12-2022

Statistical Challenges and Methods for Missing and Imbalanced Data

Rose Adjei
Utah State University

STATISTICAL CHALLENGES AND METHODS FOR MISSING AND IMBALANCED DATA

by

Rose Adjei

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences

(Statistics specialization)

Approved:

John R. Stevens, Ph.D. Major Professor	Christopher Corcoran, Ph.D. Committee Member
Daniel Coster, Ph.D. Committee Member	Ricardo Ramirez, Ph.D. Committee Member
Richard Cutler, Ph.D. Committee Member	Janis Boettinger, Ph.D. Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

Statistical Challenges and Methods for Missing and Imbalanced Data

by

Rose Adjei, Doctor of Philosophy

Utah State University, 2022

Major Professor: John R. Stevens, Ph.D.
Department: Mathematics and Statistics

The concept of missing data is present in almost every field of research. It constitutes considerable challenges in statistical analyses and interpretation of results and can weaken the ability to make valid statistical conclusions. The challenges associated with missing data can stem from the extent of data that are missing, the type of missingness present (whether Missing at Random (MAR), Missing Completely at Random (MCAR) or Missing Not at Random (MNAR)), as well as knowing the appropriate method to choose to effectively deal with the missingness in the analysis. The overall aim of this dissertation is to provide insight into missing data across different fields of study and address some of the above mentioned challenges of missing data through simulation studies and application to real datasets. This dissertation is in multi-paper format. The first paper of this dissertation addresses the dropout phenomenon in single-cell RNA (scRNA) sequencing through a comparative analyses of some existing scRNA sequencing techniques. Dropouts are technically considered missing data but are represented as zeros in scRNA sequencing. This can be very problematic when conducting any scRNA analyses as they are not easily identifiable from the true biological zeros and can introduce bias. The second paper of this work focuses on using simulation studies to assess whether it will be appropriate to address the issue of non-detects in data using a traditional substitution approach, imputation, or a non-imputation based approach. In an attempt to address this, these methods were compared at varying magnitudes of non-detects based on their Type 1 error and power effects. The final paper of this dissertation presents an efficient strategy to address the issue of imbalance in data at any degree (whether moderate or highly imbalanced). The primary technique employed to create this efficient strategy is combining random under-sampling with different weighting strategies for imbalanced data.

(95 pages)

Public Abstract

Statistical Challenges and Methods for Missing and Imbalanced Data

Rose Adjei

Missing data remains a prevalent issue in every area of research. The impact of missing data, if not carefully handled, can be detrimental to any statistical analysis. Some statistical challenges associated with missing data include, loss of information, reduced statistical power and non-generalizability of findings in a study. It is therefore crucial that researchers pay close and particular attention when dealing with missing data. This multi-paper dissertation provides insight into missing data across different fields of study and addresses some of the above mentioned challenges of missing data through simulation studies and application to real datasets. The first paper of this dissertation addresses the dropout phenomenon in single-cell RNA (scRNA) sequencing through a comparative analyses of some existing scRNA sequencing techniques. The second paper of this work focuses on using simulation studies to assess whether it is appropriate to address the issue of non-detects in data using a traditional substitution approach, imputation, or a non-imputation based approach. The final paper of this dissertation presents an efficient strategy to address the issue of imbalance in data at any degree (whether moderate or highly imbalanced) by combining random undersampling with different weighting strategies. We conclude generally, based on findings from this dissertation that, missingness is not always lack of information but interestingness that needs to investigated.

This dissertation is dedicated, first and foremost, to God Almighty, my source of inspiration, wisdom, knowledge and understanding. I dedicate this work to my dear husband, Emmanuel Nketia Boateng, who has been a great support and encouragement during the challenges of graduate school and life. I am truly grateful for your patience during this journey.

I would like to also dedicate this dissertation to my lovely mother, Madam Dora Boateng. Mama, we made it!. Thank you for all the prayers and words of encouragement. Finally, I dedicate this work to the memories of my late father, Mr. Daniel Adjei, and my late big sister, Winifred Ama Serwaa Bonsu. I miss you all very much and I will continue to make you proud. Keep resting in the bosom of the Lord Almighty.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. John R. Stevens. I am truly grateful for the extreme patience, valuable comments, advice, support and superb guidance throughout the difficult journey of creating this dissertation and successfully completing my PhD journey. Thank you for all the opportunities you provided me to grow and excel both professionally and academically. You were not only an amazing advisor but a wonderful father figure. Your words of encouragement blended with humor motivated me to keep pushing to the end even in times when I felt overwhelmed. I promise not to forget my oxford commas as I continue writing and to "Be Still (Psalm 46:10)" as I advance in my professional career and life. It was an honor to work and learn from you. I would also like to thank my committee members, Dr. Richard Cutler, Dr. Daniel Coster, Dr. Ricardo Ramirez, and Dr. Chris Corcoran. Thank you for your invaluable contributions and feedback that helped me finish my dissertation in a timely fashion. I am truly grateful to the Department of Mathematics and Statistics at Utah State University for giving me this opportunity and providing me with the needed support to complete my doctorate degree successfully. I am especially grateful to the department heads, secretaries, graduate program coordinators, professors, IT personnel, and fellow graduate students particularly Dr. Yan Sun, Dr. Brennan Bean, Dr. Eric Mckinney, Gary Tanner, Ting Xue, Kelly Seipert, Sara Poulsen, Gaby Hainsworth, Karl Dyches, and Isabel Jenson for their kindness and indispensable professional support. I would also like to acknowledge all professors and faculty members from my alma maters, Northern Arizona University and Kwame Nkrumah University of Science and Technology, for providing me with a solid statistical foundation to be able to pursue a PhD degree. Special thanks to Dr. Brent Burch, Dr. Jin Wang, and Dr. St. Laurent.

My deepest gratitude also goes to Prof. Gabriel Asare Okyere, and Dr. Yao Elikem Ayekple for the immense support they provided me to be where I am today.

This work would also not have been possible without the support and resources from the Center for High Performance Computing at the University of Utah. A special thank you to Dr. Martin Cuma, whose high performance computing expertise made many of the simulation results documented in this dissertation possible.

I would also like to appreciate my professional and academic mentors for their insightful contributions and professional help. Mores specifically, Michael R. Quigley (Lead Quantitative Analyst, Wells Fargo Bank), Sheri Hsueh (Senior Vice President, Corporate Risk, Wells Fargo Bank), Prof. Loni

Tabb (Dornsife School of Public Health, Drexel University), and Dr. Yvonne Kumi (Quantitative Analytics Consultant, Wells Fargo Bank).

I am grateful to all my friends and family for the daily check-ins, prayers, tremendous support and encouragement throughout this journey. I could not have made it without you.

Finally, I would also like to thank my dear husband, Emmanuel, for being patient and understanding. Thank you for being there and pushing me out of my comfort zone. We did it hun! I love you.

Rose Adjei

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

OVERVIEW OF MISSING DATA

1.1 Introduction

Missing data is a common statistical phenomenon that is encountered in almost every field of research and can affect making valid statistical conclusions. Missing data (also missing values) refer to those data values that are not stored for the variable in an observation of interest.
A lot of factors can contribute to missingness in data which include:

- Dropouts
- Survey non-responses
- Technical problems (low or no capture efficiency)
- Data entry mistakes
- Improper data collection by the researcher
- Natural phenomenon (e.g. disaster, rain, death, etc.)

Missing data if not properly handled can pose a number of problems. First of all, having missing values in your data can lead to unbalanced data. Second, it can result in biased estimates being produced. Third, it can lead to a reduction in statistical power, the probability of rejecting the null hypothesis when, in fact, it is false. These issues in effect will produce misleading results and affect the efficiency of the study (Kang, 2013).

1.2 Types of Missing Data

Understanding the nature of missingness is key in the analysis of missing data. When the researcher gets to know why the data values are missing and why they matter, it will be helpful in deciding on the approach to take when handling the missing data. The problems associated with missing data and the solution to these problems may differ depending on the type of missing data being dealt with. We will consider three major mechanisms of missing data based on the relationship between the observed variables and the (unobserved) missing data.

- **Missing Completely At Random (MCAR)**: The probability of missingness is unrelated to the process under study and occurs entirely at random. This means that missing data is completely independent of observable variables and unobservable parameters of interest. MCAR data is regarded as a random sample of the target/full population which is an unrealistically strong assumption. In reality, MCAR data are rare and analysis performed on this kind of missing data produces unbiased results (Gebregziabher, 2019).

- **Missing At Random (MAR)**: The probability of missingness depends only on the observed variables but not on the unobserved (missing) values. With this mechanism, the missing data has a relationship with other variables in the dataset. Due to this dependency, the MAR data is not a random sample of the full population. MAR data can induce bias in the estimation of parameters depending on the method that is used for analysis. MAR is more common and realistic as compared to MCAR. This missing data mechanism is said to be ignorable, which basically means that there is no need to model the missing data mechanism as part of the estimation process. Ignorability is not an inherent characteristic of MAR instead it depends on the analytic model being used (using correct model on the observed, missing data mechanism is not needed) (Gebregziabher, 2019).

- **Missing Not At Random (MNAR)**: This mechanism is also called nonignorable nonresponse/missing. This is neither MCAR nor MAR. The pattern of missingness depends on both observed and unobserved data. The pattern of missingness is related to other variables in the dataset, but in addition, the values of the missing data are not random. Probability of missingness varies for reasons that are unknown to us. (Unobserved measurements influence the process governing missingness, in addition to influences coming from observed measurements and/or covariates). MNAR assumptions cannot fully (empirically) be verified from data.

> Valid inferences require joint modelling of the response and the missing data mechanism hence rendering most standard methods of analyses invalid (Gebregziabher, 2019).

Knowing these distinctions are important because the validity of an analysis will depend on the missing data mechanism.

1.3 Missing Data Techniques

Missing data may or may not be problematic depending on whether your conclusions gets affected or not. Missing data being problematic (i.e leading to invalid results and conclusions) may not be outrightly seen and can only be justified after careful analysis of your missing dataset using appropriate techniques (D. Hess, 2020). When handling missing data, there are three main approaches that are taken:

- Omission - samples with invalid data or missing values are excluded from further analysis.
- Imputation - assigning values to the missing data.
- Full analysis - directly applying methods that are unaffected by missing values.

1.3.1 Missing data methods that throw away data (Omission)

To basically resolve the problem of missing data, most people tend to rely on methods that throw away or omit data in order to simply avoid having to deal with missingness in data. Throwing away data leads to reduced sample size which can further produce estimates with large standard errors and result in distorted inferences about the sample population (Gelman and Hill, 2006). The sample observations that remain may not be a true representation of the population.

Complete case analysis/listwise deletion is where an entire record/case/unit is excluded from the dataset if any single value is missing before performing further analysis. This is by far the most common method when handling missing data (Kang, 2013). This method only uses cases that have complete data. Therefore, complete case analysis has the potential to introduce bias into the estimates and lead to invalid inferences. However, depending on the type of missingness present in the data the problem of biasedness may be avoided. Listwise deletion may be a reasonable (optimal) strategy for the researcher to consider, provided that the sample data is large enough (where power is not a problem) and the assumption of MCAR is satisfied. Even if the MCAR assumption is met, complete case analysis can be immensely inefficient if the sample data is not large enough (power issues arising) (Gebregziabher, 2019). Researchers using listwise deletion may have to deal with the problem of losing large amounts of data especially where there are a high number of missing cases.

Pairwise deletion / available-case analysis is another simple approach of handling missing data by throwing away data. Unlike listwise deletion, pairwise deletion enables researchers to use as much of data as possible. This method does not include a particular variable when it has a missing value, but it can still use the case when analyzing other variables with non-missing values (Support, 2020). A researcher simply excludes a variable or set of variables from the analysis because of their missing-data rates (sometimes called "complete-variables analyses"). This means that pairwise deletion allows you to use more of the data and perform different statistical analyses using different subsets of the data (Solution, 2020).

A much simpler way to think of how pairwise deletion works is to think of correlation matrix. A correlation measures the strength of the relationship between two variables. For each pair of variables for which data is available, the correlation coefficient will take that data into account. Thus, pairwise deletion maximizes all data available by an analysis by analysis basis (Solution, 2020). Even though this method has the ability to preserve more information, the populations of each analysis would be different and possibly non-comparable (Salgado et al., 2016). Hence researchers may have challenges with drawing inferences to the total sample (Solution, 2020). Pairwise deletion is known to work well with datasets that satisfy MCAR or MAR assumptions thereby producing less biased results (Kang, 2013). Different sample sizes are considered under pairwise deletion which may be problematic as the results obtained will not generalize to the entire original population.

