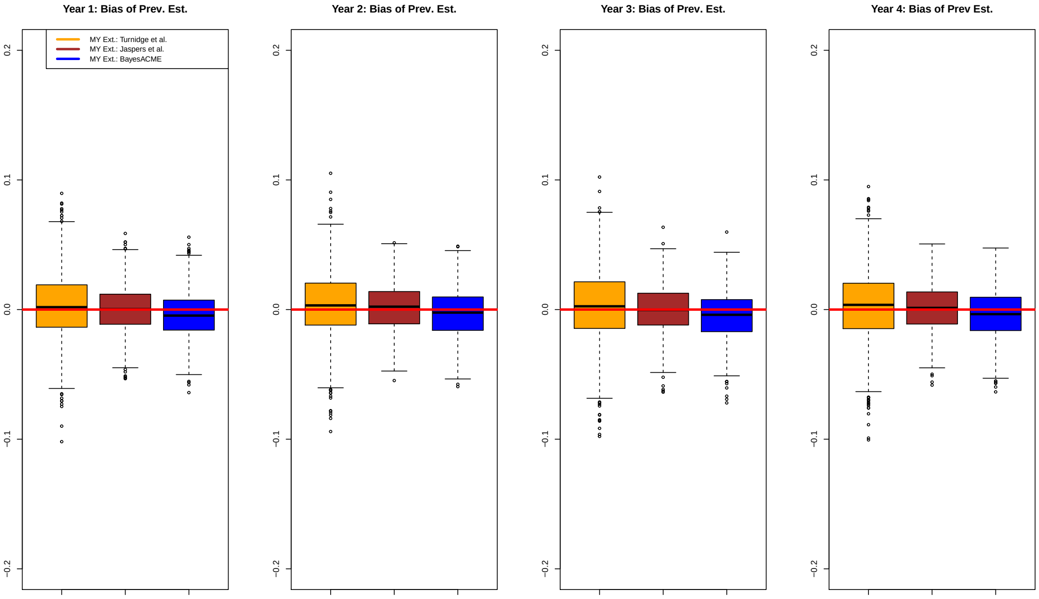


Figure G.2. Normal with Low Contamination at Size 600

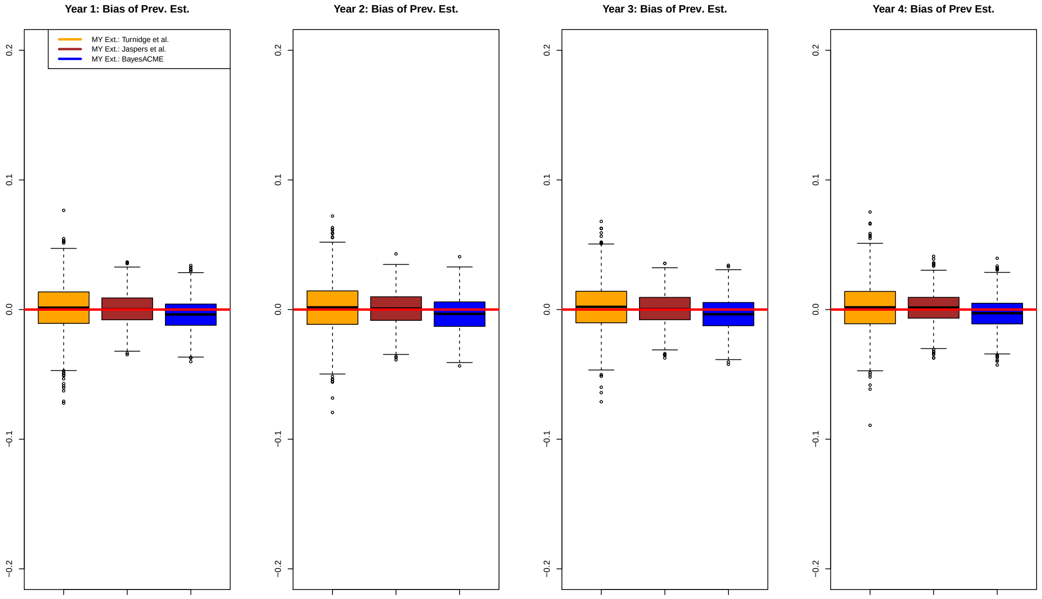


Figure G.3. Normal with Low Contamination at Size 1200

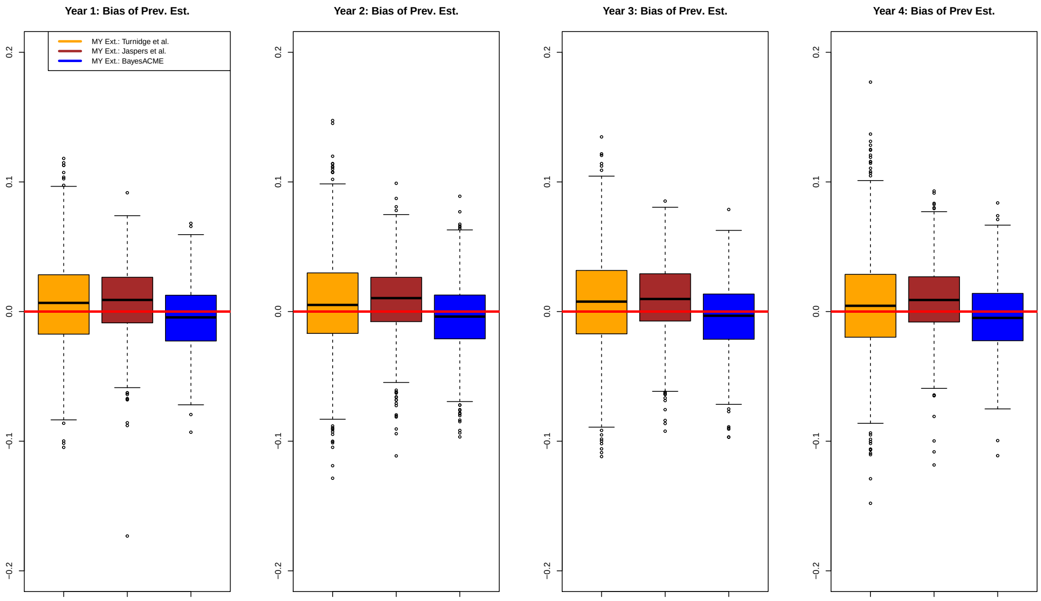


Figure G.4. Normal with Medium Contamination at Size 300

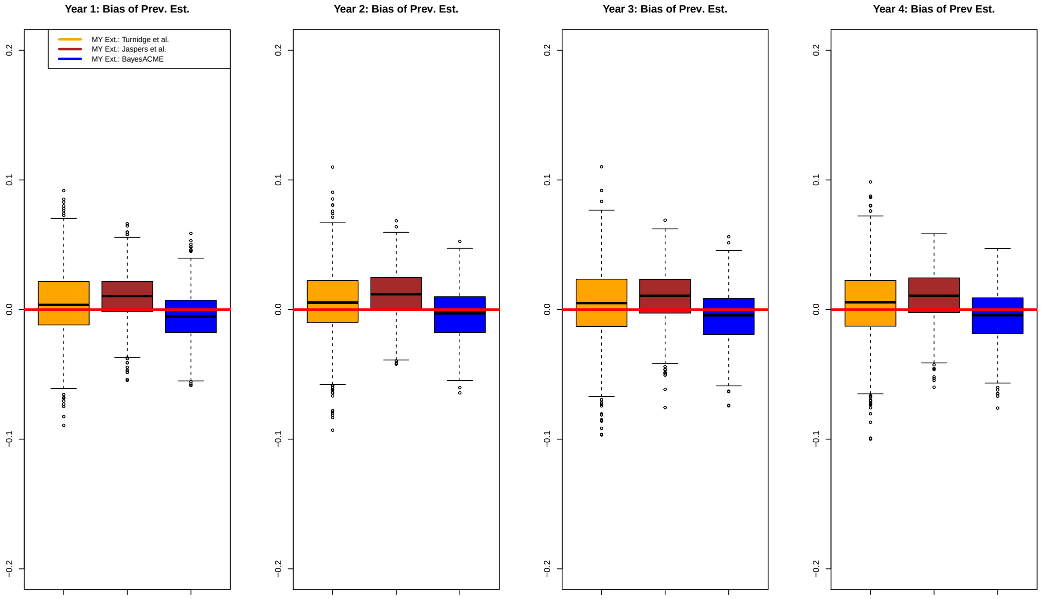


Figure G.5. Normal with Medium Contamination at Size 600

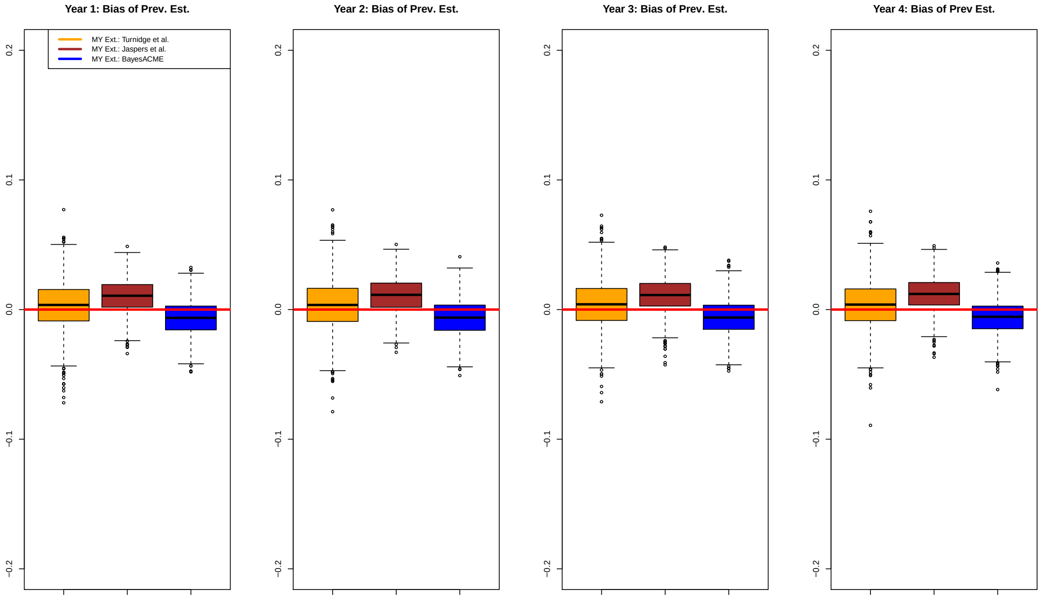


Figure G.6. Normal with Medium Contamination at Size 1200

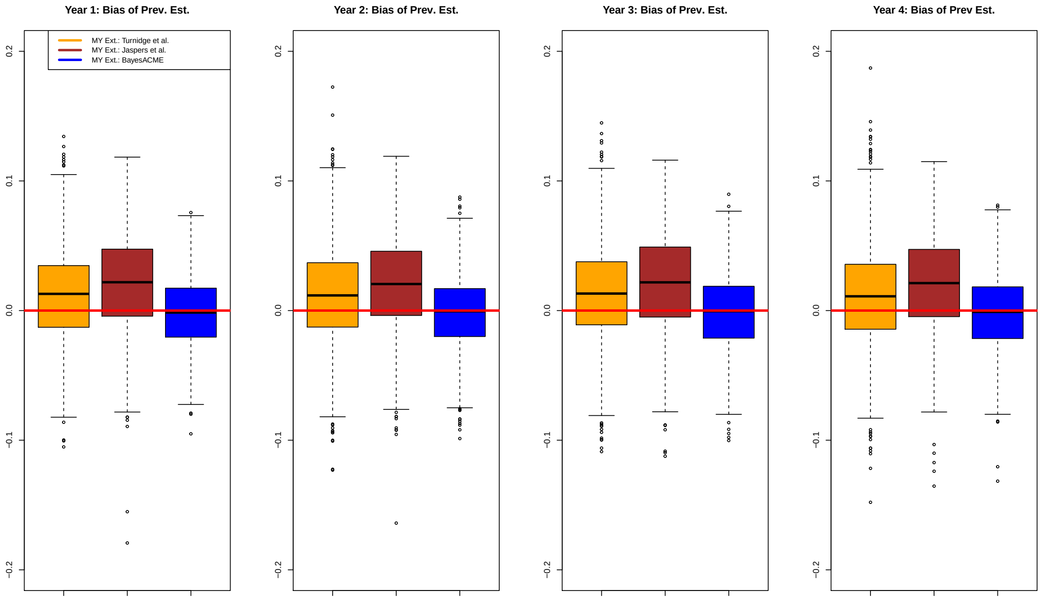


Figure G.7. Normal with High Contamination at Size 300

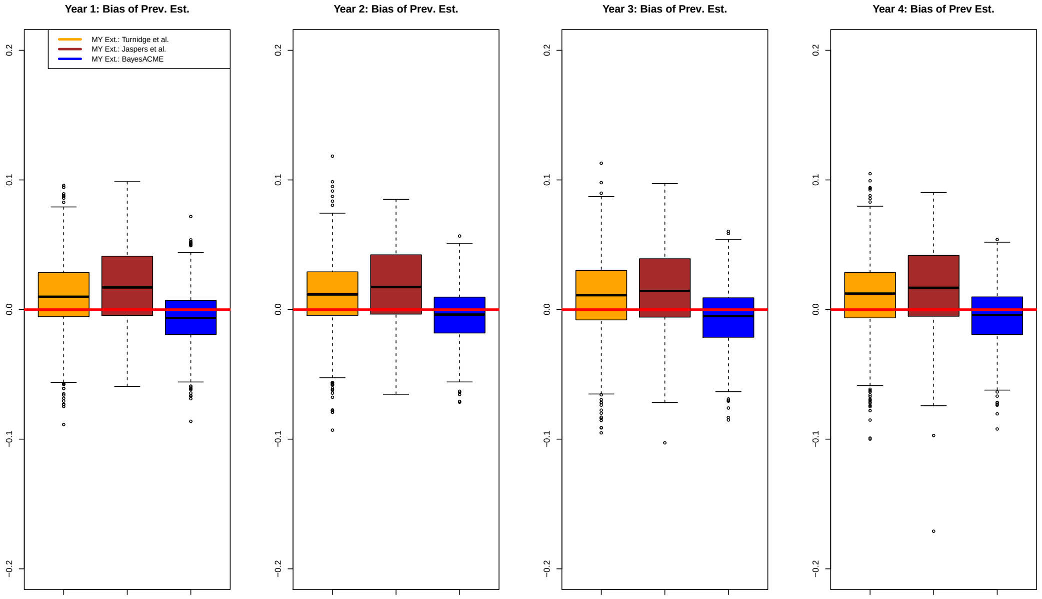


Figure G.8. Normal with High Contamination at Size 600

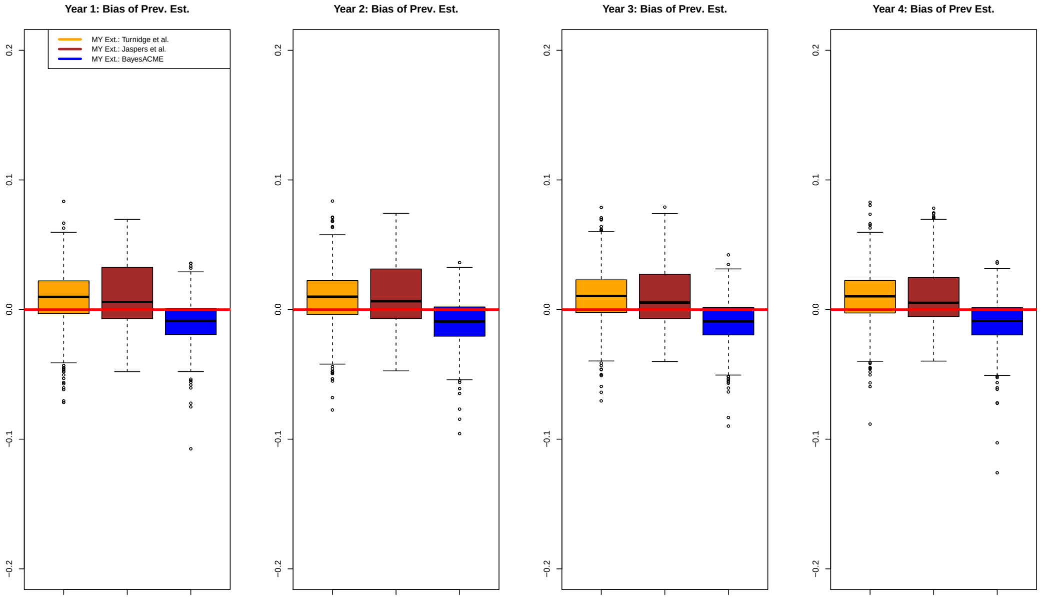