Weighting is another data omission method that can be considered. As seen with complete case analysis, excluding cases with missing data can distort the representativeness of the sample and produce biased estimates. Weighting is therefore a method carried out to compensate for the missingness (nonresponse), restore the representativeness of sample, and reduce the bias in estimates produced (Raghunathan, 2004a) & (Gelman and Hill, 2006).

Under this procedure, a model is built to predict the non-response of the variables with missing data using the other variables with complete data. The weights assigned are the inverse of the probability of response and are mostly used in regression models, e.g., weighted logistic regression models. The inverse of predicted probabilities of response from this model could then be used as survey weights to make the complete-case sample representative (along the dimensions measured by the other predictors) of the full sample (Gelman and Hill, 2006). Results produced may be unbiased if the data satisfy the MAR assumption where the observed data are a random sample in the weighting class (Gebregziabher, 2019).

Weighting becomes more complicated when there is more than one variable with missing data (Gelman and Hill, 2006). Again, weighting may not be the best option and may lead to biased results if the respondents differ significantly from the non-respondents in the class(Gebregziabher, 2019). Lastly, there is the potential that standard errors will become erratic if predicted probabilities are close to 0 or 1.

1.3.2 Missing data approaches that retain all data / assign values to missing data (Imputation)

Marginal mean imputation is a fast and simple fix to handling missing data. This method replaces the missing value by the mean of the observed values for that variable (Van Buuren, 2018). This could be problematic as the distribution of the variable imputed can be distorted severely. Again, mean imputation can lead to discrepancies in summary statistics which include underestimation of variance since it keeps the sample size at its full value and distortion of the relationship between variables by pulling the estimates of the correlation towards zero (Gelman and Hill, 2006). Using mean imputation tends to introduce bias in the estimate of the mean if the data is not MCAR. Hence this method should probably be used as a quick fix when there is just a very small number of missing values (Van Buuren, 2018). The more the missing data the larger the underestimation of the variance which may artificially lead to very small p-values and increase the possibility of type I errors (Grace-Martin, 2020a).

Last Observation Carried Forward (LOCF) takes previous observed value as a replacement for missing data. When multiple values are missing in succession, this procedure searches for the last observed value to use as a replacement for the missing values (Van Buuren, 2018). The missing value is replaced by the last observation from the same subject. LOCF is a highly popular imputation method used in longitudinal studies / time series analysis / clinical trials in which the subjects are repeatedly measured over a series of time-points (Kang, 2013).

LOCF assumes that outcomes do not change over time when a subject becomes missing and that a single data point can be used to estimate the distribution of potential values. Hence LOCF may be appropriate to use in situations where measured variables are fixed and are known not to change over time (Shoop, 2015) or in cases where we are certain of what the missing values should be (Van Buuren, 2018). These underlying assumptions are occasionally justifiable and hence the use of LOCF can be problematic when these assumptions are not met (Shoop, 2015).

One advantage of LOCF is that it is convenient since it generates a complete dataset. Also, this method is easy to understand and communicate between statisticians and clinicians or between a sponsor and the researcher or non-statisticians (Kang, 2013). LOCF is sometimes traditionally viewed as preferred method in clinical trials as it is considered conservative (one that would lead to underestimation of the true treatment effect) and less prone to selection (Van Buuren, 2018). It must be noted that under certain restrictive assumptions LOCF can produce unbiased estimates of treatment effect. On the other hand, LOCF produces biased estimates of the treatment effect and underestimates the variability of the estimated result unless the above mentioned assumptions are met (Kang, 2013). Now, LOCF can result in biased estimates even under MCAR.

It is therefore recommended that LOCF needs to be followed by a proper statistical analysis method that can distinguish between the real and imputed data. However, this is typically not done. The Panel on Handling Missing Data in Clinical Trials recommends that LOCF should not be used as the primary approach for handling missing data unless the assumptions that underlie it are scientifically justified (National Research Council 2010, 77) (Van Buuren, 2018).

Zero Imputation is one of the simplest and most intuitive ways of compensating for missing values in a dataset. This method is carried out by replacing missing values with zeros. Even though Zero imputation may come off as a simple and efficient technique which retains the full dataset, it has the tendency to artificially create erroneous relationships between variables thereby leading to high estimation errors (Shi, 2007).

Regression Imputation incorporates the knowledge of other variables with the idea of producing smarter imputations. The first step in this procedure involves building a model from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data. The imputed values correspond to the most likely values under the model (Van Buuren, 2018). There are a number of advantages associated with using regression imputation. First of all, unlike the listwise or pairwise deletion, regression imputation retains a great deal of data and avoids significantly altering the standard deviation or the shape of the distribution (Kang, 2013). Regression imputation yields unbiased estimates of the means under MCAR, just like mean imputation, and of the regression weights of the imputation model if the explanatory variables are complete. Moreover, the regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model (Van Buuren, 2018).

On the other hand, imputing predicted values has an effect on the correlation. One may be led to believe that we're to do a good job by preserving the relations between the variables. In reality however, regression imputation artificially strengthens the relations in the data. Correlations are biased upwards, and the variability of the imputed data is systematically underestimated, making the imputations too good to be true. Regression imputation is a recipe for false positive and spurious relations. In view of these pitfalls, it is recommended that regression imputation should be used when more than 10% of the data is missing and when the data contains highly correlated variables (Lodder, 2013).

KNN is a non-parametric algorithm that is used to replace missing values in a dataset. This method is sometimes referred to as the "Nearest Neighbor Imputation". KNN can be used for data that are continuous, discrete, ordinal, and categorical which makes it particularly useful for dealing with all kind of missing data (MAR, MCAR, MNAR). The KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. When using KNN for missing values, it is assumed that a point value can be approximated by the values of the points that are closest to it, based on other observed variables (Obadiah, 2017). Usually KNN algorithm uses the Euclidean distance as a distance metric to identify neighbouring points (Malarvizhi and Thanamani, 2012). The KNN algorithm identifies different groups ('K' samples) in the dataset that are similar to each other and use these 'K' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'K'-neighbors found in the dataset (Vidhya, 2020).

When using KNN algorithm to impute, it is important to carefully select K. The K in the KNN algorithm is a user defined constant which represents the number of neighbors to be specified. Choosing a very low K can increase the influence of noise and make the results less generalizable. On the other hand, a very high K will tend to blur local effects which are exactly what we are looking for. It is recommended that an odd K is selected for binary classes to avoid ties. The KNN algorithm automatically normalizes the data when both numeric and categorical variable are provided. This allows every attribute/variable the same influence in identifying neighbors when computing certain type of distances like the Euclidean one.

KNN is very easy to implement and simple to understand as only two parameters are required to implement (the value of K and the distance function). Again, KNN algorithm requires no training before making predictions, hence new data can be added seamlessly which will not impact the accuracy of the algorithm. However, this imputation method may not be always appropriate as it does not work well with large datasets. The cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm. Also, it does not work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension. Lastly, KNN is very sensitive to noisy data and outliers (Kumar, 2019).

The above imputation techniques (mean imputation, LOCF, zero imputation, regression imputation, and KNN) on the whole are classified as single imputation methods. This is because these methods generate a single estimate for the missing value. All single imputation methods underestimate standard errors (Grace-Martin, 2020b).

Multiple imputation as the name suggests is a simulation-based technique for handling missing data that creates multiple versions of imputed data. This method is a more advanced technique compared to the other imputation methods mentioned above. Multiple imputation creates several complete versions of the data by replacing the missing values by plausible data values. These plausible values are drawn from a distribution specifically modeled for each missing entry. The imputed datasets created (say size m) are identical for the observed data entries, but differ in the imputed values. The next step is to estimate the parameters of interest from each imputed dataset. This is typically done by applying the analytic method that we would have used had the data been complete. The results will differ because their input data differ. It is important to realize that these differences are caused only because of the uncertainty about what value to impute. The last step is to pool the m parameter estimates into one estimate, and to estimate its variance (Van Buuren, 2018). Standard errors are calculated using Rubin's (1987) formula that combines variability within and between data sets (Allison, 2012a) Under the appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties and hence are used as the final imputed values to complete the dataset (Van Buuren, 2018).

Over the last decade, multiple imputation has rapidly become one of the most widely-used methods for handling missing data. However, one of the big uncertainties about the practice of multiple imputation is how many imputed data sets (m) are needed to get good results. A simplified rule of thumb that is mostly applied is that the number of imputations should be similar to the percentage of cases that are incomplete. So if 27% of the cases in your data set have missing data on one or more variables in your model, you should generate about 30 imputed data sets. With large data sets and many variables in the imputation model, this can become burdensome as getting more data sets will require more computing time (Allison, 2012a). Multiple imputation can be used with any kind of data and model with conventional software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates (Soley-Bori, 2013). It is possible to do multiple imputation when data are missing not at random (MNAR), but to do that, you first need to specify a model for the missing data mechanism—that is, a model of how missingness depends on both observed and unobserved quantities (Allison, 2014).

One of the benefits of multiple imputation is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference (Roy, 2019). Another reason to use multiple imputation is that it separates the solution of the missing data problem from the solution of the complete-data problem. The missing-data problem is solved first, the complete-data problem next. Though these phases are not completely independent, the answer to the scientifically interesting question is not obscured anymore by the missing data. The ability to separate the two phases simplifies statistical modeling, and hence contributes to a better insight into the phenomenon of scientific study (Van Buuren, 2018).

Multiple imputation on the other hand, can be challenging to implement given the different steps one has to go through to arrive at the final imputation. The validity of the multiple imputation results will be questionable if there is an incompatibility between the imputation model and the analysis model, or if the imputation model is less general than the analysis model (Raghunathan, 2004b). Lastly, multiple imputation is computationally intensive and involves approximations. Some algorithms need to be run repeatedly in order to yield adequate results, and the required run length increases when more data are missing (Sterne et al., 2009).

1.3.3 Full Analysis methods

There are methods which take full account of all information available, without the distortion resulting from using imputed values as if they were actually observed. The full analysis methods are direct methods unaffected by, and robust to, missing values. These methods are sometimes referred to as likelihood based functions since it employs the likelihood function in its analysis of data. Full analysis methods are more advanced and modern. **The Maximum Likelihood (ML)** approach of estimating missing data is sometimes referred to as “Full Information Maximum Likelihood(FIML)”, “direct maximum likelihood” or “raw maximum likelihood” (Newsom, 2020). The underlying assumptions that needs to be met before using ML is that the data should be missing at random (MAR) and should follow a multivariate normal distribution (Statistics, 2020). Under maximum likelihood estimation, the full, incomplete dataset is analyzed without imputing data (Grace-Martin, 2020c).

In carrying out ML estimation, the value of some population parameter is estimated by determining the value that maximizes the likelihood function (actually the log of this function) based on the sample data that is available (Statistics, 2020). Rather than imputing the data values, this method uses each case available data to compute maximum likelihood estimates. The maximum likelihood estimate of a parameter is the value of the parameter that is most likely to have resulted in the observed data. When data are missing, we can factor the likelihood function. The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables. These two likelihoods are then maximized together to find the estimates. That is, the missing data may be estimated by using the conditional distribution of the other variables (Kang, 2013). This method gives unbiased parameter estimates and standard errors.

One advantage of ML estimation is that it does not require the careful selection of variables used to impute values that multiple imputation requires. It is, however, limited to linear models (Grace-Martin, 2020c). Under the identified assumptions above being satisfied, ML produces estimates that are consistent, asymptotically efficient and asymptotically normal (Allison, 2012b).

Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods (Roy, 2019). The essential idea behind the EM algorithm is to calculate the maximum likelihood estimates for the incomplete data problem by using the complete data likelihood instead of the observed likelihood because the observed likelihood might be complicated or numerically infeasible to maximize (Unknown, 2020).

Under the Expectation-Maximization algorithm, there are two main steps involved: Expectation-step (E-step) and Maximization-step (M-step). This method begins with the Expectation-step (E-step) where the observed data is augmented with manufactured data so as to create a complete likelihood that is computationally more tractable. We then replace, at each iteration, the incomplete data, which are in the sufficient statistics for the parameters in the complete data likelihood, by their conditional expectation given the observed data and the current parameter estimates (Unknown, 2020). It must be noted that the parameter estimates are used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data (Roy, 2019). The new parameter estimates are obtained from these replaced sufficient statistics as though they had come from the complete sample (Roy, 2019). Alternating E- and M-steps, the sequence of estimates often converges to the MLEs under very general conditions (Unknown, 2020).