Figure G.9. Normal with High Contamination at Size 1200

Normal: Slow Trend

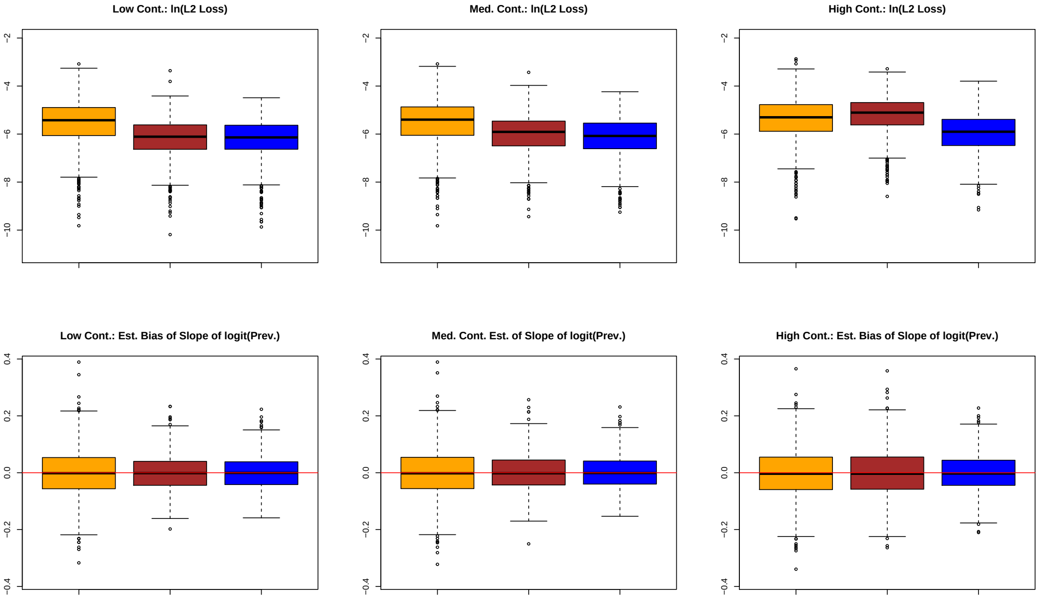


Figure G.10. Trend for Normal case with Slow Rate of Decline at size 300

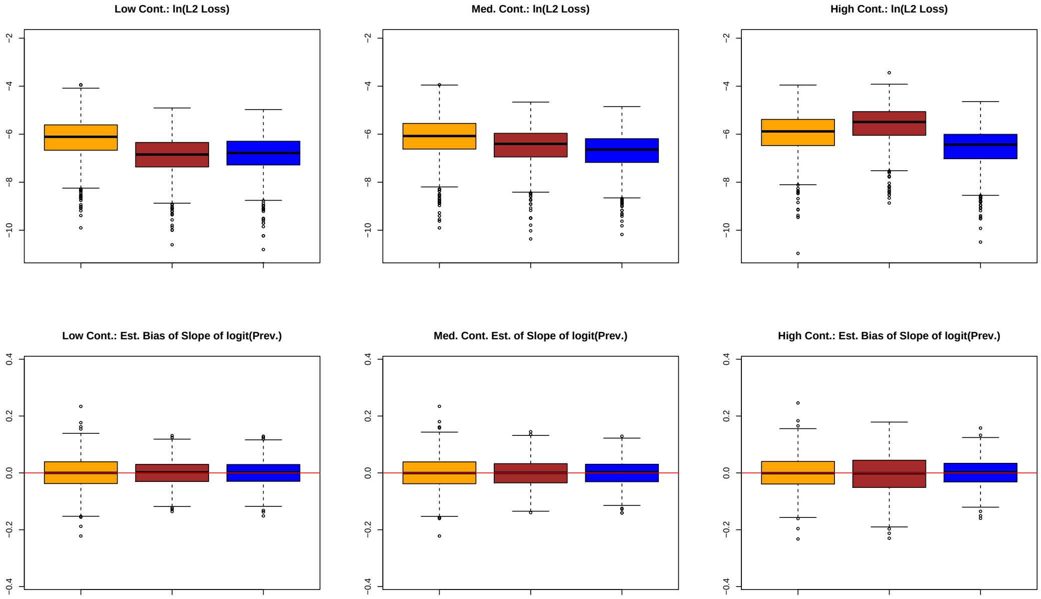


Figure G.11. Trend for Normal case with Slow Rate of Decline at size 600

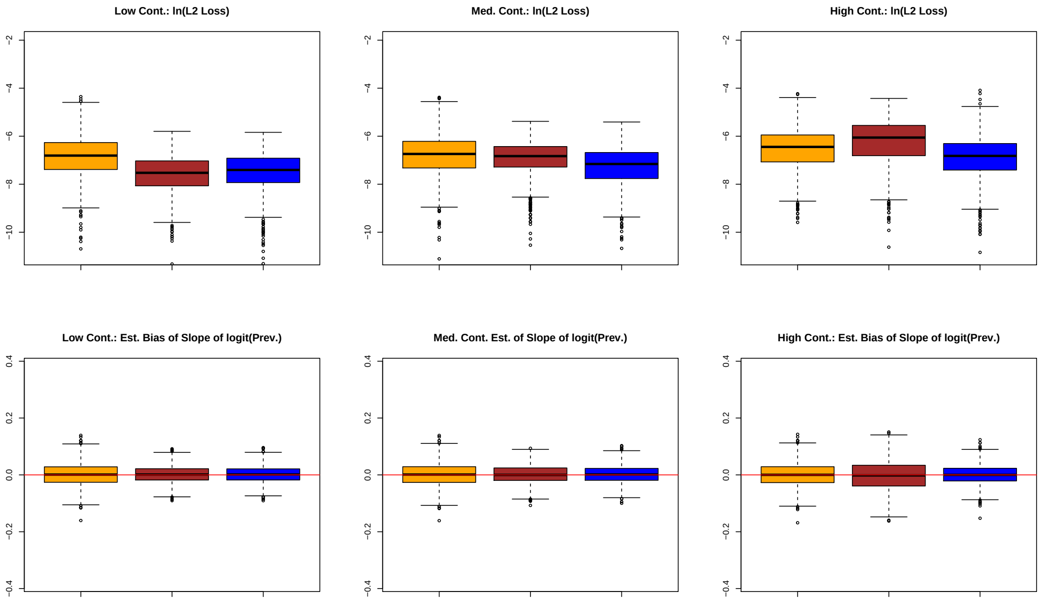


Figure G.12. Trend for Normal case with Slow Rate of Decline at size 1200

Fast Rate

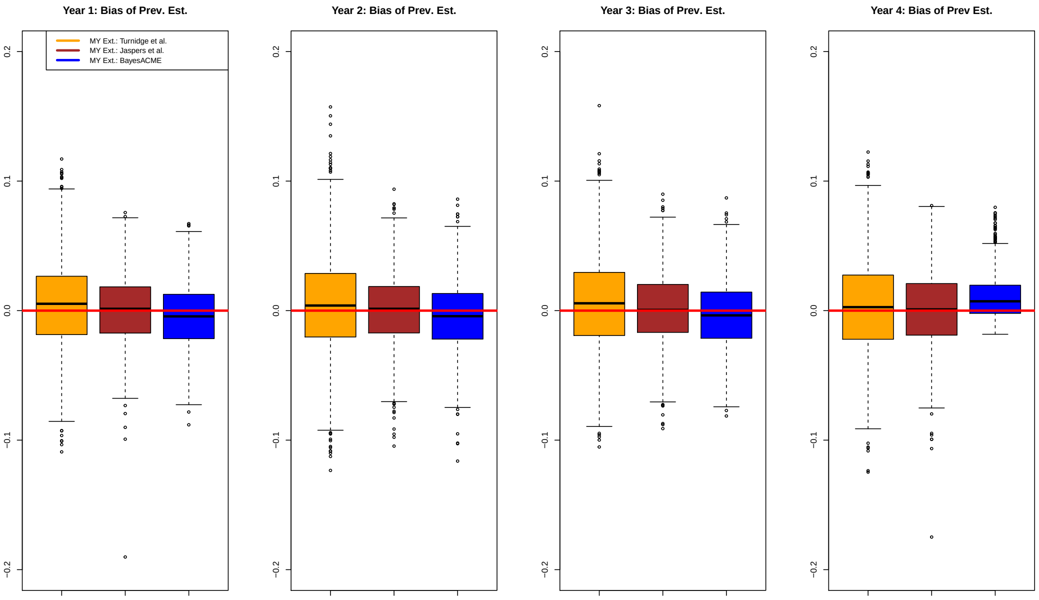


Figure G.13. Normal with Low Contamination at Size 300

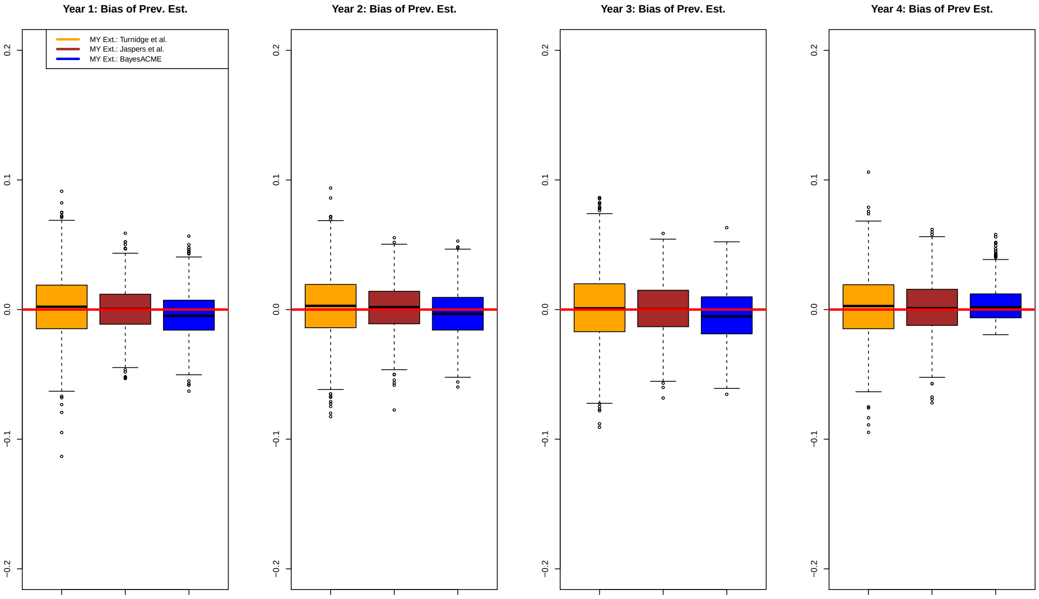


Figure G.14. Normal with Low Contamination at Size 600

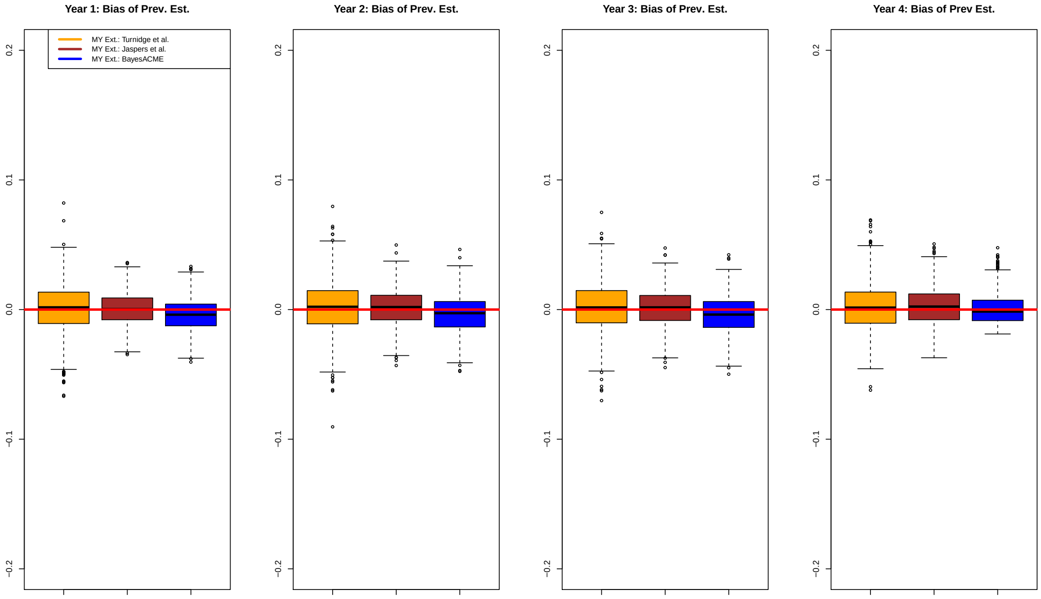


Figure G.15. Normal with Low Contamination at Size 1200