Expectation maximization is applicable whenever the data are missing completely at random (MCAR) or missing at random (MAR), but unsuitable when the data are not missing at random (Moss, 2016). Aside missing data, the EM Algorithm can be used for the latent variables (i.e variables that are not directly observable and are actually inferred from the values of the other observed variables) in order to predict their values with the condition that the general form of probability distribution governing those latent variables is known to us (Bhadauria, 2019). It is always guaranteed under this algorithm that likelihood will increase with each iteration (Bhadauria, 2019).

The EM algorithm can be slow to convergence. It works best when you only have a small percentage of missing data and the dimensionality of the data isn't too big. The higher the dimensionality, the slower the E-step; for data with larger dimensionality, you may find the E-step runs extremely slow as the procedure approaches a local maximum (Glen, 2015). Again, the EM algorithm cannot be used to obtain direct estimates of linear model parameters. Although the imputed values are optimal statistical estimates of the missing observations, they lack the residual variability present in the hypothetically complete data set; the imputed values fall directly on a regression line and are thus imputed without a random error component (Enders, 2001).

The Expectation-Maximization algorithm forms the basis of many unsupervised clustering algorithms in the field of machine learning (Bhadauria, 2019). It is applicable in the estimation of Gaussian Mixture Models (GMM) (Glen, 2015). A mixture model is a model comprised of an unspecified combination of multiple probability distribution functions. The Gaussian Mixture Model, or GMM for short, is a mixture model that uses a combination of Gaussian (Normal) probability distributions and requires the estimation of the mean and standard deviation parameters for each. The EM algorithm is therefore an appropriate approach to use to estimate the parameters of the distributions (Brownlee, 2020c). The EM algorithm is also frequently employed to estimate model parameters for imputation in scRNA seq methods such as scImpute, scRNA-seq complementation (SCC), Zero-Inflated Factor Analysis (ZIFA). [(Hu et al., 2020), (Harel et al., 2014)]

1.4 Motivating Example for Missing Data Imputation: Single-Cell RNA Data

Single-Cell RNA sequencing (scRNA-seq) is becoming an increasingly popular method in the fields of Cellular Biology and Bioinformatics and across a large number of other biological disciplines including Genomics, Developmental Biology, Neurology, Oncology, and Immunology. scRNA-seq is a high throughput analysis that enables researchers to understand at the single-cell level what genes are expressed, in what quantities, and how the gene expression levels differ across thousands of cells within a heterogeneous sample(s) (Company, 2020). Single-cell RNA sequencing (scRNA-seq) technology provides an effective way to study cell heterogeneity, discover new cell types, and understand cell development at single-cell resolution. More specifically, scRNA sequencing is a method for measuring the transcriptome-wide gene expression in single cells (van Djik et al., 2018).

One major challenge associated with the analysis of the scRNA-seq data is that the scRNA-seq data is characterized by a high percentage of missing values. This is mainly as a result of low capture efficiency and stochastic gene expression (Yang et al., 2018). From the roughly 100,000 to about 300,000 mRNAs present in a cell, only 10% - 40% are captured using current scRNA-seq protocols (van Djik et al., 2018). For example, in a mouse embryo cell, the missing rate can reach nearly 30%, even after noise reduction (Yang et al., 2018). Single-cell RNA-seq (scRNA-seq) data usually contain many zero expression values. This is different from the typical structure for a missing dataset in a lot of studies which mostly contain blanks for missingness. The zeros that are biologically driven (such as genes that do not express RNA at the time of measurement) are referred to as the true zeros, whilst those that are technically driven (such as genes that express RNA, but not at a sufficient level to be detected by sequencing technology) are referred to as dropout zeros. These dropout zeros are designated as the missing data values in scRNA-seq data (Yang et al., 2018). This means that not all observed zeros are considered missing data values in scRNA-seq data. The dropouts increase the cell-to-cell variability, leading to signal influence on every gene, and obscuration of gene-gene relationship detection thereby affecting downstream analyses (Chen et al., 2019).

Given the high level of missingness associated with scRNA seq data, there is the need to carefully address the missingness so as to avoid the key statistical problems that come with missing data as outlined above. With the high fraction of the missing data values present in scRNA seq data, omission methods such as pairwise deletion and complete case analysis discussed previously may not be appropriate techniques to use when handling scRNA data. The direct deletion of the missing data can result in a loss of valuable information and affect the downstream analyses of the data (Yang et al., 2018). Due to technical limitations and biological factors, scRNA-seq data are noisier and more complex. The high variability of scRNA-seq data raises computational challenges in data analysis. To effectively handle the high variability of scRNA-seq data, attention should be paid to choosing appropriately analytical approaches (Chen et al., 2019).

Several technologies have been developed in recent years which attempt to address the challenge of sparsity in scRNA-seq data. Imputation is a common and useful approach that is employed to handle the issue of missingness in scRNA-seq data. However, one challenge when imputing expression values is to distinguish true zeros from missing values (dropout zeros) (Andrews and Hemberg, 2018a). Many of the scRNA-seq imputation methods such as SAVER (Huang et al., 2018a), scImpute (Li and Li, 2018), and DrImpute (Kwak et al., 2017) use models of the expected gene expression distribution to distinguish true biological zeros from zeros originating from technical noise. Alternatively, some scRNA-seq imputation methods like MAGIC (van Dijk et al., 2017) perform data smoothing to reduce noise present in observed values by using information from neighboring data points (Andrews and Hemberg, 2018a). The above mentioned scRNA-seq data analysis methods have their pros and cons which are briefly discussed below based on their mechanisms.

SAVER (Single-cell Analyses Via Expression Recovery) (Huang et al., 2018a) is an expression-recovery method for denoising single-cell RNA sequencing data by borrowing information across genes and cells to provide accurate expression estimates for all genes. SAVER assumes that the count of each gene in each cell follows a Poisson-Gamma mixture, also known as a negative binomial model. Instead of specifying the Gamma prior, we estimate the prior parameters in an empirical Bayes-like approach with a Poisson Lasso regression using the expression of other genes as predictors. Once the prior parameters are estimated, SAVER outputs the posterior distribution of the true expression, which quantifies estimation uncertainty, and the posterior mean is used as the SAVER recovered expression value (Huang et al., 2018b).

SAVER takes advantage of gene-to-gene relationships to recover the true expression level of each gene in each cell, removing technical variation while retaining biological variation across cells. This method is able to accurately recover the gene expression distribution, which is important for identifying rare cell types, identifying highly variable genes (Huang et al., 2018b). However, SAVER may lead to expression changes of the genes unaffected by dropouts introducing new bias and possibly eliminating meaningful biological variation (Li and Li, 2018). SAVER tends to treat all zero expressions as missing values. This is inappropriate since some of them may reflect true biological non-expression and be a result of gene expression stochasticity (Li and Li, 2018). Finally, because SAVER relies on Markov Chain Monte Carlo algorithms to infer parameters, it is computationally intensive and might not be suitable for large datasets (Li et al., 2019).

MAGIC (Markov Affinity-based Graph Imputation of Cells) (van Dijk et al., 2017) is a method for imputing missing values restoring structure of large biological datasets (van Dijk et al., 2017). MAGIC leverages the large sample sizes in scRNA-seq (many thousands of cells) to share information across similar cells, via data diffusion, to denoise the cell count matrix and fill in missing transcripts (van Dijk and Gigante, 2019). MAGIC imputes likely gene expression in each cell, revealing the underlying biological structure. It uses signal-processing principles similar to those used to clarify blurry and grainy images (Van Dijk et al., 2018).

MAGIC is effective at recovering gene-gene relationships and additional structures (Van Dijk et al., 2018). This imputation method is able to impute complex and non-linear shapes of interactions(van Dijk et al., 2017). Despite these advantages, MAGIC can lead to over-smoothing and may remove natural cell-to-cell stochasticity in gene expression shown to lead to biologically meaningful variations in gene expression. This is because the imputation is based on information shared across similar cells (Huang et al., 2018b). Lastly, MAGIC tends to treat all zero expressions as missing values. This is inappropriate since some of them may reflect true biological non-expression and be a result of gene expression stochasticity (Li and Li, 2018).

DrImpute is an ensemble method based on consensus clustering method. It performs clustering many times and conducts imputation by the average value of similar cells. More specifically, under the drImpute algorithm, the cell-cell distance matrix is computed using Spearman and Pearson correlations, followed by the cell-wise clustering based upon the distance matrix over a range of expected number of clusters k (k ranging from 10 to 15 by default). Each combination of distance metric (Spearman or Pearson) and k, the clustering results, are used to impute dropout entries. Then the averaged estimation over all combinations are taken as the final imputed values (Gong et al., 2018).

DrImpute can effectively separate the dropout zeros from true zeros. Also, DrImpute can significantly improve the performance of existing tools like PCA and t-SNE in visualizing scRNA-seq data by imputing dropout events (Gong et al., 2018). One of the limitations of DrImpute is that considers only cell-level correlation leaving out gene-level correlation (Gong et al., 2018).

ScImpute (Li and Li, 2018) is developed to accurately and robustly impute the dropout values in scRNA-seq data. It simultaneously determines which values are affected by dropout events in data and perform imputation only on dropout entries. scImpute also detects outlier cells and excludes them from imputation (Li and Li, 2018). Just like MAGIC, scImpute directly estimates the true expression levels by relying on pooling the data for each gene across similar cells (Huang et al., 2018b). ScImpute models log-normalized expression values as a mixture of gamma-distributed dropouts and normally-distributed true observations (Andrews and Hemberg, 2018a).

ScImpute can impute the dropout values without introducing new biases through using the information from the same genes unlikely affected by dropouts in other similar cells. Also, scImpute is able to effectively separate the dropout zeros from true zeros (Chen et al., 2019). ScImpute enhances the clustering of cell subpopulations and improves the accuracy of differential expression analysis (Chen et al., 2019). However, scImpute can lead to over-smoothing and may remove natural cell-to-cell stochasticity in gene expression shown to lead to biologically meaningful variations in gene expression. This is because the imputation is based on information shared across similar cells (Huang et al., 2018b).

Looking at the scRNA-seq tools discussed(SAVER, MAGIC, DrImpute, scImpute), it is evident their algorithms employ the basic imputation techniques in handling missing data discussed previously. DrImpute uses the mean imputation technique by averaging the expression values from similar cells. MAGIC and ScImpute have embedded the nearest neighbor algorithm (KNN) by imputing missing expression values by sharing information across similar cells. SAVER uses some element of regression imputation in its algorithm by Poisson Lasso regression using the expression of other genes as predictors.

1.5 Below-limit-of-detection

Data sets containing values below the limit of detection (LOD) are known as ‘censored data sets' (Barescut et al., 2011). The problem of censored data, in which the observed value of some variable is partially known, is related to the problem of missing data, where the observed value of some variable is unknown. Missing data due to limit of detection is a common obstacle in epidemiological and biomedical research (Harel et al., 2014). The Limit of detection or LOD is the lowest true value that can be measured (detected) with statistical significance by means of a given analytical procedure. This analytical threshold, LOD, is practically determined on the basis of noise level using a given measurement procedure (ScienceDirect, 2020). This measured quantity value should be distinguishable from background noise or the absence of that quantity (blank value) with some stated level of confidence (generally 99%) (Harel et al., 2014). Mostly observations that fall below the LOD are not reported and are thus creating missingness in data for further data analysis. This incomplete setup might cause bias, inefficiency and in most cases will make the analyses more complex (Harel et al., 2014).

It is very necessary to decide how values that fall below the limit of detection should be treated as they can affect statistical analyses to some extent. The most common technique used when handling below the limit of detection cases is the Complete Case Analysis (CCA), where observations that fall below the limit are simply eliminated. Another popular method is single imputation where every value below the LOD is replaced by a constant such as $LOD/2$ or $LOD/\sqrt{2}$ (Harel et al., 2014) or substituting with a value between zero and LOD value. In other instances, values below the LOD are treated as zeros (Barescut et al., 2011). More complex methodology like Maximum Likelihood (ML), Bayesian analysis, and Multiple Imputation have become more prominent in recent years to account for below the limit of detection cases (Harel et al., 2014).

Though CCA, single imputation, and the other substitution methods mentioned are easy to employ, the resultant non-censored (complete) data set will have certain limitations and deficiencies when it comes to subsequent data analysis and interpretation. With CCA, subsequent analysis commonly results in biased estimation, while analysis using single imputation will result in impaired estimates of variances and covariances (Harel et al., 2014). Substituting with zero will bias the estimates low and using LOD as substitutes could bias the results high (Morton and Lion, 2016). CCA and substitution methods may be appropriate to use when there is a very low percentage of the dataset falling below the limit of detection (Barescut et al., 2011). The United States Environmental Protection Agency (EPA) Unified Guidance suggests that the substitution method can be acceptable when only a small portion of the data set (10-15 percent) consists of below the limit of detection

(ITRC, 2013).