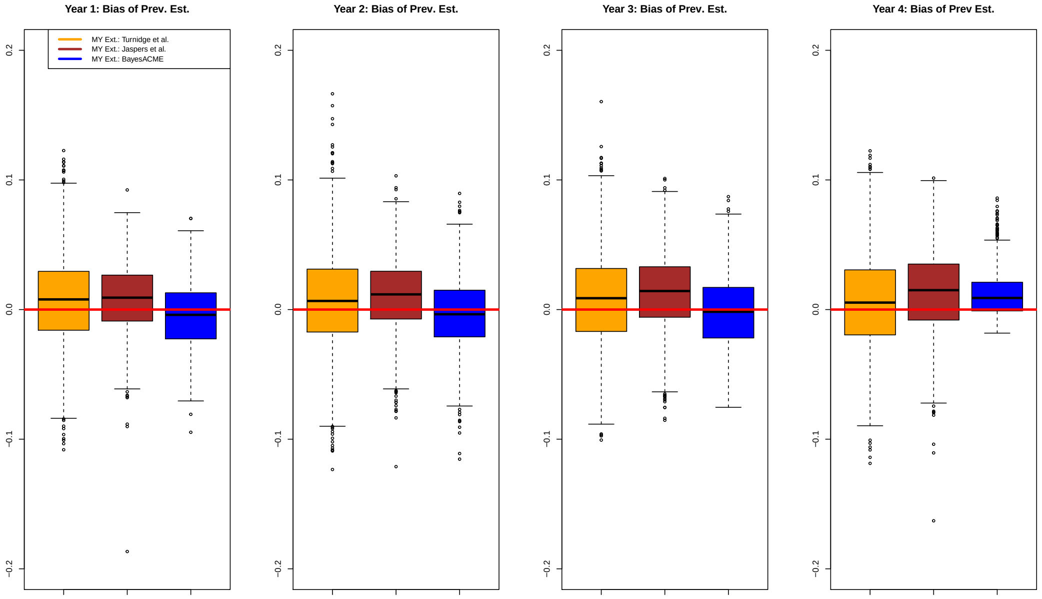


Figure G.16. Normal with Medium Contamination at Size 300

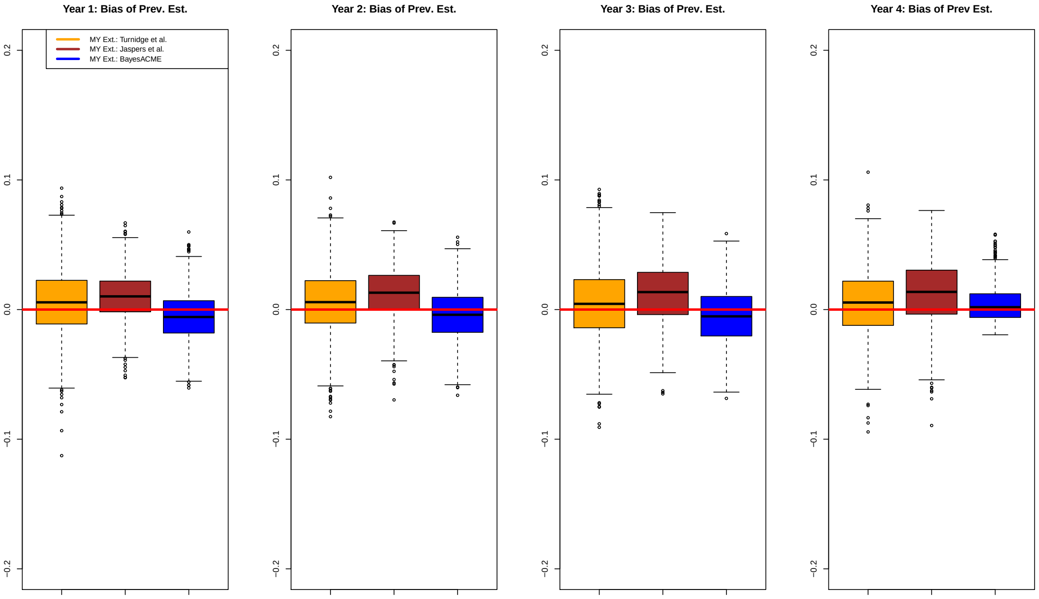


Figure G.17. Normal with Medium Contamination at Size 600

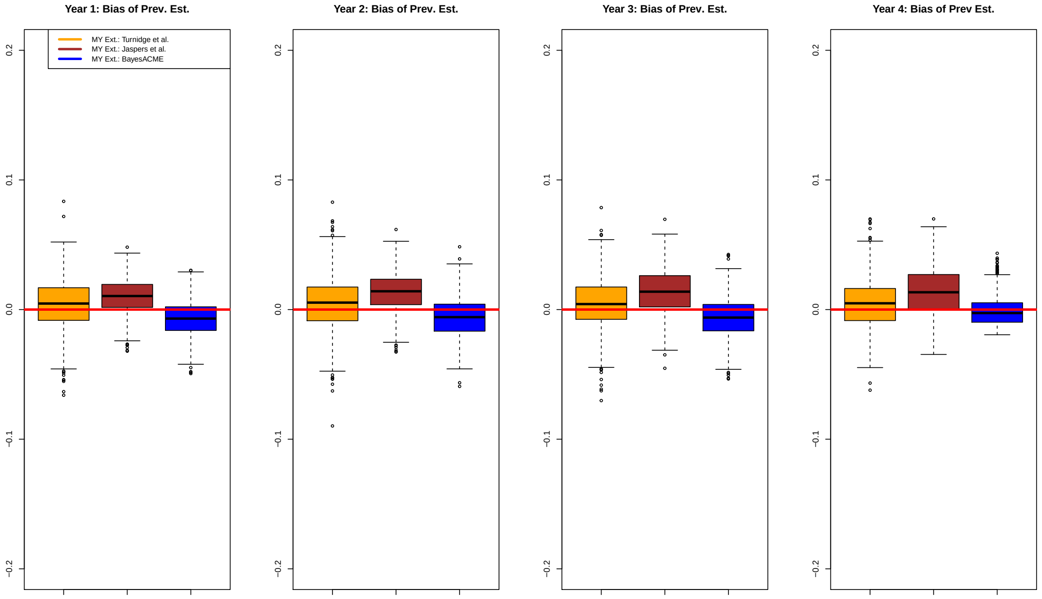


Figure G.18. Normal with Medium Contamination at Size 1200

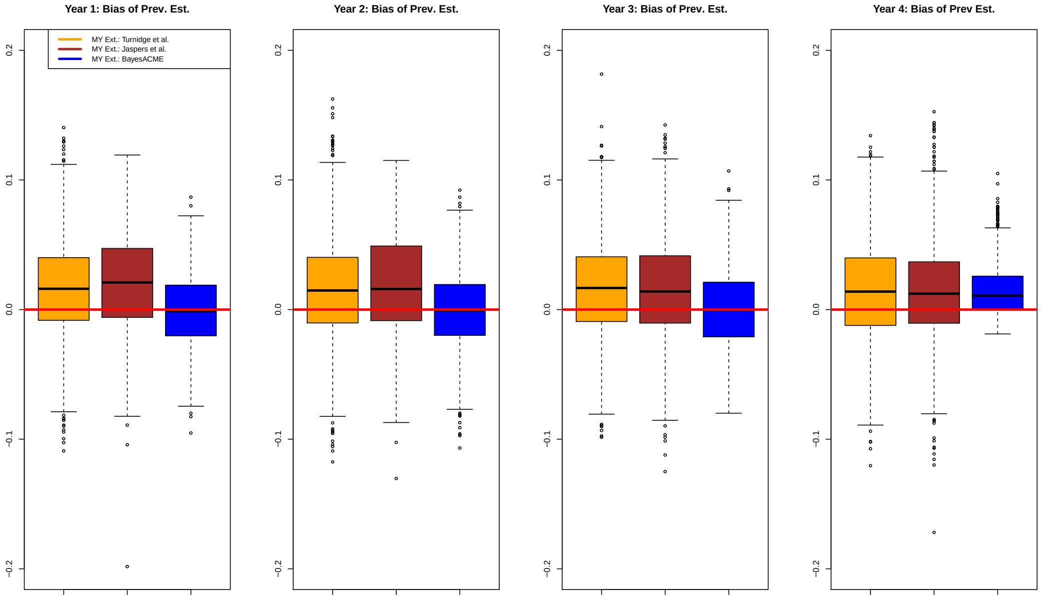


Figure G.19. Normal with High Contamination at Size 300

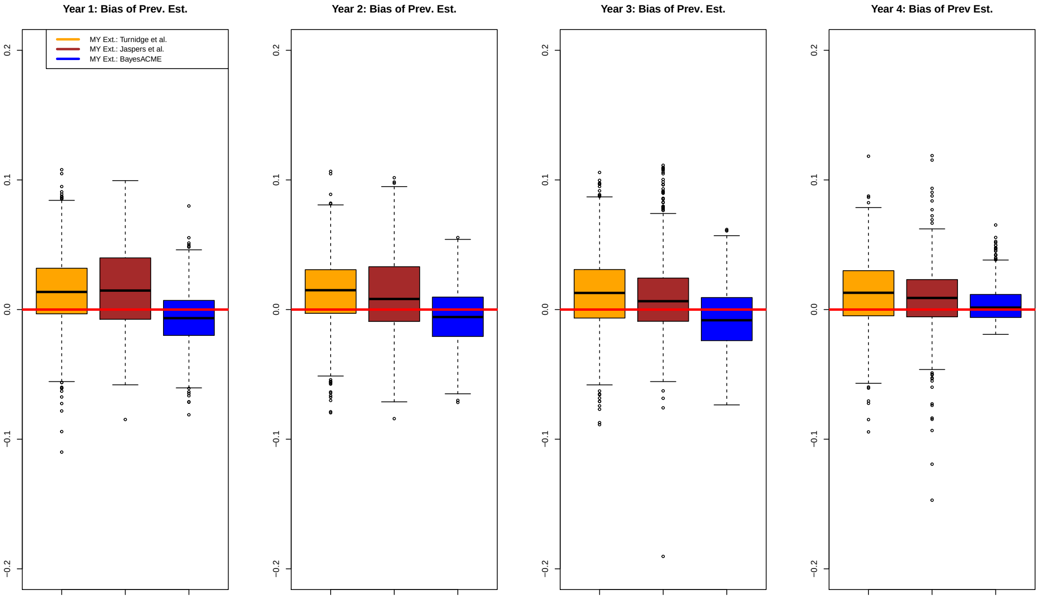


Figure G.20. Normal with High Contamination at Size 600

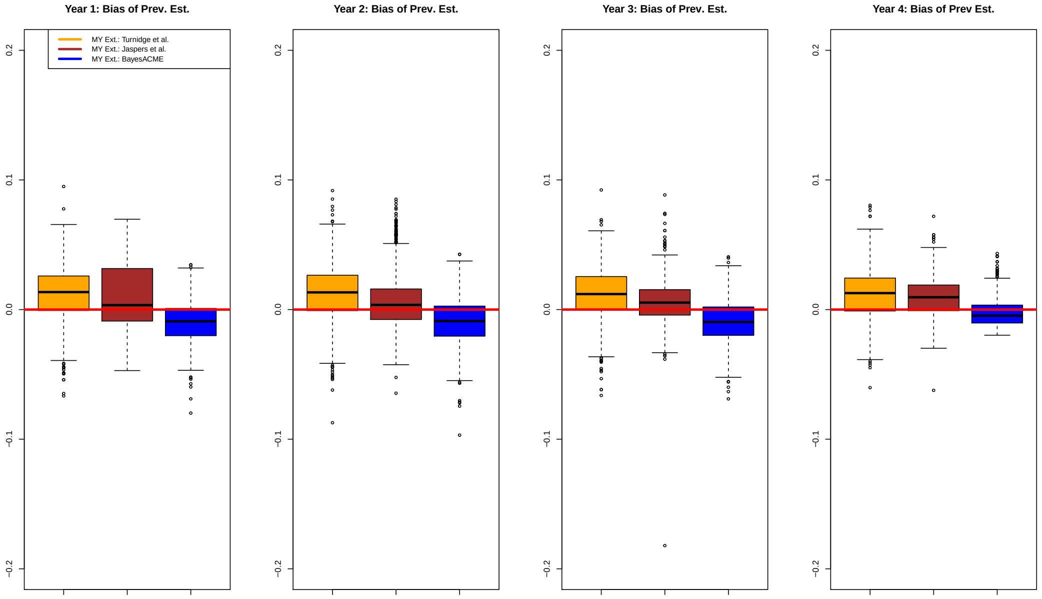


Figure G.21. Normal with High Contamination at Size 1200

Normal: Fast Trend

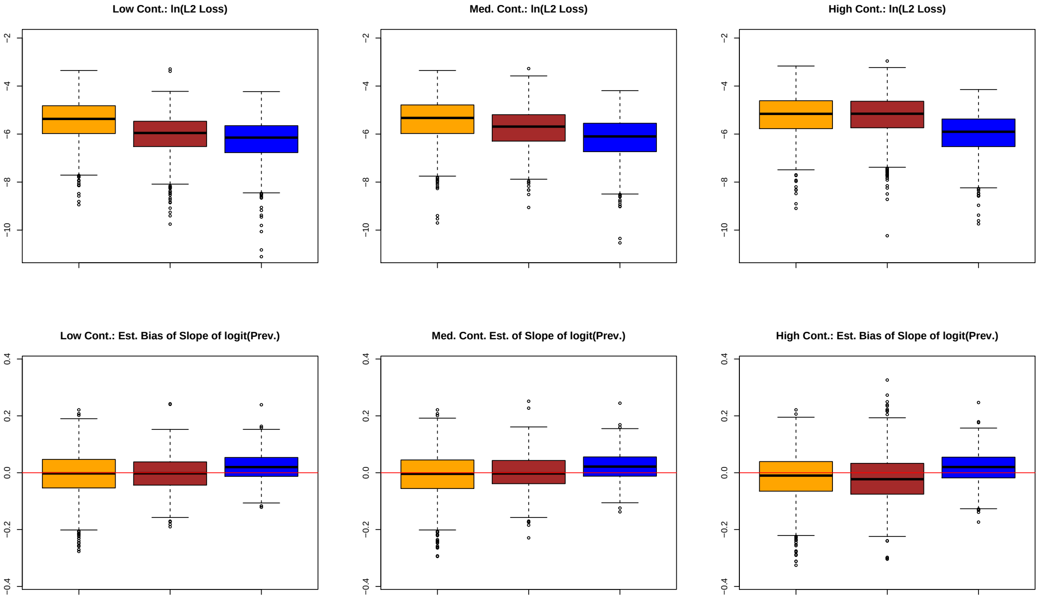


Figure G.22. Trend for Normal case with Fast Rate of Decline at size 300

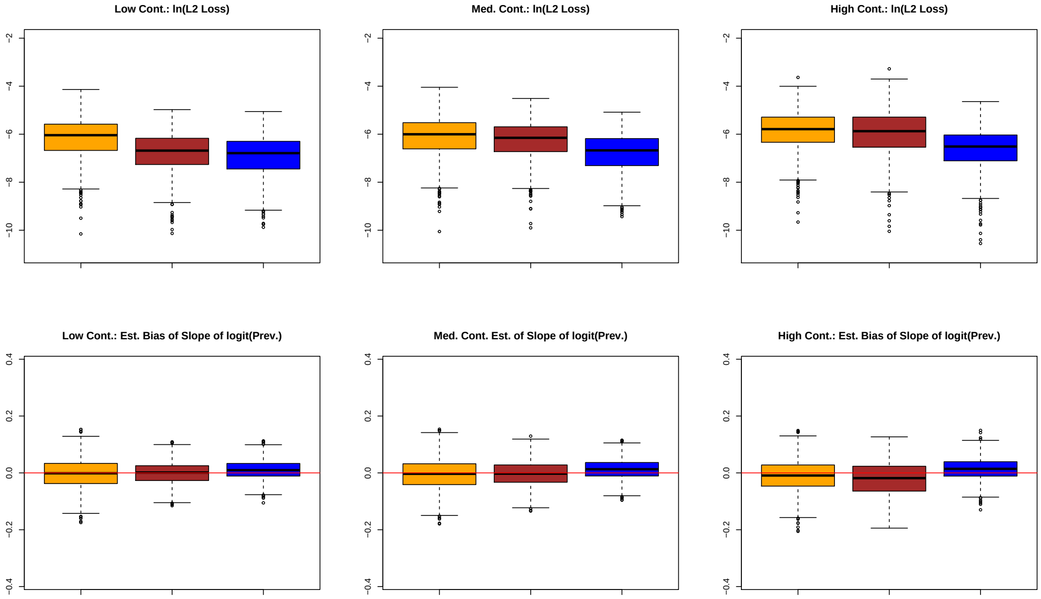


Figure G.23. Trend for Normal case with Fast Rate of Decline at size 600

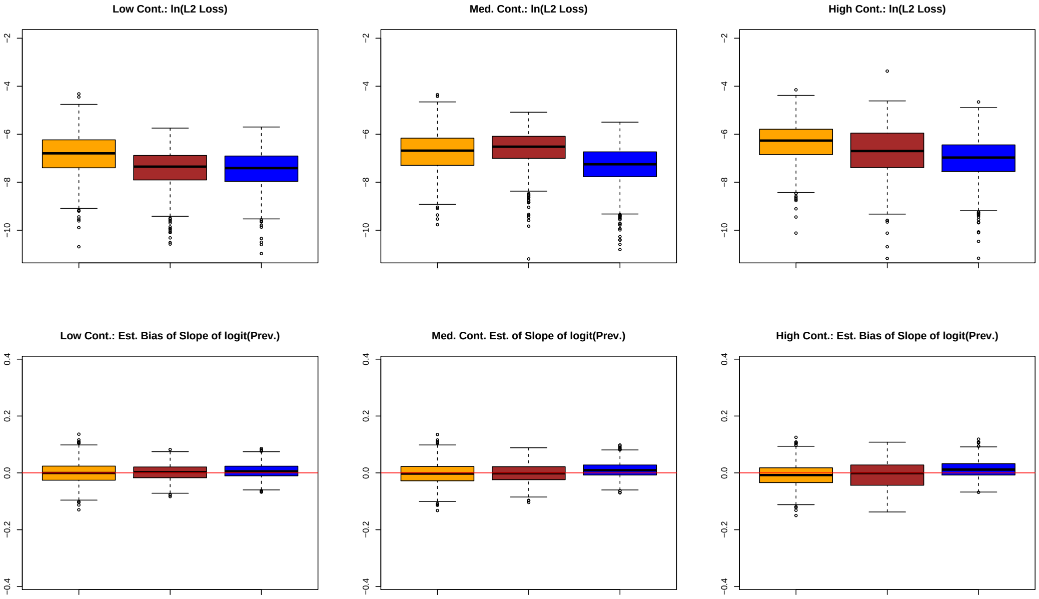


Figure G.24. Trend for Normal case with Fast Rate of Decline at size 1200

log2gamma Distribution

Slow Rate

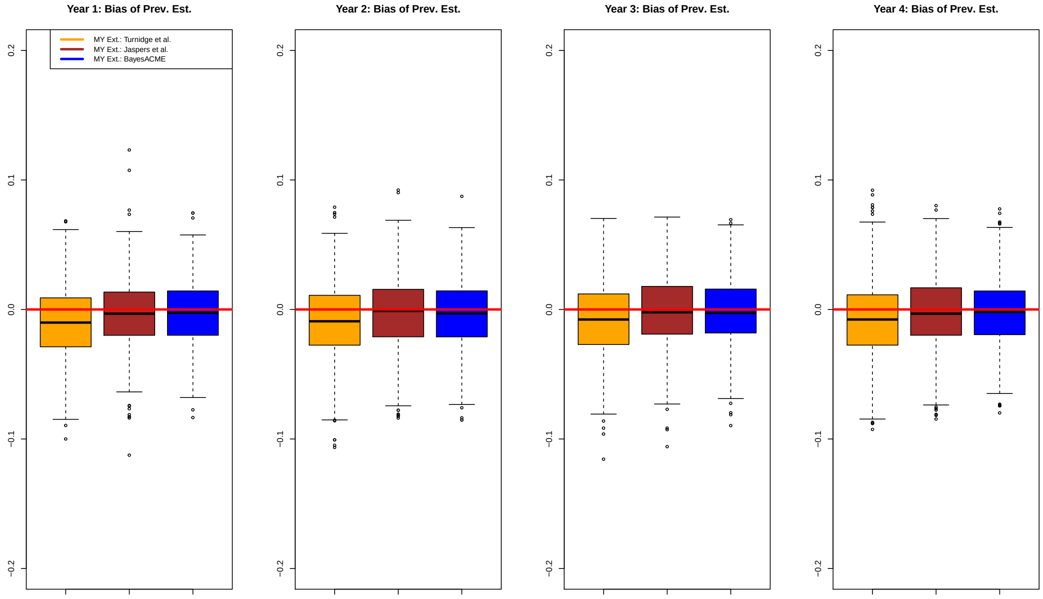


Figure G.25. log2gamma with Low Contamination at Size 300

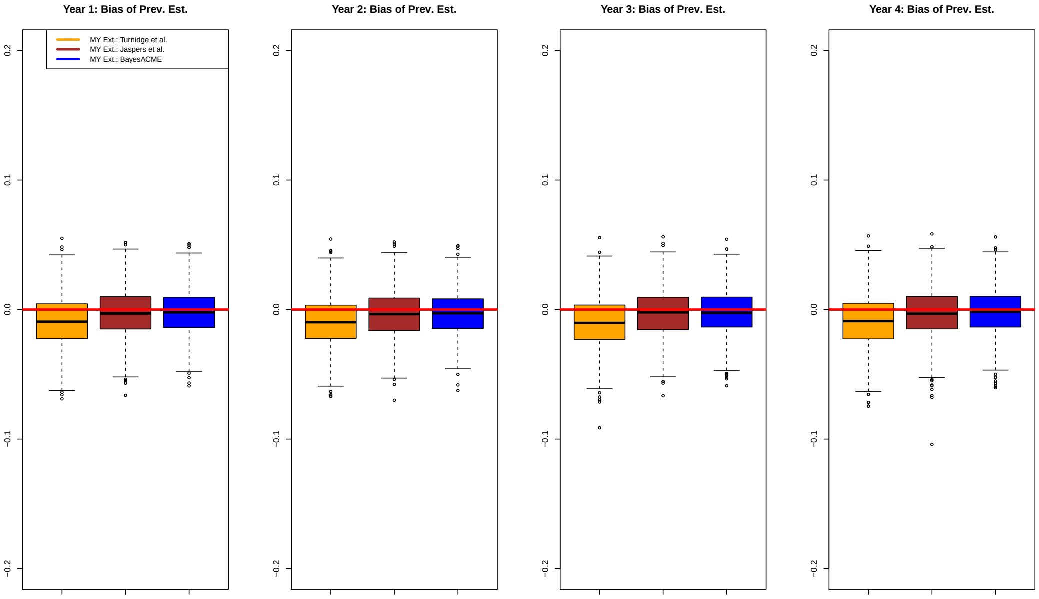


Figure G.26. log2gamma with Low Contamination at Size 600

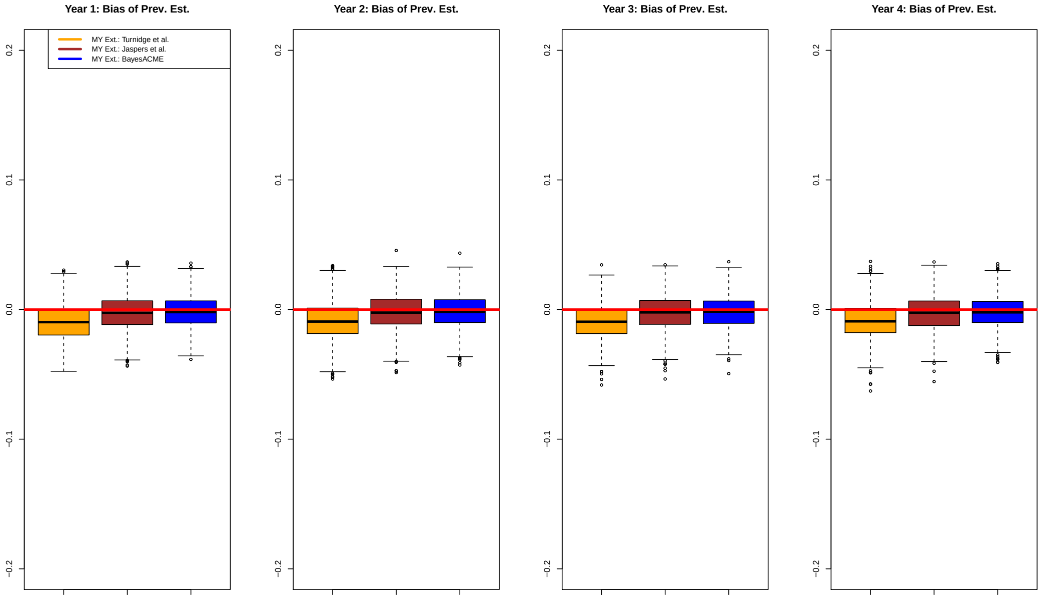


Figure G.27. log2gamma with Low Contamination at Size 1200

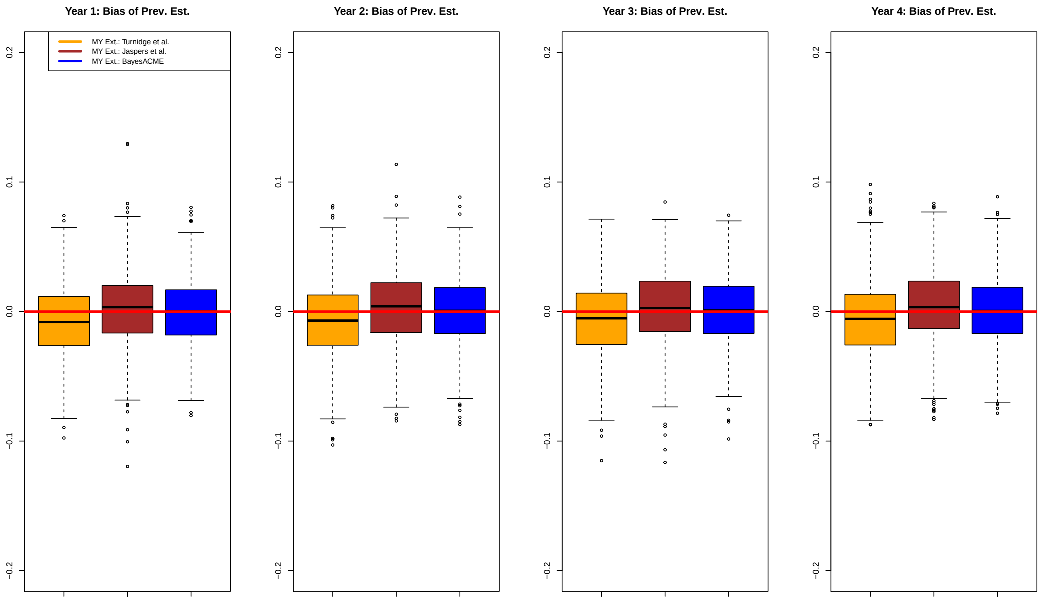


Figure G.28. log2gamma with Medium Contamination at Size 300

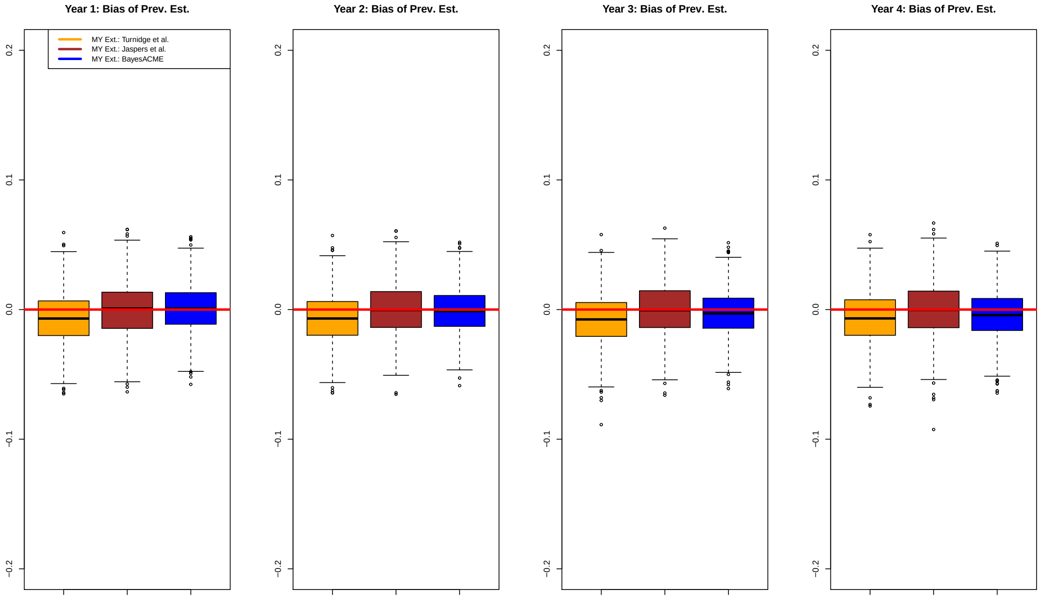