Bayesian and ML methods can provide unbiased and more efficient estimation but are often dependent on strong assumptions and are more difficult to apply in practice (Harel et al., 2014). Multiple imputation (MI) methods yield valid and robust parameter estimates and explicit imputed values for variables that can be analyzed as outcomes or predictors. The distribution-based MI method is a valid and feasible approach to analyze bivariate data with values below LOD, especially when explicit values for the non-detections are needed (Chen et al., 2011).

A variety of survival analysis methods have also been proposed for handling values below the LOD depending on the purpose of analysis. The non-parametric Kaplan-Meier method, MLE, and robust Regression on Ordered Statistics (ROS) can be used to estimate the summary statistics for censored data. For group comparisons in censored data, methods like Generalized Wilcoxon test and censored regression (with 0/1 indicator) will be appropriate. Lastly, when performing a linear regression on a censored dataset, logistic regression or proportional hazard (Cox) regression can be used (Barescut et al., 2011).

It must be noted that data below the LOD are informative, for instance, in chemical analysis, as they indicate that the analyte (component of interest in a sample) has a concentration between 0 and LOD, and simply excluding such values from analyses may substantially bias results (Chen et al., 2011). It is therefore necessary to handle below the limit of detection cases carefully.

1.6 Zero-inflated vs Missing data

We are often quick to leave out values that are zeros or missing (NA) when analyzing data because we think they “lack information”. However, these data points could be critical pieces of information when analyzing a research problem such that the "lack" of information is actually information. Sometimes zero in a dataset may mean “nothing” or a situation where the outcome of interest does not occur but is not unknown (missing). Data can be missing for a number of reasons but understanding why it is missing or zero can be critical in learning more and making better decisions (Ivanecky, 2020). It is therefore important as a researcher to understand the origin of the dataset to know why there are zeros or missingness (NAs). You should only replace missing values by zero if you have good reason to believe that the actual values, were they known, would be zero. In any other circumstance, this will not be appropriate and will lead to biased results (Schechter, 2016).

Zero-inflated data is very common in a wide variety of disciplines including Biostatistics, Bioinformatics, Psychology, Environmental sciences, etc. Zero-inflated data refers to count data that contain excessive number of zeros. These excessive zeros are usually too large that the data do not readily fit standard distributions (e.g. normal, Poisson, binomial, negative-binomial, and beta) (Zuur et al., 2009). Two kinds of zeros are thought to exist in a zero-inflated dataset, "true zeros" and "excess zeros"(zero counts greater than what is expected by the distribution used to model the data) (Gebregziabher, 2019). In lay terms, true zeros occur where there could have been an event, but there was none whilst excess zeros refer to those zeros where there could not be any event (Wayne, 2011).

In a classic example of zero-inflated data (Statistical Consulting Group, 2020), suppose state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked whether or not they have a camper, how many people were in the group, were there children in the group, and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish (Statistical Consulting Group, 2020). With this example there is an excess zero problem of not knowing whether a person fished or not.

Generally, since the excess zero problem will always be present at some point when analyzing zero-inflated data, it will be necessary that the researcher carefully understand what kind of zero is present. It will be therefore be helpful if methods are developed to help researchers distinguish straight away true zeros from excess zeros given any zero-inflated dataset (a possible topic for future work). Identifying the source of zeros will be key in order to select the most appropriate statistical model. scRNA data is a typical example of data with excessive number of zeros (zero-inflation). Such data consist of true zeros and dropout zeros (missing), as well as the potential for data below the LOD. Due to the excess zero problem, a number of methods have been developed in recent years to detect true zeros from dropout zeros under scRNA technology (see section 1.4) .

1.7 Dissertation Format

This is a three-paper dissertation. The format of this dissertation will include an introduction or overview (the current chapter), followed by three main chapters and a summary chapter or conclusion. The introductory chapter discusses an overview of missing data, which is the central theme of this dissertation, including relevant background and conceptual framework. The three main chapters that follow the overview or introductory chapter will be standalone papers that will be submitted to various publication outlets. Each of the three papers will have an abstract, a literature review, methodology, results and future work sections.

The paper in Chapter 2 addresses the dropout phenomenon in single-cell RNA (scRNA) sequencing through a comparative analyses of some existing scRNA sequencing techniques. In Chapter 3, the paper considers a simulation study to assess whether it will be appropriate to address the issue of non-detects in data using a traditional substitution approach, imputation, or a non-imputation based approach. Lastly, the paper in Chapter 4 presents an efficient strategy to address the issue of imbalance in data at any degree (whether moderate or highly imbalanced).

A discussion section was added that directly discusses the implications and interpretations of the chapter findings as well as identifies limitations of the study. The final chapter, which is the conclusion, links the findings of all three papers into a coherent research contribution in the field of data analysis with missing data.

CHAPTER 2

COMPARATIVE ANALYSIS OF STATISTICAL METHODS FOR SINGLE-CELL RNA SEQUENCING

Abstract

The dropout phenomenon in single-cell RNA sequencing (scRNA-seq) data has provoked a lot of controversial questions to be asked amongst researchers in fields utilizing this kind of data. The dropout zeros, designated as the missing values, are mostly not distinguishable from the true zeros (i.e., zeros arising from biological factors) and can bias the results of downstream analyses in scRNA sequencing. The question: "whether imputation is a necessary step in scRNA-seq analysis?" remains unresolved as there are different schools of thought on how to handle dropouts in scRNA. To answer this, a simulation study was conducted evaluating the average Type I error rate and power from four popular scRNA-seq imputation methods (MAGIC, SAVER, DrImpute, and scImpute) combined with three differential test methods (DESingle, MAST, and Seurat). A case of no imputation before differential expression testing was also considered. MAGIC was found to consistently outperform all the other methods (including no imputation scenario) when combined with all the differential testing methods, yielding the lowest Type I error and highest power across 100 simulation runs. It was therefore concluded that imputation is a crucial step in scRNA-seq analysis.

2.1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a high throughput analysis that enables researchers to understand the gene expression within a single cell, in what quantities the genes are expressed, and how they differ across thousands of cells within a heterogeneous sample, e.g., early embryo development (Company, 2020). Single-cell RNA data are mostly characterized by a high percentage of missing values (or zeros) due to technical limitations and stochastic gene expression. This could pose a major problem as the missing data can introduce bias and affect downstream analyses. According to van Djik et al. (2018), from the roughly 100,000 to about 300,000 mRNAs present in a cell, only 10% - 40% are usually captured using current scRNA-seq protocols. Following this statistic, scRNA data can contain as high as 90% zeros. This high proportion of zero read counts that are expressed by many genes in scRNA data has therefore become another major concern for discussion. The zeros present in a scRNA dataset may be biological or technically driven as discussed in section 1.4. A major challenge mostly encountered when dealing with these zeros is that they are not easily distinguishable without biological knowledge or spike-in control. The biological or true zeros are seen to carry meaningful information about cell states while the technical zeros or dropouts represent missing values artificially introduced during the generation of scRNA-seq data (Jiang et al., 2022).

The sparse nature of scRNA data has raised a lot of computational and interpretability concerns. According to He et al. (2021), the zero inflation can introduce bias in differential expression test analysis and pose challenges in detecting differentially expressed (DE) genes. Jiang et al. (2022) added that the dropouts (missing values) can impede the full and accurate interpretation of cell states and the differences between them. This has given rise to varied opinions on how the dropout situation should be handled in the scRNA-seq field. There is a proposal by Qiu (2020) to embrace dropouts as useful information rather than treating them as a problem to be fixed. To do this, they generated a dropout pattern by binarizing the single-cell RNA-seq count data (i.e. turning all non-zero observations into one), and based on the dropout pattern used a co-occurrence clustering algorithm to identify cell populations. Qiu (2020) recommended that recognizing the utility of dropouts would provide an alternative direction for developing computational algorithms for single-cell RNA-seq analysis.

Several imputation methods have also been developed over the years, to address the issue of missingness (dropouts) in single-cell data. However, a paper by Andrews and Hemberg (2018b) argued that imputation of scRNA data can introduce circularity that can generate false-positive results and that SAVER was the least likely to generate false or irreproducible results and thus should be favoured over alternatives if imputation is necessary. This raises an important question: Is imputation a necessary step in single-cell data analyses?

Another tough question that comes up when performing any single-cell data analysis is "which technique is best for analysing scRNA data?". This question is often asked as researchers are faced with a number of methods to choose from and also as every method has it own pros and cons. This can be with regards to performing an imputation, testing for differential expression, or even visualizing scRNA data in an analysis. Many researchers have attempted to do a comparative analysis on which single-cell technique (e.g., imputation technique, differential test method) is best. This however remains controversial, as more advanced imputation techniques are being developed. Some studies have assessed the reliability of some scRNA-seq methods based on metrics such as False Discovery Rate (FDR), receiver operating characteristic (ROC) curve, and the number of differentially expressed genes detected. For instance, a study by Jaakkola et al. (2016), measured reproducibility by comparing the precision and recall of the detection of all DE genes between the full data set and its subsets. Another study by Ganna et al. (2014) focused on the reproducibility of methods in terms of rediscovery rate (RDR).

In this chapter, the strength of some popular imputation and differential testing methods were assessed through simulation study. This will help answer the two key questions raised above. In detail, the imputation methods were combined with the differential test methods, and their performances were assessed based on Type I Error rate (False Positive Rate) and power. To further address the controversial issue of whether imputation is a necessary step, a case of performing differential testing without imputing scRNA data was considered. The various method combinations were evaluated based on their ability to recover the true expression of ERCC spike-ins in scRNA data, which is the "ground truth". ERCC spike-ins are a set of external RNA controls that enable performance assessment of a variety of technology platforms used for gene expression experiments (Technologies, 2012). This study to a large extent would save researchers the time when performing any differential test analysis with these selected methods. This is because the study will provide them a better insight and guidance into choosing the most appropriate technique combination to use to obtain preferably unbiased results.

2.2 Methodology

2.2.1 Motivating scRNA Data

A real data, mouse Embryonic Stem Cell (mESC) single-cell RNA-seq data with ERCC spike-ins by Buettner et al. (2015), was used in this study to simulate data to evaluate the performance of scRNA sequencing methods. This data is publicly available in the scRNAseq R package under Bioconductor. The Buettner mESC data count matrix consists of 38,293 genes (rows) and 288 cells (columns) with ERCC spike-ins (i.e, 92 bacteria genes spiked in at known concentrations). Counts of ERCC spike-ins are made up of 92 rows and 288 columns. The 288 cells are made up of 3 cell-cycle fractions (G1, S, and G2M phase) with 96 cells per cell phase.

2.2.2 Methods

Although there are several imputation and differential test methods specifically designed for scRNA-seq data, this study will only focus on a few of the popular methods. Four imputation methods: MAGIC (Markov Affinity-based Graph Imputation of Cells) (van Dijk et al., 2017), DrImpute (Kwak et al., 2017), scImpute (Li and Li, 2018), and SAVER (Single-cell Analyses Via Expression Recovery) (Huang et al., 2018c) were combined with three differential expression (DE) test methods: DESingle (Miao et al., 2018), MAST (Model-based Analysis of Single-cell Transcriptomics (Finak et al., 2015), and SEURAT (Satija et al., 2015).

The scRNA-seq imputation techniques mentioned in this paper have been discussed in section 1.4, highlighting their pros and cons, as well as how their imputation algorithms work on scRNA-seq data. The differential test method, MAST, fits two-part, generalized linear models that are specially adapted for zero-inflated single cell gene expression data and tests differential expression between groups based on likelihood ratio testing (Finak et al., 2015). DESingle, on the other hand, uses the Zero-Inflated Negative Binomial model (ZINB) to detect differentially expressed genes and can discriminate between real and true zeros (Miao et al., 2018). By default, Seurat performs differential expression based on the non-parameteric Wilcoxon rank sum test. It also supports other differential tests such as DESeq2 and MAST and can be used to pre-filter genes or cells before performing any analysis (Satija et al., 2015).

Each imputation method was applied to either the raw mESC read count matrix or log-transformed form as appropriate or intended by the authors. All the imputation methods except for DrImpute required a log-transformed matrix per documentation. The workflow for each simulation analysis in this study was highly dependent on the DE test method under consideration. MAST expects that a log-transformed approximately scale-normalized data is provided for DE testing (McDavid, 2022). On the contrary, Seurat and DESingle expect a raw read count matrix. To use the raw read count matrix in Seurat, the raw data must first be converted to a Seurat object using the CreateSeuratObject function under the Seurat package. To test for DE with Seurat on the raw imputed and non-imputed data, data pre-processing steps like filtering genes with low expression, normalization, and log transformation were carried out using functions in the Seurat package. By default, Seurat implements a global-scaling normalization method “LogNormalize” that normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result Drou et al. (2022). Normalization and log-transformation of the raw imputed and non-imputed data in Seurat was done using the NormalizeData function and differential testing with Seurat was done with the FindMarkers function.

All imputed and non-imputed data were log-normalised before getting passed to MAST using the scran package by Lun et al. (2016). Differential testing of expression levels between cell phases (G1, S, G2M) was performed using the LRTest function in the MAST package. DESingle automatically performs data pre-processing steps like normalization and filtering out genes with low expression when testing for DE so there was no need to perform these steps prior. The only data pre-processing step that was done before passing the imputed or non-imputed data to DESingle was filtering out low quality cells using the scran package by Lun et al. (2016). Differential testing was then carried out using the DESingle function. Imputed data from DrImpute was unlogged before getting passed to DESingle for testing. This is because DESingle expects the input data to be on a raw unnormalized scale.

Unlike MAST and Seurat, DESingle integrates parallel computing functionality with the BiocParallel package under Bioconductor. Due to the high dimensionality of the simulated data, parallelization was enabled under DESingle before testing. The run times for MAST and Seurat were therefore shorter than DESingle. Parallel computing is highly recommended for SAVER and scImpute in order to reduce computation time and avoid running into memory issues, as these imputation methods are computationally intensive. All parallel computing was done in this paper by setting the number of cores to 16. Unless stated otherwise all imputation methods were run with default parameters (see Table 2.1).

Table 2.1: Summary of scRNA-seq Imputation methods

Method	Input	Model	Simulation Parameter	Reference
MAGIC	Raw count	–	knn = 10	(van Dijk et al., 2017)

ScImpute	Raw Count	Gamma-Normal mixture	drop.thre =0.5
			Kcluster = 3
				(Li and Li, 2018)

SAVER	Raw Count	Poisson-Gamma Mixture	size-factor = 1	(Huang et al., 2018c)
DrImpute	$log_2(\text{Raw Count}+1)$	ZINB	ks = 10:15	(Kwak et al., 2017)

2.2.3 Simulation

ERCC spike-ins were used to simulate the scRNA-seq data in this study because they are synthetic RNA molecules with known concentrations, which can be manipulated to serve as a ground

Table 2.2: Summary of Differential Test methods

Method	Input	Model	Test	Reference
DeSingle	Raw count	ZINB	Likelihood Ratio Test	(Miao et al., 2018)
MAST	log_2(Raw Count $+ 1$)	Generalized linear hurdle model	Likelihood Ratio Test	(Finak et al., 2015)
Seurat	log_2(Raw Count $+ 1$)	–	Wilcoxon Rank Sum Test	(Satija et al., 2015)

truth for the performance evaluation of the various methods discussed. To ensure that results obtained after a single run of analysis were not just by chance, simulations were run multiple times (i.e 100 simulation runs). In the end, the accuracy of the methods were evaluated based on the average Type I error and average power across the 100 simulation runs. The Buettner mESC data were analysed in two ways as detailed below.

2.2.4 Ground Truth Simulation

For the Type I error estimation case, the read count data for the 92 ERCC spike-ins were first permuted (within spike-in gene) to simulate no differential expression between the cell phases. The expression values of the first 92 genes in the original Buettner mESC count matrix were then replaced with the permuted ERCC spike-ins. All necessary data pre-processing steps were carried out on simulated data, and imputed data were passed on to MAST, Seurat, and DESingle to test for differential expression (DE). The false discovery rate (FDR) was controlled at a 0.05 level with the Type I error rate estimated by the percent of the 92 spike-ins that were called significant after FDR correction.

To evaluate performance based on the power, the first 92 gene expression in the original Buettner mESC data were replaced by the permuted 92 ERCC spike-ins, as dicussed above. After, each cell phase (G1, S, G2M) for the first 92 "spike-in" genes in the matrix, was multiplied by the factors (5, 8, 10) respectively. The multiplication was to introduce differential expression at specific magnitudes (fold changes). All necessary data pre-processing steps were then carried out, with imputed data passed on to the DE methods. The FDR was controlled at 0.05 and the power was computed as the percent of 92 spike-ins found to be significant.

Using these as ground truths, the power and Type I error that each method combination contributed were estimated.

2.2.5 Data Pre-processing

Single-cell RNA data tend to be very noisy due to the prevalence of dropouts and other technical factors. As a result, the quality of scRNA data is typically poor. It is therefore not advisable to use raw scRNA data for analysis without accounting for the noise. This can pose computational challenges and result in biased downstream analyses. In view of this, the data pre-processing steps Quality Control (QC) and Normalization were carried out on the simulated data before testing for differential expression.

Quality control to remove low quality cells was done using the scran (Lun et al., 2016) package from Bioconductor. Low quality cells in scRNA-seq data need to be removed to ensure that technical effects do not distort downstream analysis results.

One measure of low quality are cells with relatively small library sizes (i.e., total sum of counts across all features (genes)). High proportions of spike-in RNA is also indicative of low quality cells since the quantity of spike-in RNA added to each cell is expected to be constant. This means that the proportion would increase upon loss of endogenous RNA in low-quality cells (Lun et al., 2016). The perCellQCMetrics and quickPerCellQC functions were used to filter out low quality cells in the simulated mESC data. Genes that were expressed in less than 10 cells were removed from simulated data. About 271 cells remained after basic quality control (94 G1, 81 S, and 96 G2M cells).

Normalization is another necessary step that helps to remove the influence of technical effects in the underlying counts, while preserving true biological variation (Hafemeister and Satija, 2019). The normalized expression values were log-transformed, except for when testing imputed or non-imputed data with DESingle because it expected raw unnormalized counts. The advantage of log-transformation is to prevent a few large observations from being extremely influential, and make the transformed values continuous, allowing for greater flexibility for modeling (Li and Li, 2018).

Due to the high dimensionality of the mESC simulated data and the number of simulations to be performed, simulations were parallelized and run through the high performance computing platform at the University of Utah. This reduced the computational time which could have taken months on a standard laptop (which requires large pools of memory) to about 10 days on the high performance computing platform. Simulated data were imputed after quality control, before differential expression was tested between pairs of the cell phases (G1, S, and G2M). The other data pre-processing steps (i.e., normalization and log-transformation) were done unique to the input requirements of the various imputation and DE methods (as shown in Tables 2.1 and 2.2).

2.3 Results and Discussion

Figure 2.1 summarizes the average Type I error rate estimation results across a 100 simulation runs focusing on the the first 92 "spike-in" genes as the ground truth. Here, MAGIC systematically had the lowest Type I error rate when combined with any of the DE methods (DESingle, MAST or Seurat). On the other hand, DrImpute consistently lost control over the Type I error across all the DE methods but had a relatively lower false positive rate with MAST. The performances of SAVER, scImpute, and no imputation were similar, yielding a controlled Type I error rate with MAST and Seurat, but a higher Type I error rate with DESingle.

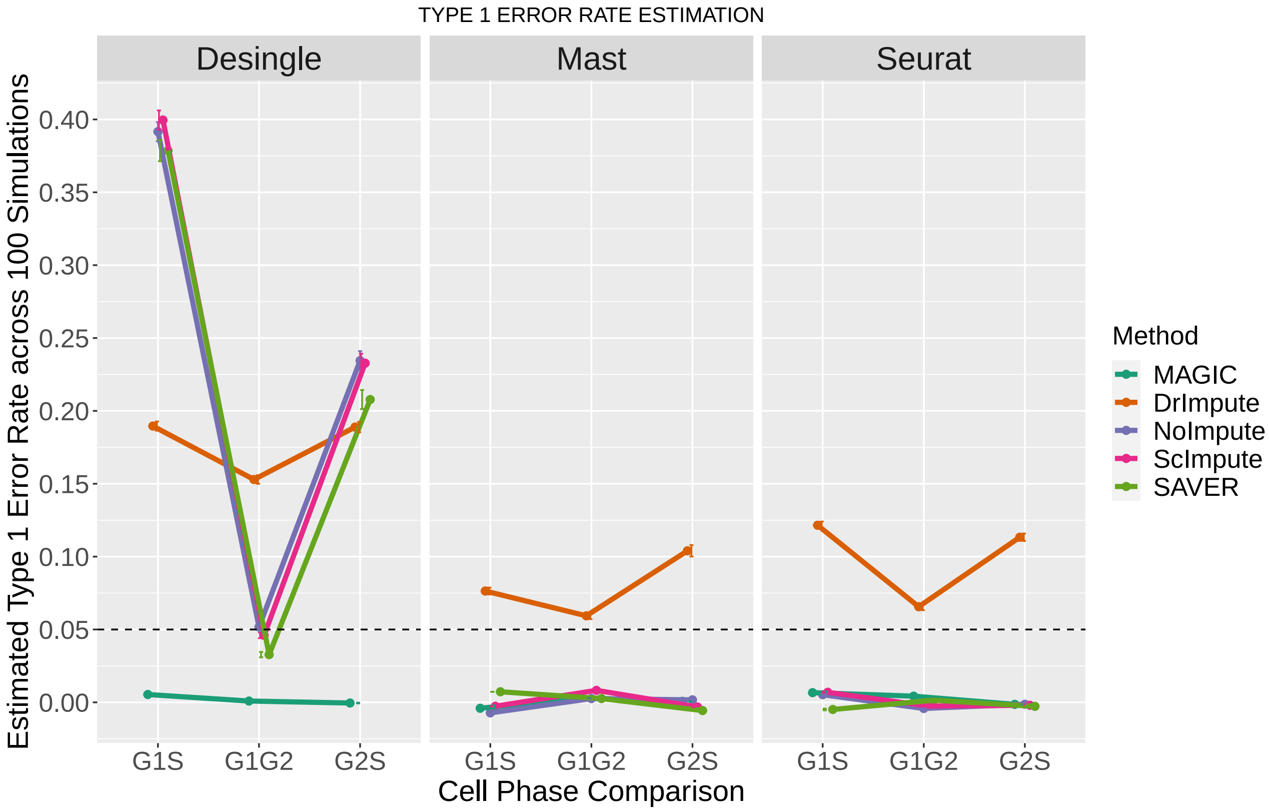


Fig. 2.1: Average Type I error across 100 simulations for the G1 vs S, G1 vs G2 and G2 vs S comparisons. Data were slightly jittered for visualization convenience.

The power estimation results in Figure 2.2, revealed that MAGIC consistently outperformed all the other imputation methods and the non-imputation method, yielding the highest power consistently across all the cell phase comparisons for each DE method. DrImpute performed better with DESingle and MAST than with Seurat. The resulting power estimates for Seurat when combined with all the imputation methods, as well as no imputation, were very poor compared to DESingle and MAST. SAVER combined with Seurat failed to detect any significant differences between the gene expression for the cell phases even though the differences existed. Recall that the cell phases (G1, S and G2M) were multiplied by the factors (5, 8, 10) respectively, hence the expected power should reflect these magnitude differences. The magnitude difference is expected to be higher for the G1 vs G2 comparison, followed by G1 vs S, and G2 vs S having the lowest difference. MAGIC matched the true profile (relative expected magnitude difference) across all the three DE methods. DrImpute performed quite well with DESingle AND MAST but failed to capture the true profile of the magnitude difference when combined with Seurat. The result for DrImpute paired with Seurat looked very different from the other methods. The power results for SAVER, scImpute and No imputation were generally below a 50%, with an interesting profile representing the magnitude difference for the cell comparisons under DESingle.