Figure G.29. log2gamma with Medium Contamination at Size 600

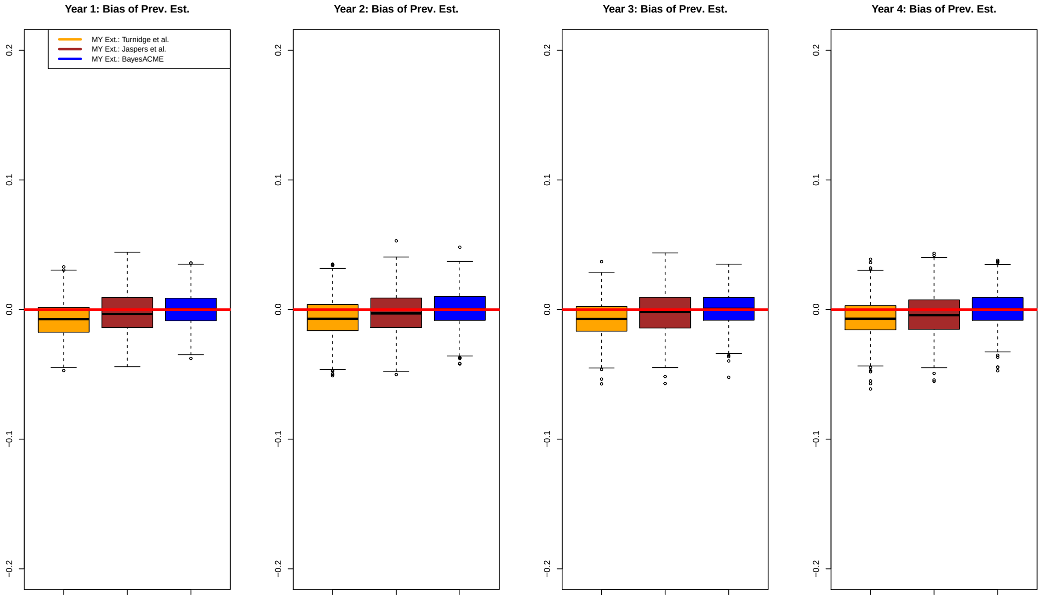


Figure G.30. log2gamma with Medium Contamination at Size 1200

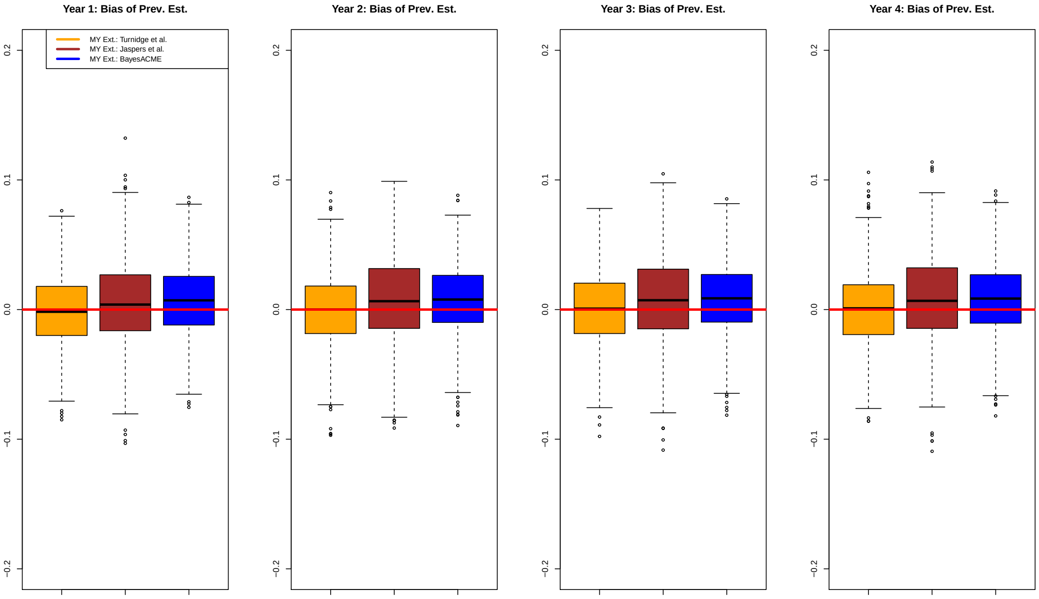


Figure G.31. log2gamma with High Contamination at Size 300

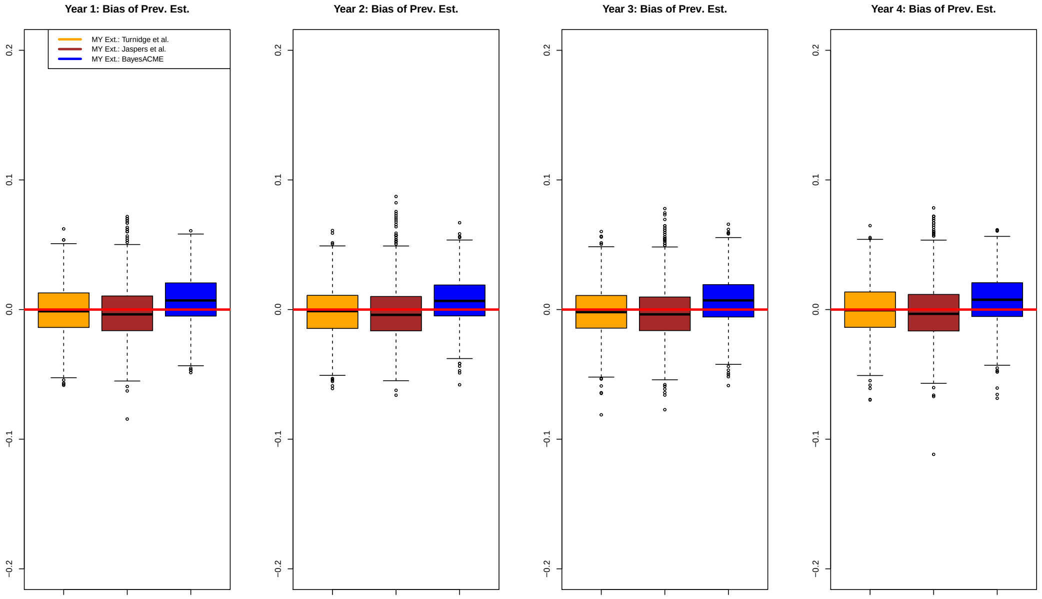


Figure G.32. log2gamma with High Contamination at Size 600

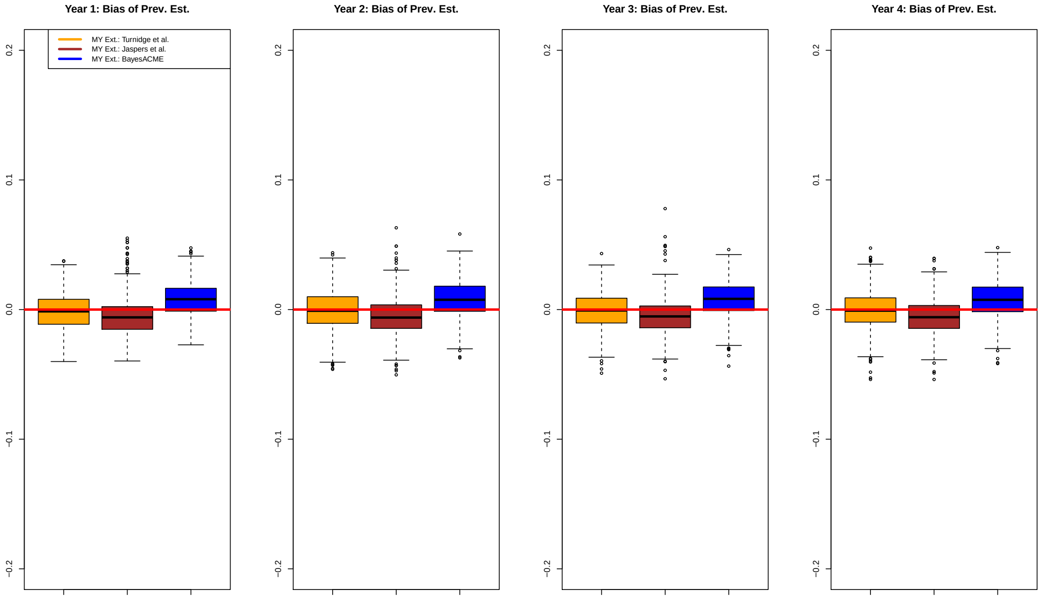


Figure G.33. log2gamma with High Contamination at Size 1200

log2gamma: Slow Trend

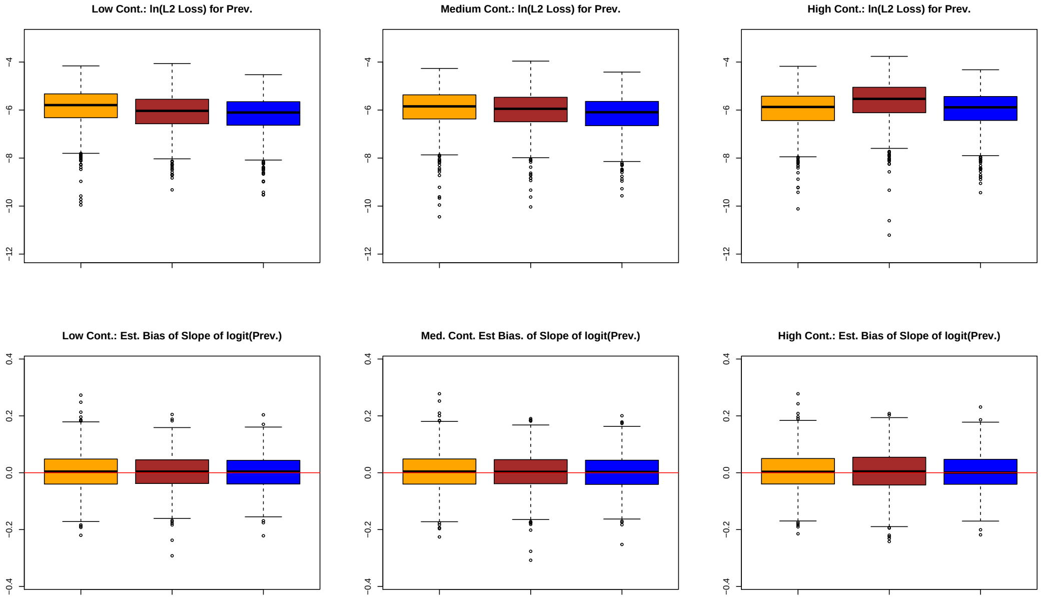


Figure G.34. Trend for log2gamma case with Slow Rate of Decline at size 300

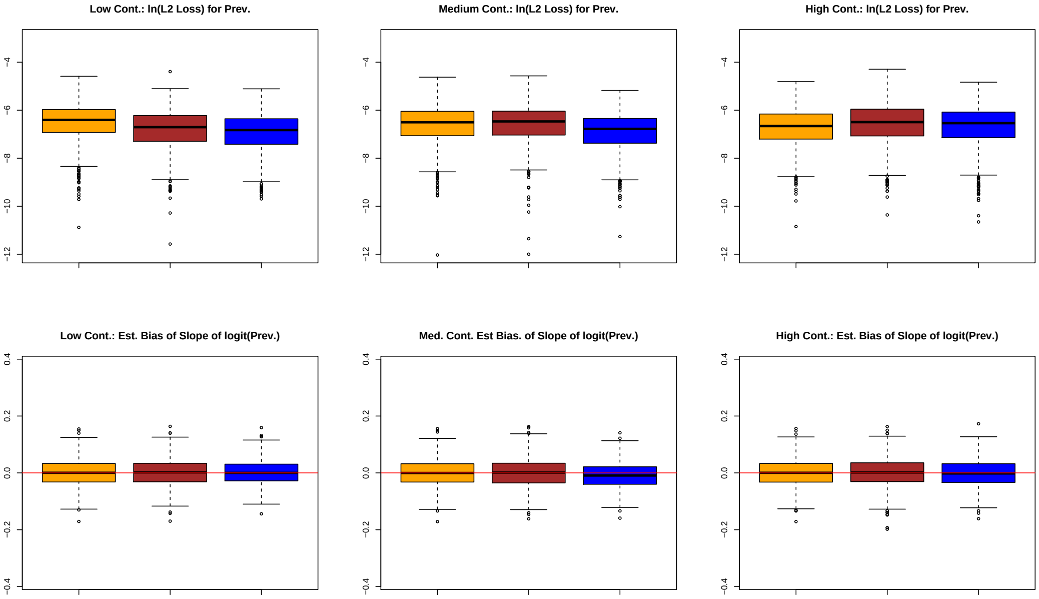


Figure G.35. Trend for log2gamma case with Slow Rate of Decline at size 600