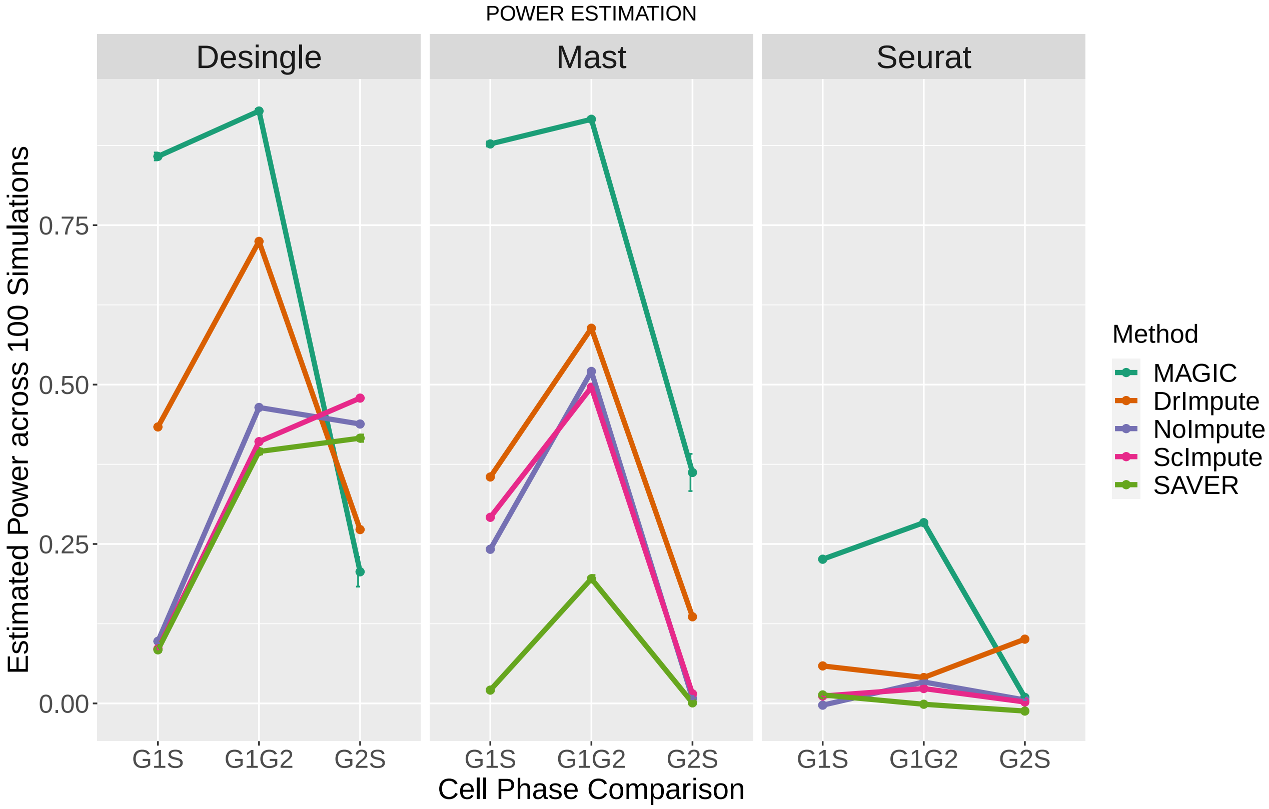


Fig. 2.2: Average Power across 100 simulations for the G1 vs S, G1 vs G2, and G2 vs S comparisons. Data were slightly jittered for visualization convenience.

The interesting findings from the power plots in Figure 2.2 spurred a further investigation to assess the accuracy of the imputation methods, as well as the no imputation scenario. The log fold change estimated after each of the different imputation methods were visualized to get a better insight into how biased the produced estimates for the differences in gene expression were. The dashed lines in Figures 2.3A, B, and C represent the true magnitude of differential expression between the cell phases. From Figure 2.3, MAGIC systematically overestimated the true magnitude of differential expression across all the cell comparisons. MAGIC consistently yielded a lower variability as compared to DrImpute, scImpute, SAVER and no imputation. The estimation of bias were pretty close to the truth for DrImpute, scImpute, SAVER and no imputation (see Figures 2.3A and 2.3C). The bias results for DrImpute, scImpute, SAVER and no imputation looked comparable. All the methods seemed to overestimate smaller magnitude differences, which is shown in the G2 vs S comparison in Figure 2.3B.

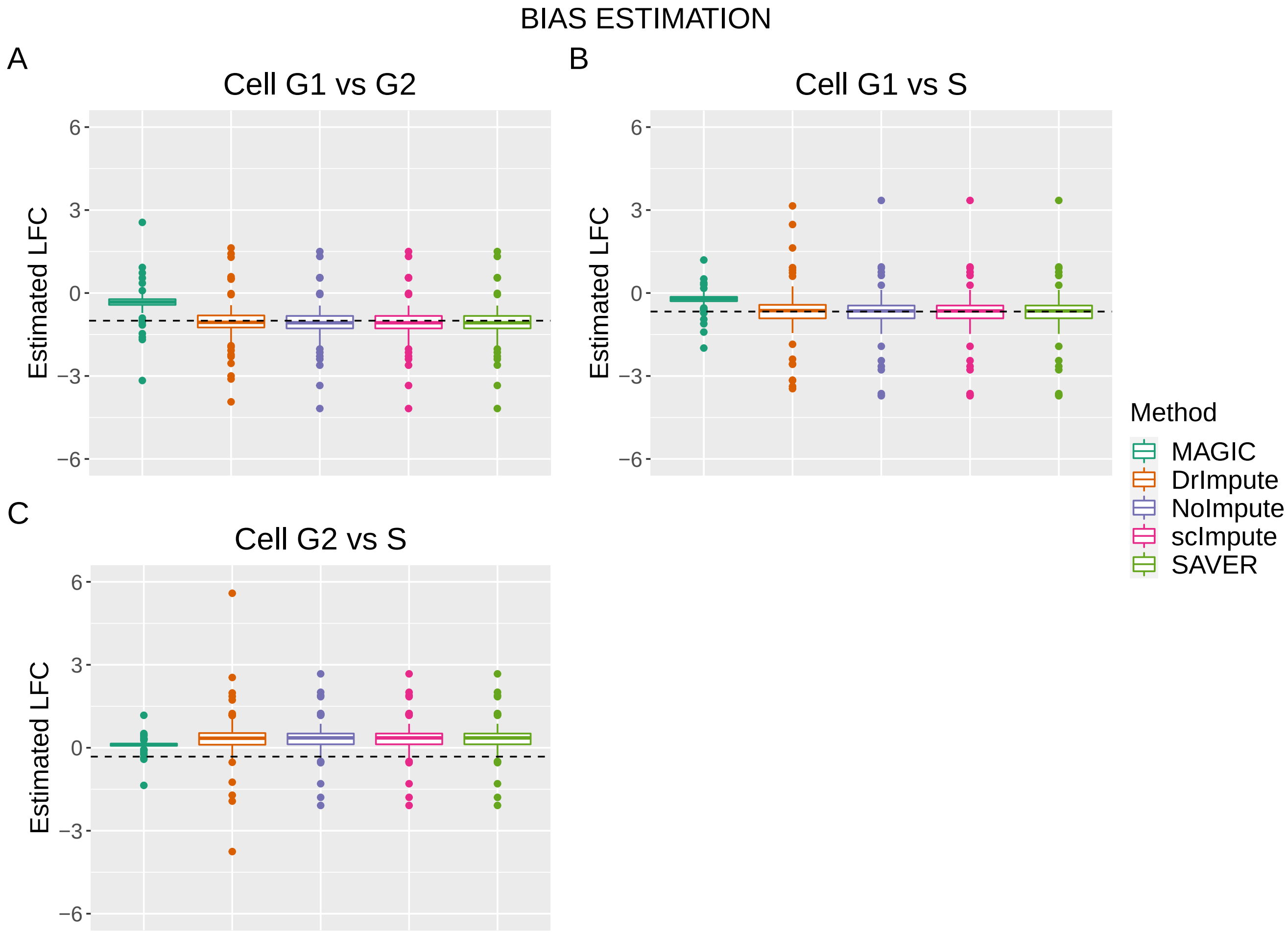


Fig. 2.3: The estimated log fold change (LFC) after different imputation methods and a no imputation case. The estimated log fold change values shown are on a log base 2 scale, and dashed horizontal reference lines indicate the true LFC for each comparison.

The conclusion reached in this study was that the best method for both imputation and differential testing varied due to different reasons. Though MAGIC had a poor bias estimation of the magnitude difference, it detected all ground truth simulated by yielding the lowest Type I error rate and highest power consistently across all DE methods. Hence, if a researcher is only interested in getting fairly higher power and lower false positive (i.e., finding whether genes are differentially expressed or not), MAGIC paired with DESingle or MAST would be recommended. This also means that the estimated log-fold change values from MAGIC cannot be fully trusted for any further analysis due to the possible overestimation (biasedness) of estimates. However, if interested in less biased estimates (rather than detection of DE genes), DrImpute would be recommended as it was the next best imputation method to yield a fairly higher power with DESingle and MAST. Judging from the results shown in Figures 2.1 and 2.2, SAVER appears to be very conservative (i.e., lower Type I error rate and fairly a lower power) and was not the best-performing as projected by some papers in literature (Andrews and Hemberg, 2018b). Results from performing differential testing without imputation generally proved that imputation is indeed a vital step in performing scRNA downstream analyses since the estimated power results without imputation were lower than that of the imputation methods used in this study.

2.4 Limitations and Future Work

A possible future consideration for this study is to investigate why the various imputation methods failed to detect smaller magnitude differences when combined with the DE methods. For instance, MAGIC was good at detecting larger magnitude differences but at smaller differences the power dropped. For example, the G2 vs S comparison had low power generally in Figure 2.2, and Figure 2.3B showed higher bias as well.

According to (Sun et al., 2019), Dimensionality Reduction (DR) is an indispensable analytic component for many areas of single cell RNA sequencing (scRNAseq) data analysis. This study proposed that proper DR can allow for effective noise removal and result in an effective downstream analyses. An extended comparative analysis could therefore be carried out extending this study, to include scRNA dimensionality reduction methods like principal component analysis (PCA), a popular technique, and t-distributed stochastic neighbor embedding (t-SNE), and how well these techniques help to improve the Type I error rate and power evaluated.

Finally, more advanced imputation methods like scRecover (Miao et al., 2019) and DCA (Eraslan et al., 2019), and differential testing techniques like Monocle (Trapnell et al., 2014) and

SCDE (Kharchenko et al., 2014) could be explored. For instance, it has been reported that scRecover when combined with other imputation methods like scImpute, SAVER, and MAGIC, not only detects dropout and real zeros at higher accuracy, but also improves the downstream clustering and visualization results (Miao et al., 2019). ScRecover would therefore be a technique worth adding to a future extension of this study and the various combinations of it with MAGIC, scImpute, and SAVER explored.

CHAPTER 3

HANDLING NON-DETECTS WITH IMPUTATION IN A NESTED DESIGN: A SIMULATION STUDY

Abstract

In this paper, a simulation study was conducted to assess whether it is ideal to address the issue of non-detects in data using a traditional substitution approach for non-detects, imputation, or a non-imputation based approach. Simulated data used were simple nested designs motivated by a real-life data in a study of bumble bee activity in a commercial cherry orchard by Kuivila et al. (2021). The simulated data were generated at different thresholds or censoring levels and at different effect sizes. For each simulated data, seven popular existing techniques to handle non-detects were applied: (i) Zero substitution, (ii) Substitution with half Limit of Detection (LOD/2), (iii) Substitution with LOD/$\sqrt{2}$, (iv) Multiple Imputation (MI), (v) Regression on Order Statistics (ROS) (Imputation approach), and (vi) Maximum Likelihood Estimation (MLE) (likelihood estimation approach) and (vii) Kaplan-Meier (KM). Multiple Imputation (MI) was not applicable as the design of the simulated data violated the assumption of having a multivariate distribution. By comparative analysis of the simulated data, substituting with LOD/2 seemed appropriate for the design simulated, as it outperformed the other techniques (i.e ROS, MLE, KM, LOD/$\sqrt{2}$, and zero substitution) by yielding a lower Type I error, lower bias, and a better power across increasing effect sizes.

3.1 Introduction

Missing data due to non-detects is a common obstacle in epidemiological and biomedical research (Harel et al., 2014). Other disciplines such as Ecology, Pharmacology, Environmental Science, and so on, are faced with the issue of non-detects too. A non-detect is an analytical sample where the concentration is deemed to be lower than could be detected using the method employed by a laboratory (Ctech.com, 1994 - 2022). More specifically, non-detects are potential low-level concentrations of organic or inorganic chemicals known only to be somewhere between zero and the laboratory's detection or reporting limits. Non-detects are sometimes referred to as left-censored data or below detection limits. Having a high level of non-detects in a dataset can be problematic as it can complicate the computations of descriptive statistics, differences among groups, correlation coefficients, and regression equations (Helsel, 2006). This in effect can lead to bias and affect drawing valid conclusions about the data. Non-detects are similar to dropouts in single-cell data as discussed in Chapter 2.

Mostly, observations that fall below the limit of detection are not reported and can create missingness in data, thus necessitating further data analysis. Unlike scRNA sequencing which represents its missing information (dropouts) as zeros, non-detects which are a form of missing data are mostly represented as “nd” in a dataset. Considering the adverse effects non-detects can pose to statistical analysis, it is very crucial that researchers understand how to properly handle non-detects and carefully select the right technique to avoid any false interpretations.