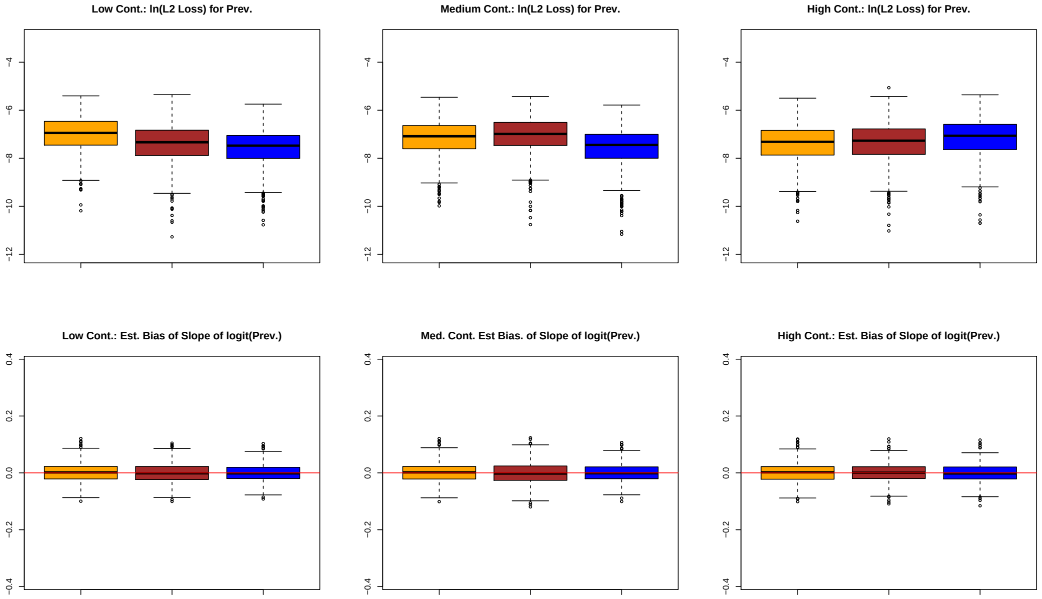


Figure G.36. Trend for log2gamma case with Slow Rate of Decline at size 1200

Fast Rate

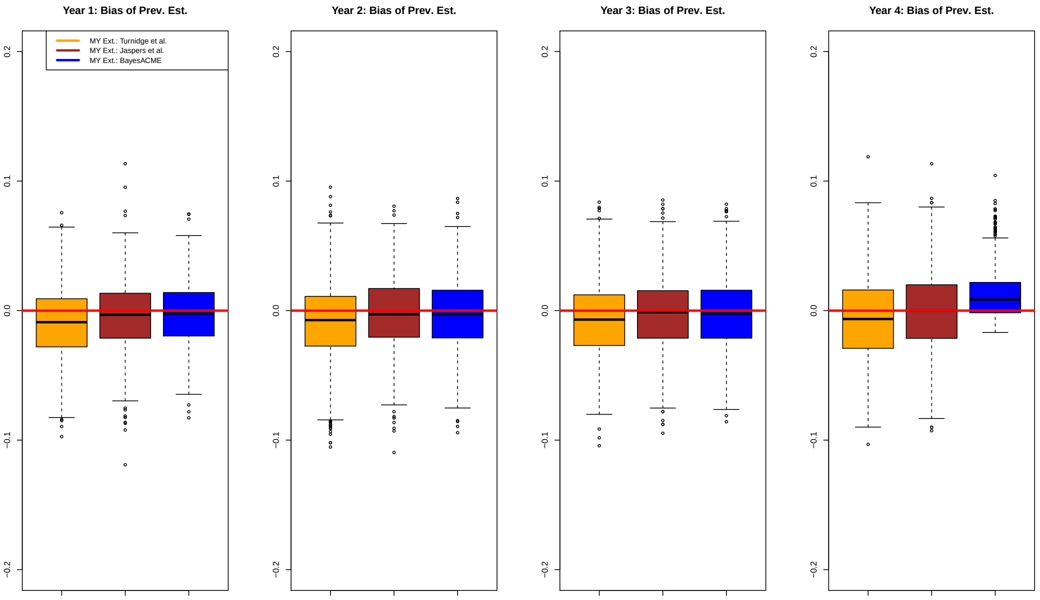


Figure G.37. log2gamma with Low Contamination at Size 300

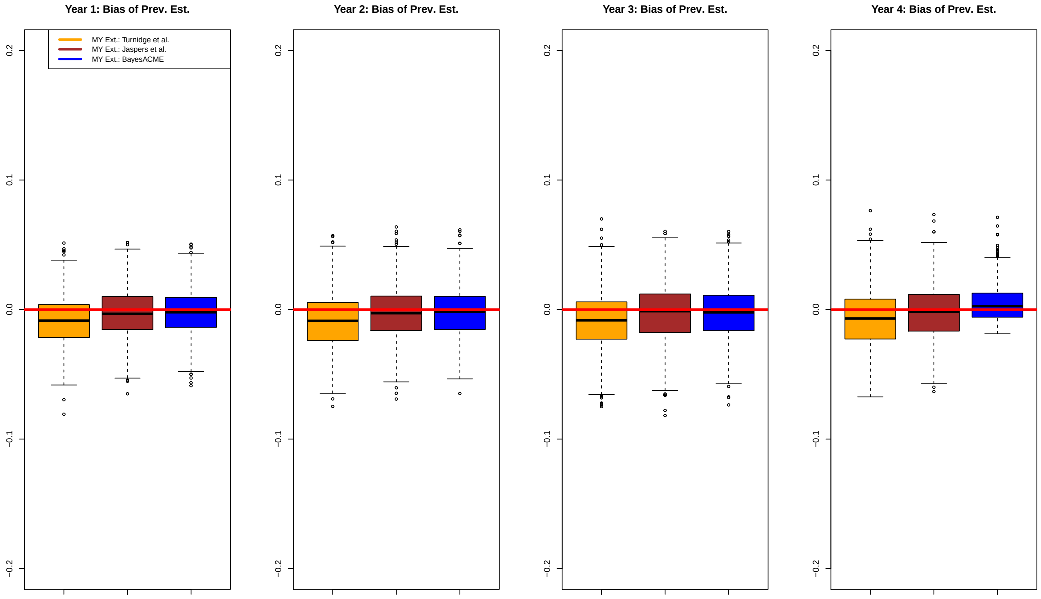


Figure G.38. log2gamma with Low Contamination at Size 600

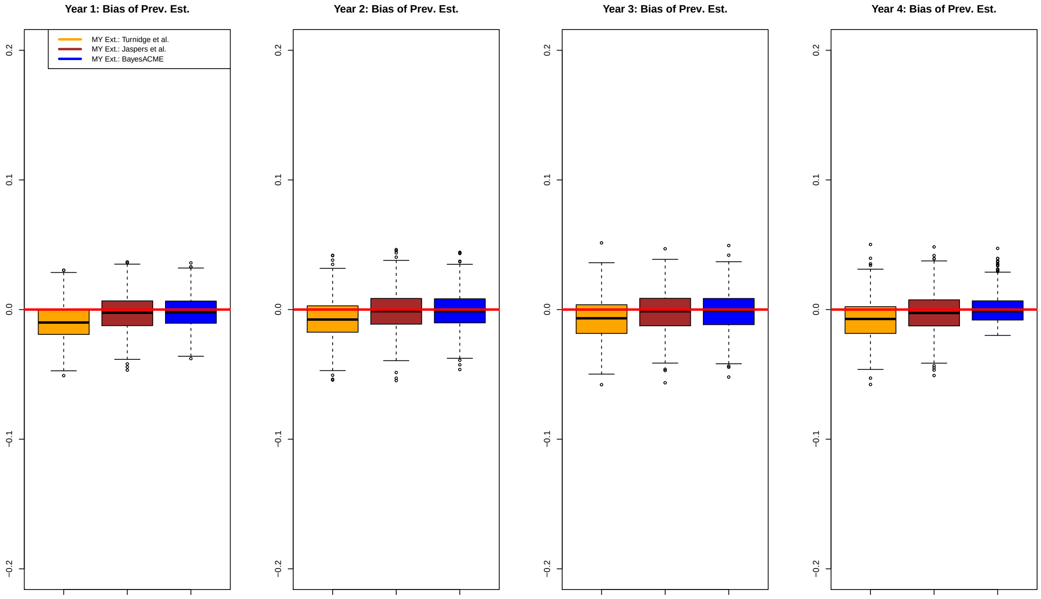


Figure G.39. log2gamma with Low Contamination at Size 1200

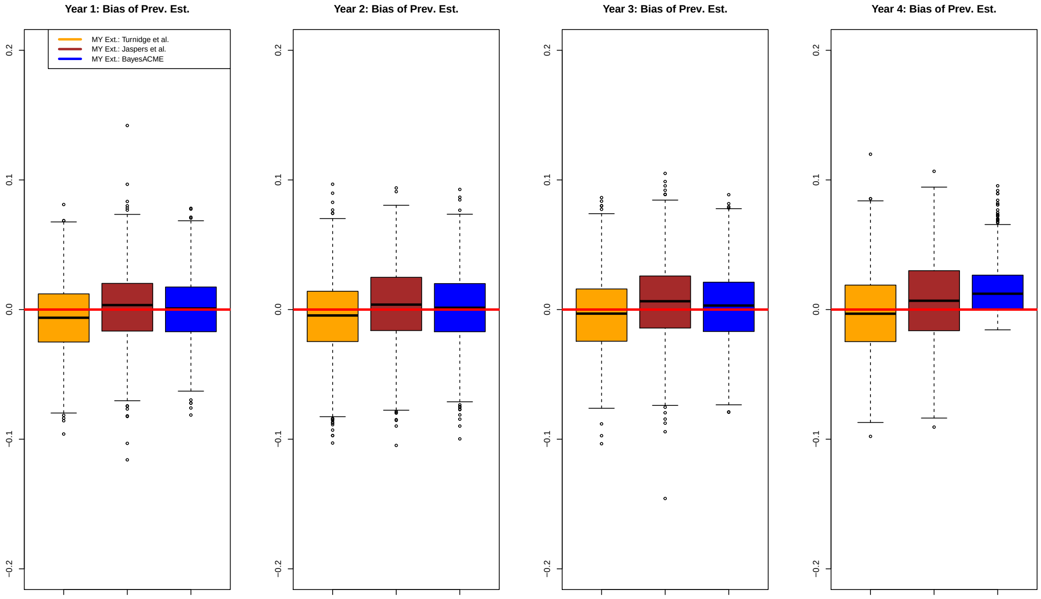


Figure G.40. log2gamma with Medium Contamination at Size 300

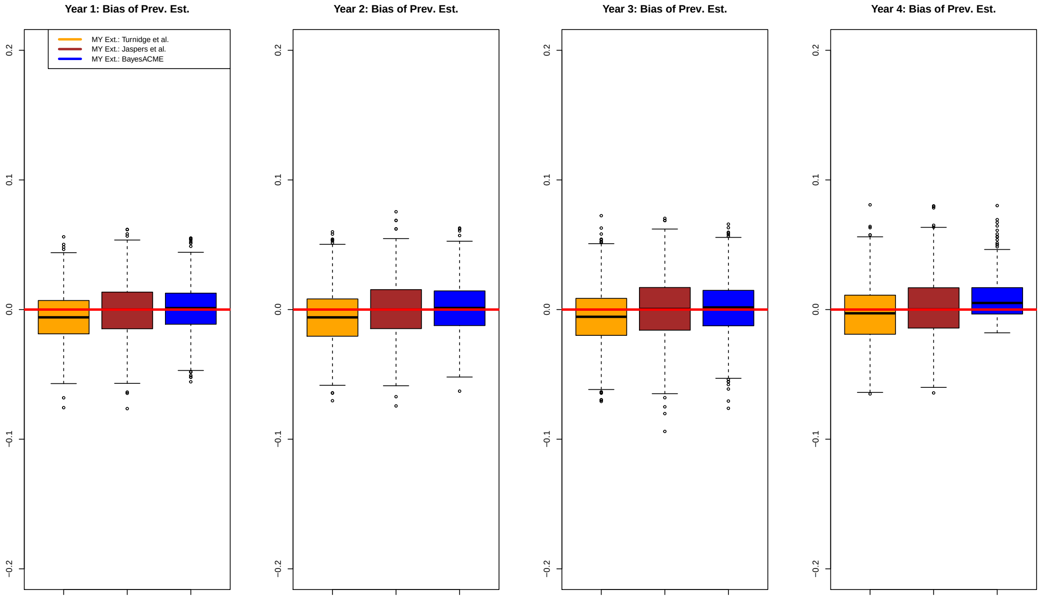


Figure G.41. log2gamma with Medium Contamination at Size 600

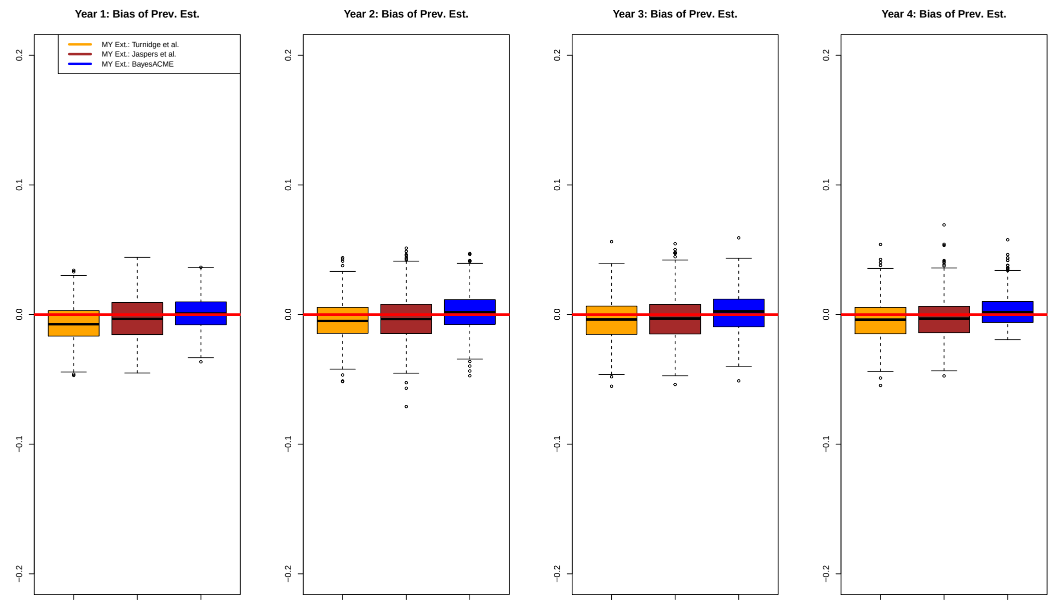


Figure G.42. log2gamma with Medium Contamination at Size 1200