There are a number of recommended techniques for managing non-detects in data. The most commonly used method for non-detects is to substitute values for non-detects (i.e, using a fraction of the detection limit, a value between zero and the detection limit value, or simply replacing non-detects with zeros). More complex methodology like Maximum Likelihood (ML), Bayesian analysis, and Multiple Imputation have become more prominent in recent years to account for below the limit of detection cases (non-detects). A variety of survival analysis methods have also been proposed to handle values below the Limit of Detection (LOD) depending on the purpose of analysis. For instance, the non-parametric Kaplan-Meier method and robust Regression on Ordered Statistics (ROS) can be used to estimate summary statistics for censored data.

Despite considerable research in recent years on handling non-detects, regulatory agencies have published no comprehensive guidance on the recommended approach to use in a particular situation (ITRC, 2013). Some existing methods in handling non-detects or censored data include non-parametric tests (such as Kruskal-Wallis test and Wilcoxon rank-sum), omission of non-detects, substitution (for instance, using LOD or Zero), Multiple Imputation, survival analysis methods (such as Kaplan-Meier (KM) method, Maximum Likelihood Estimation (MLE), and Regression on Order Statistics (ROS). The worst practice when dealing with non-detects is to omit or delete them (Helsel, 2006). This method is not recommended because omitting non-detects from a statistical analysis can bias outcomes and prevent the statistical tests from detecting real differences (thus decreasing the statistical power of the method) (ITRC, 2013).

In Farnham et al. (2002), the procedures for handling censored data depend on the technical application involved. The study went on to say that the best method to use generally depends on the amount of data below the detection limit, the size of the dataset, and the probability distribution of the measurements. When the number of “$< LOD$” observations is small, replacing them with a constant (e.g., LOD/2, 0 etc.) is generally satisfactory. Distributional methods such as the marginal maximum likelihood estimation or more robust techniques are often required when a large number of "$< LOD$" observations are present (ITRC, 2013). The United States Environmental Protection Agency (EPA) Unified Guidance suggests that the substitution method can be acceptable when only a small portion of the data set (10-15 percent) consists of below the limit of detection (ITRC, 2013).

According to Helsel (2006), substituting values for non-detects should be used rarely and generally be considered unacceptable in scientific research. The justification the paper made was that, two decades of research has shown that this fabrication of values produces poor estimates of statistics, and commonly obscures patterns and trends in the data. Also, papers using substitution may conclude that significant differences, correlations, and regression relationships do not exist, when in fact they do. Another paper by Shoari and Dubé (2018) arguably states that, substituting non-detects with constants such as 0, LOD/2 etc. can deliberately diminish data representativeness and statistical results may be incorrectly interpreted because the uncertain measurements (i.e., non-detects) are treated as actually observed values. Some studies also claim that the MLE method is considered the "gold standard" method for dealing with non-detects, provided the data are well described by a lognormal distribution (Ganser and Hewett, 2010). In light of all these assertions, this paper seeks to clearly illustrate whether or not it is ideal to use an imputation method rather than simply substituting non-detects with constants during statistical analysis. In addition to this comparative analysis, non-imputation methods like Maximum Likelihood Estimation (MLE) and Kaplan-Meier (KM), are also considered in handling non-detects in this study. Some methods specifically designed for the imputation of scRNA data (as discussed in Chapter 2) will be applied to a simulated censored data (data with non-detects) and the results will be compared to some existing imputation techniques designed expressly for censored data. In subsequent analyses, variations of limit of detection and effect sizes (increasing from low to high) were considered to evaluate how well each non-detect method performed at different non-detect percentages using the Type I error rate and the power as metrics.

3.2 Methodology

3.2.1 Data

Synthetic data simulated from a real data were used for all analyses in this study. Real data were used as the basis for simulating data so as to get realistic distributional characteristics and mirror approximate real-world results. The real data used were from the study "Field-Level Exposure of Bumble Bees to Fungicides Applied to a Commercial Cherry Orchard" by Kuivila et al. (2021).

Their study evaluated bumble bee exposure to fungicides by quantifying concentrations of boscalid and pyraclostrobin in nectar and pollen collected by colonies of *Bombus huntii Greene*, 1860 (Hunt bumble bee) deployed in a commercial cherry (*Prunus avium L.*) orchard. The concentrations of boscalid in nectar varied by bee colony with significantly higher concentrations detected in colonies in the treated block (sprayed) than in the control block (unsprayed). In light of this, the data design for this simulation study is a simple nested design with treatment as the fixed factor and Group (bee colony) as the random factor nested within treatment. All non-detects in the Kuivila et al. (2021) study were accounted for using substitution with half Limit of Detection (LOD/2). Was the paper justified to use LOD/2 substitution for their study design?

3.2.2 Simulation Setup

In the Kuivila et al. (2021) study, there were about 59 nectar samples of $\leq 0.25ml$ from 13 bee colonies (7 Control colonies and 6 Treated colonies) which were analyzed over the 12 days of their experiment. Pyraclostrobin levels were detectable in only 2 of 13 nectar samples from the control block and in 8 of 46 samples from the treated block (roughly 83% non-detects). On the other hand, boscalid levels were detectable in 38 of 59 nectar samples with about 36% non-detects. For the purpose of this study, one type of fungicide concentration (i.e boscalid concentration) from the Cherry Orchard nectar data was used to simulate data based on distributional characteristics. The simulated data were generated using a Gamma distribution which best defines the distribution fit of the log boscalid concentrations in the observed nectar data since the concentrations were heavily right-skewed in the original study by Kuivila et al. (2021). The sample size for the simulated data were 100 nectar samples (boscalid concentrations in nectar) with 13 bee colonies (6 Treated colonies vs 7 Control colonies). The number of replicates within each group (bee colony) varied from 2 to 22 similar to what is shown in the original study. For instance, looking at the original data, the colonies from the control had a higher replication frequency than that of the treated colonies.

To simulate the non-detects in the data, the range of values (0.375, 0.5, 0.75, 1.125, 1.5, 2, 2.5) were considered as LODs (thresholds) to create varying censoring levels for analyses. The evaluation metrics, Type I error and power, were used to assess the strength of the non-detect methods after the methods were applied to the simulated data. At each censoring level, data simulation was run 1,000 times for reproducibility purposes. The Type I error and power results obtained were then averaged across the 1,000 simulation runs and the methods were compared. The null hypothesis under consideration was that the mean boscalid concentration from the treated colonies are not different from the control colonies. This null hypothesis was introduced into the simulated data by shuffling the data within the bee colonies to break any existing relationships between them. For the Type I error scenario, any significant result detected after hypothesis testing was deemed a false positive.

The power was assessed by adding in some magnitude of treatment effect (i.e., effect size) at increasing levels to the permuted data. The effect sizes added were 0.01, 0.1, 0.5, 0.75, 1 and 3. The original Kuivila et al. (2021) study used a two-sample linear rank test with Peto-Peto test to test for differences in exposure to boscalid between groups of bee colonies. The method they used is similar to a Wilcoxon rank sum test for two sample comparisons based on censored data but it did not account for the nesting feature which was present in their study design. A nested ANOVA was therefore used in this study, which accounts for nesting in the simulated data while testing for hypothesis in the Type I error and power scenarios. The data design for this simulation study was a simple nested design with treatment as the fixed factor and Group (bee colony) as the random factor nested within treatment. Given that simulated data were highly right-skewed and ANOVA assumes a normal distribution, transformation of data was necessary. Log transformation was therefore done after non-detect methods were applied to simulated data. The simulated data from the gamma distribution represent log-scale boscalid concentrations, and a log transformation of those data was still necessary to achieve approximate normality.

The results of the Type I error and power were quantified and visualized across the different non-detect methods. A lower average Type I error and a higher average power indicated a closer step to detecting the ground truth. To further justify what the best method was in this study, the magnitude of fold change (treatment effect estimate) was also computed and visualized across the different degrees of censoring. The fold change was computed as the differences in mean estimates. This was to help assess potential bias introduced by the various non-detect methods, their variability levels and also justify the power effect plots.

3.2.3 Methods

Non-detects can be considered a Missing Not at Random (MNAR) (see section 1.2) phenomenon so it is essential that one chooses the appropriate technique carefully in order to avoid any false interpretations. Even though there are a number of existing techniques, methods that are commonly used to handle non-detects in data were selected for this study. The methods used in this paper are summarized in this section and listed in Table 3.2.3.

Method **Abbreviation**	**Name**	**References**
Half_LOD	Substitution with half Limit of Detection (LOD/2)	(ITRC, 2013)
Sqrt_LOD	Substitution with LOD/$\sqrt{2}$	(ITRC, 2013)
Zero_sub	Zero substitution	(ITRC, 2013)
ROS	Regression on Order Statistics	(ITRC, 2013), (Helsel, 2005)
KM	Kaplan-Meier	(ITRC, 2013)
MLE	Maximum Likelihood Estimation	(ITRC, 2013), (Helsel, 2005)

Substitution Methods

Substitution is the most simplistic procedure when it comes to handling non-detects in data. Non-detects are usually replaced with values such as LOD, LOD/2, LOD/$\sqrt{2}$ or zero. This method is easy and requires little statistical knowledge. Several studies in past years have raised concerns about using substitution methods. For instance, Helsel (2010) argues that substitution can introduce a pattern that is alien to the pattern of the original data. Another study states that, simple substitution is OK only if few non-detects exist and only if the limit is so low relative to most measurements that it really does not make a statistical difference whether substitution is done with a zero, with half of the reporting limit, or with the reporting limit itself (Thomas, 2006).

The US Environmental Protection Agency (USEPA) has recommended substitution of one-half the detection limit when censoring percentages are below 15%. They added that, however, if simple substitution of values below the detection limit is proposed when more than 15% of the values are reported as not detected, nonparametric methods should be considered or a test of proportions should be used to analyze the data (USEPA, 1998). These recommendations by USEPA were not strongly backed by any published paper so it remains debatable. The direction of this paper considers how substitution methods compared to other methods in hypothesis testing at all levels of censoring not just at a low degree of censoring as stated by USEPA. LOD/2, LOD/$\sqrt{2}$ and 0 were used as substitution methods in this study.

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation is a fully parametric method which assumes that data are normally or lognormally distributed. It should be noted that MLE is not an imputation method for non-detects; instead, it estimates summary statistics (mean and standard deviation) for the full data accounting for the non-detects or censored values (ITRC, 2013). MLE solves a "likelihood equation" to find the mean and standard deviation values that are most likely to have produced both nondetect and detected data (Helsel, 2005). Application of MLE assumes that non-detects are distributed in a manner similar to the detected values. Hence MLE will perform poorly and produce misleading results if a well-fitted or closely matching distribution cannot be found to model the underlying population (ITRC, 2013). According to Helsel (2010), a few departures may be tolerated provided that the data distribution is not too far away from that assumed by the MLE.

It is therefore very crucial that the assumed distribution is accurate, as MLE may not be robust to any misspecification of data distribution. Again, MLE is most generally applicable to larger data sets ($n > 50$) with high detection frequencies (limited to up to 80% censoring) (ITRC, 2013). However, if the data follow a known distribution, MLE may work well for small data sets ($n < 50$) because it is using correct distributional information (Helsel, 2010). Using the NADA package in R with cenmle function, MLE was used to analyze the left-censored data in this simulated study. The distribution assumed when using the cenmle function was a lognormal which is very similar to the distribution of the simulated data that follows a gamma distribution. To perform hypothesis testing, the estimated mean from MLE is used to code the nested ANOVA manually in R.

Regression on Order Statistics (ROS)

Regression on order statistics (ROS) is a simple imputation method that fills in nondetect data on the basis of a probability plot of detects (Helsel, 2005). This method calculates a linear regression line in order to estimate values for the non-detects (ITRC, 2013). Using the ros function in NADA package in R, this analysis was performed. The output from the ros function included the imputed data which were used for the nested ANOVA in R using the aov function. During the data imputation with ROS, the method could not handle data with non-detect percentage greater than 80%, which is reflected in the Type I error and power plots in this study.

Kaplan-Meier (KM)

Kaplan-Meier (KM) is a non-parametric method so it does not require an assumption regarding the underlying distribution of the data. Just like MLE, Kaplan-Meier does not impute non-detects in data, instead, it estimates a cumulative probability distribution function to calculate summary statistics like means and variances. When applied as an intermediate step to calculate parametric statistics, Kaplan-Meier assumes that all data values come from a single underlying (non-negative) statistical population. In particular, contaminants are assumed to be present in non-detects at some low level not readily quantified by the analytical method. Kaplan-Meier can accommodate multiple reporting limits and is routinely used with data sets having 50 - 70% detection frequency. One weakness of Kaplan-Meier is that it cannot rank censored data points with reporting limits above the highest detected concentration (ITRC, 2013). In light of this, the thresholds were carefully selected not to exceed the highest detected boscalid concentration. The cenfit function under the NADA package in R, was helpful in using KM to handle non-detects in the simulated data.