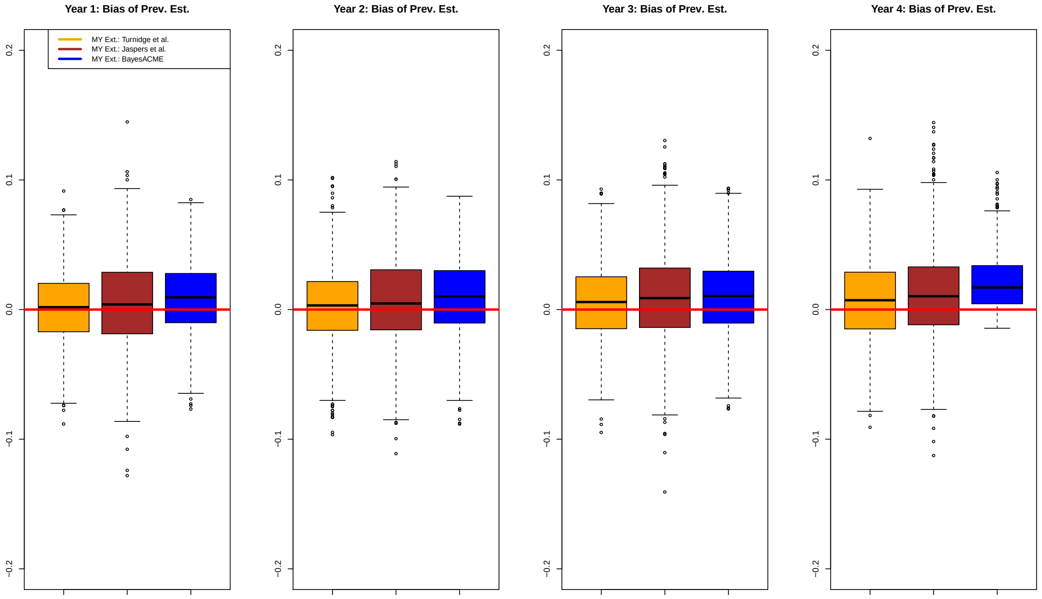


Figure G.43. log2gamma with High Contamination at Size 300

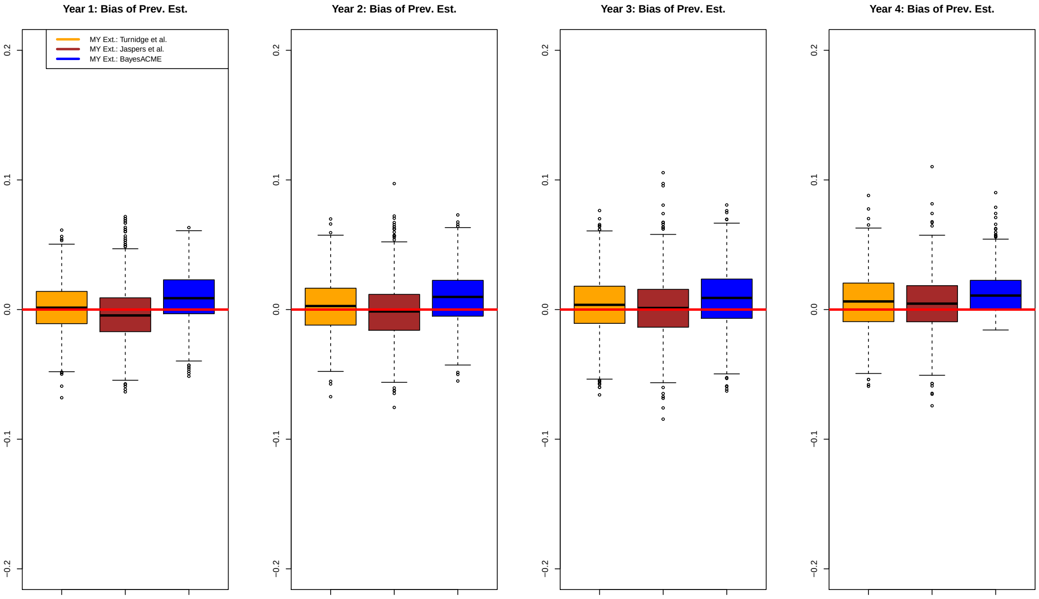


Figure G.44. log2gamma with High Contamination at Size 600

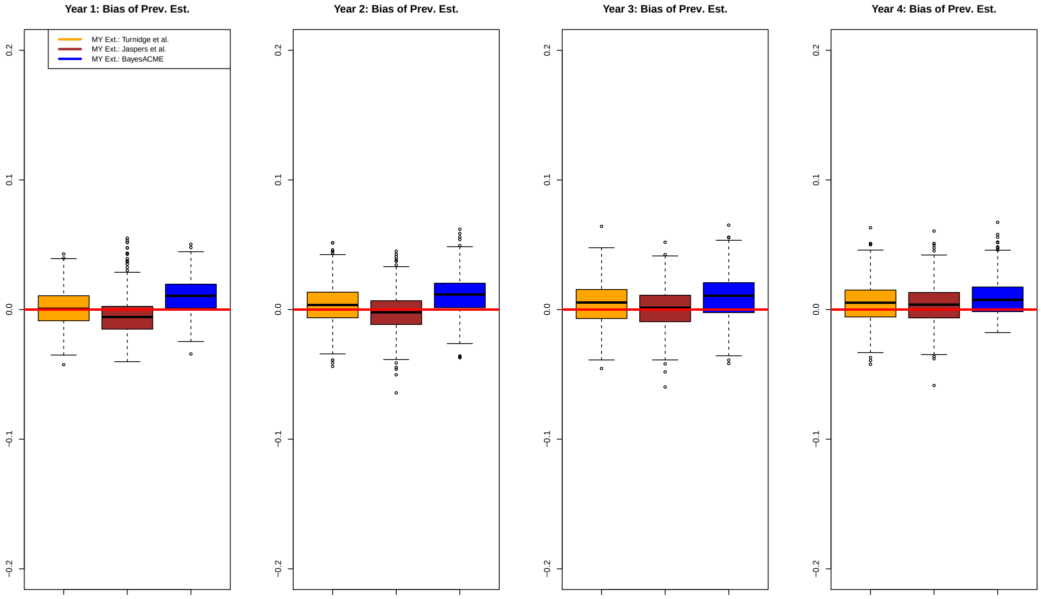


Figure G.45. log2gamma with High Contamination at Size 1200

log2gamma: Fast Trend

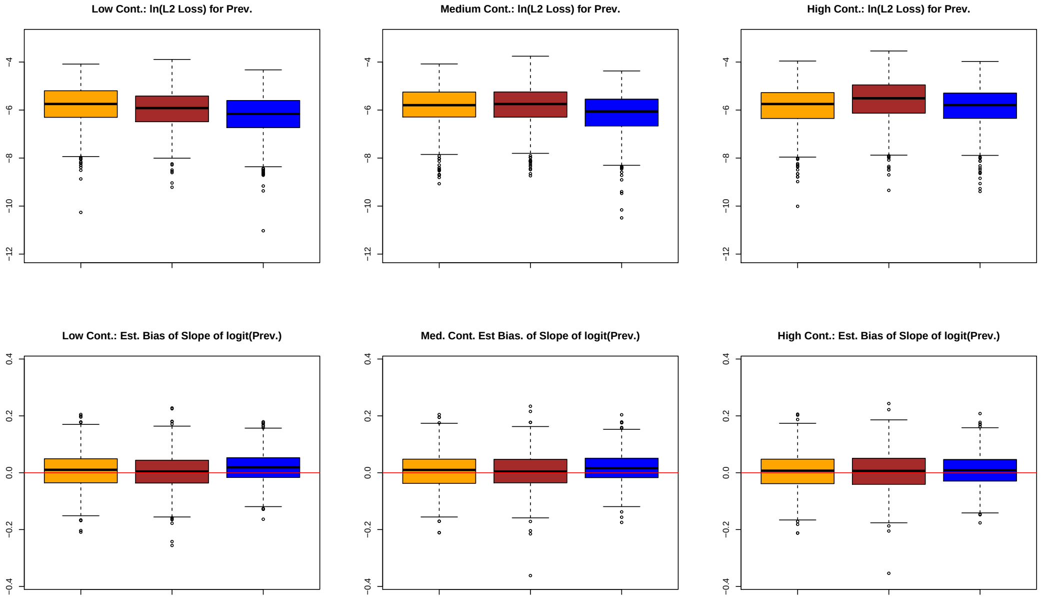


Figure G.46. Trend for log2gamma case with Fast Rate of Decline at size 300

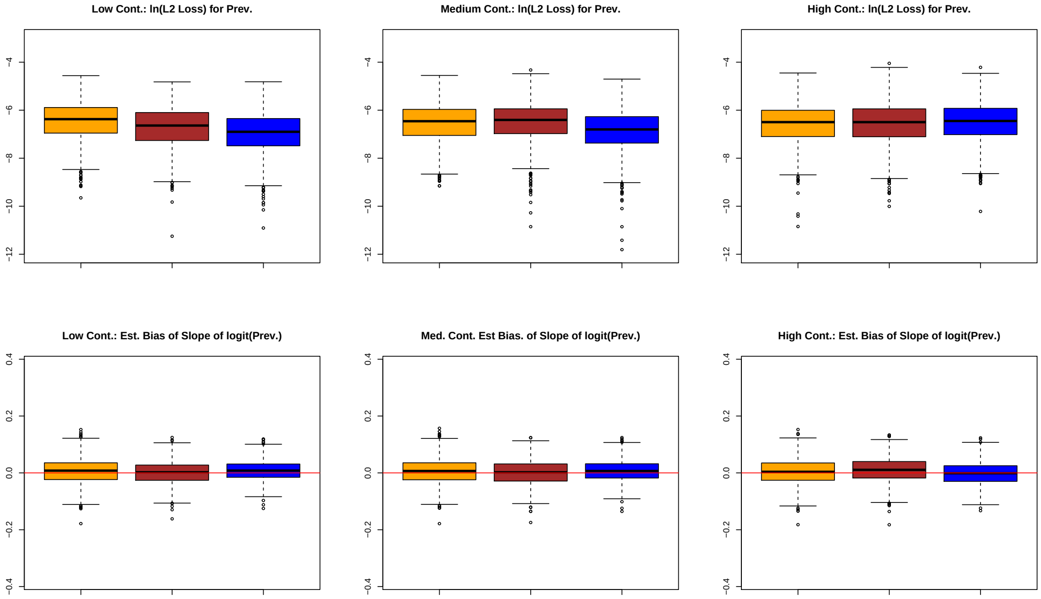


Figure G.47. Trend for log2gamma case with Fast Rate of Decline at size 600

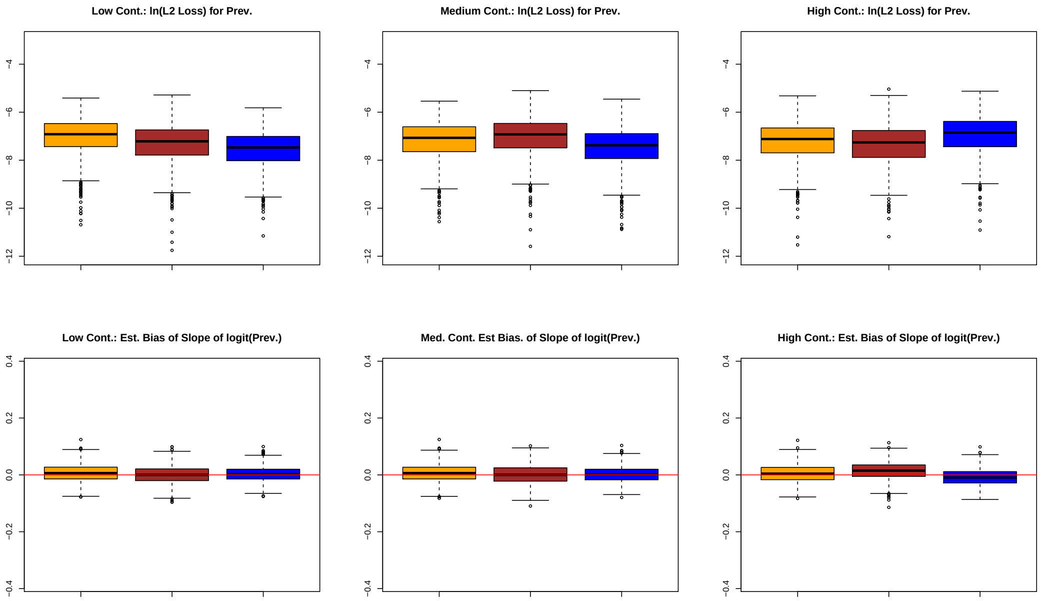


Figure G.48. Trend for log2gamma case with Fast Rate of Decline at size 1200

VITA

Will Eagan was born in Boston, Massachusetts. In 2011, he graduated with his B.A. in Mathematics from Hamilton College in Clinton, New York, U.S.A. He restarted his graduate studies at Purdue University in August 2013 after illness. He obtained his M.S. in Mathematical Statistics from the Department of Statistics at Purdue University in December 2015. In 2018, he received an American Statistical Association (ASA) Biopharmaceutical Scholarship Award for his research and service. In 2019, he led the Purdue team to a first place tie in the ASA Leadership Competition. In 2020, he received an Institute of Mathematical Statistics (IMS) Hannan Graduate Student Travel Award. Currently, he serves on the ASA Committee on Membership Recruitment and Retention.