3.3 Results and Discussion

In real-world data with control treatments, there is the possibility that the non-detect rate would become confounded with treatments as the effect size increased, with higher non-detect rates anticipated in the control group. Figure 3.1 shows that there is an inverse relationship between the average non-detect percentage and the effect sizes. As the effect size increased the average non-detect percentage dropped. However, as the limit of detection increased, the average non-detect percentage increased, which is indicative of a positive association.

In comparison of the measures of performance based on the Type I error and power plots shown in Figures 3.2 and 3.3, KM and MLE generally did not perform well as the other methods used in these simulations. They lost control of the Type I error and yielded a lower power compared to the substitution methods and ROS. More specifically, KM lost control of the Type I error after reaching 2.5 LOD which corresponds to an average non-detect of about 70%, looking at Figure 3.1 (because the Type 1 error rate is considered when the effect size is 0). The KM results confirm what has been discussed in literature (see 3.2.3).

Following various publications and literature on non-detect methods, the substitution methods (LOD/2, LOD/$\sqrt{2}$ and zero substitution) did not perform as poorly as expected, they out-

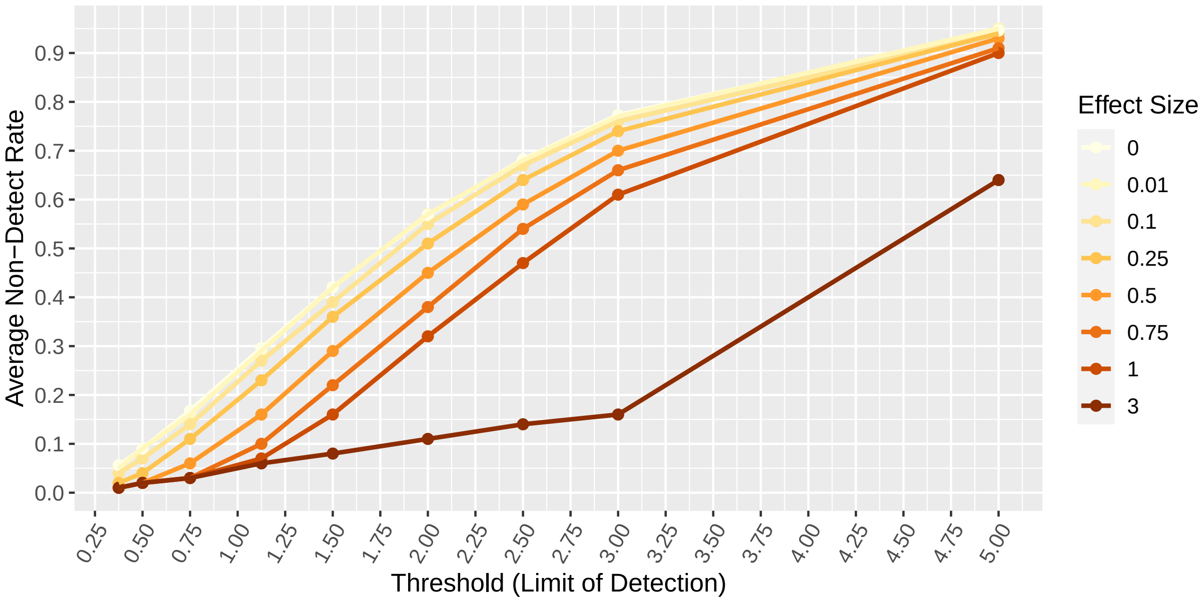


Fig. 3.1: Visualization of average non-detects at varying effect sizes across 1000 simulation runs per limit of detection

performed the "choice" techniques KM and MLE. The substitution methods and ROS had a good control of the Type I error. However, there were some interesting dynamics in the power plots. At lower effect sizes (0.01, 0.1, 0.25), the average power was very poor across the increasing levels of LODs. However, as the effect sizes got relatively higher, the average power significantly improved. At effect size 3, the average power plateaued around an average power of 1 for the substitution methods and ROS. At effect sizes 0.5, 0.75, and 1, the average power showed a drastic decline as the LODs increased.

These interesting patterns from the power warranted a further investigation. The "Bias" plots in Figures 3.4 and 3.5 were then generated to support the investigation. The bias plots were constructed from the estimated fold changes obtained from the different non-detect methods. Two thresholds were selected: 0.375 (lowest) and 2.5 (highest). For these two thresholds, the estimated fold changes for the effect sizes (0.01, 0.5, 0.75, 1) were visualized since they had some thought-provoking patterns occurring in the power plots. Zero substitution had the highest variability compared to all of the other methods (see Figures 3.4 and 3.5. Again, the zero substitution method overestimated the treatment effect at relatively high effect sizes 0.5, 0.75 and 1 (falls above the dashed line in Figure 3.4) and underestimated (falls below dashed line) the treatment effect at low threshold with relatively higher effect sizes, more prominently in Figures 3.4 C and D. This

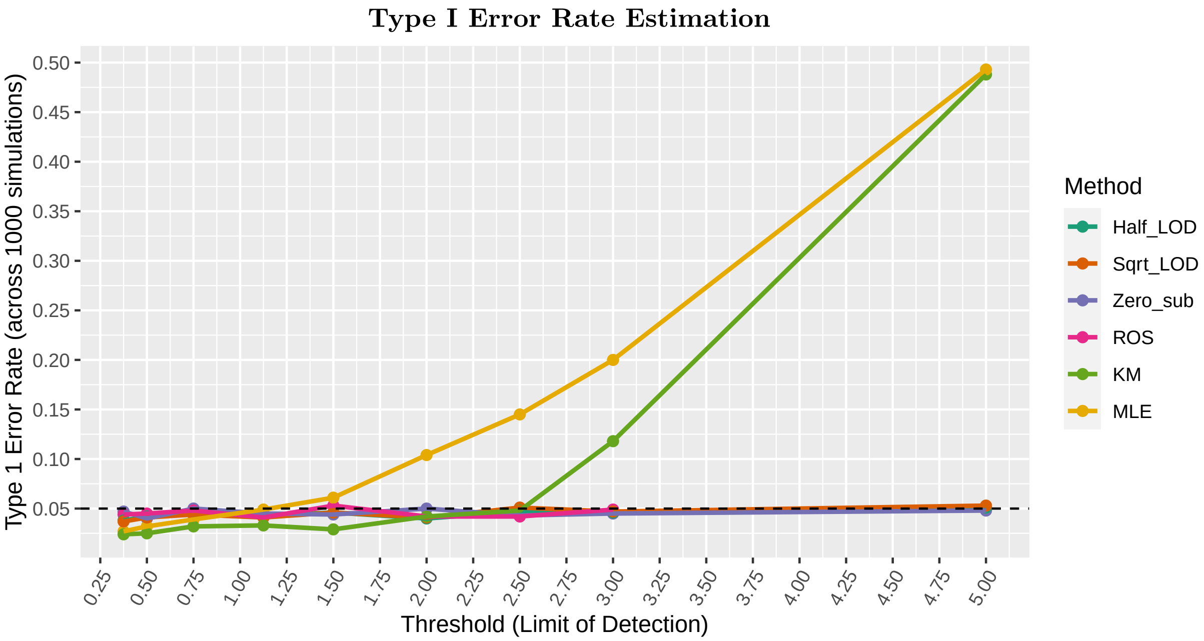


Fig. 3.2: Comparison of Type I Error Rate by method for handling non-detects and Limit of Detection level. Method name abbreviations are summarized in Table 3.2.3.

means that zero substitution in this study, tend to bias high at a high LOD and bias low at a lower LOD, for relatively higher effect sizes. A lot of outliers were detected with zero substitution for the low threshold bias plot in Figure 3.4.

Looking at Figure 3.4, at a low threshold of 0.375, the substitution methods (LOD/2, LOD/$\sqrt{2}$ and zero substitution) were more variable compared to ROS, KM and MLE. In general, ROS had the lowest variability but it underestimated the treatment effect as the effect size got higher for both low and high thresholds. Even though KM and MLE resulted in a lower variability with a better estimation of the treatment effect across all the bias plots, they will not be considered as ideal techniques for this simulation study since they could not control for the average Type I error. At a high threshold of 2.5, substitution by LOD/2 or LOD/$\sqrt{2}$ had a similar performance with very low variability and few outliers. These two techniques performed the best generally for the bias estimation, However, taking into account little departures that occurred in the bias estimation plots, substitution by LOD/2 performed better relatively.

Judging from the power, Type I error, and bias plots, substitution by LOD/2 emerged the best overall followed by substitution by LOD/$\sqrt{2}$ and then ROS. Even though zero substitution had a good control of the Type I error and relatively better power, it lost control of the bias which makes the results obtained from this method quite questionable and unreliable proven by literature. From

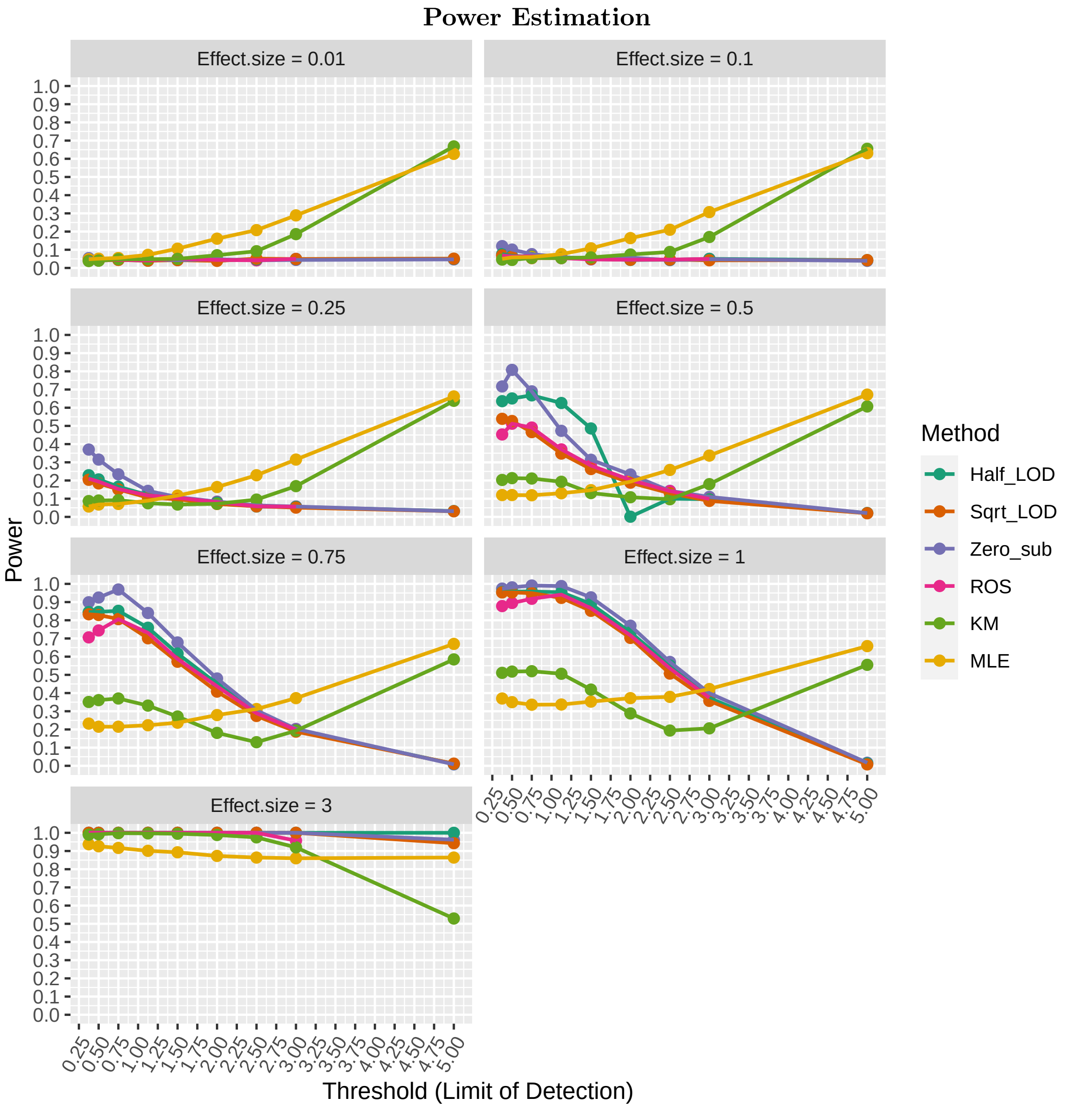


Fig. 3.3: Comparison of Power estimation by method for handling non-detects Limit of Detection level and Treatment Effect sizes. Method name abbreviations are summarized in Table 3.2.3.

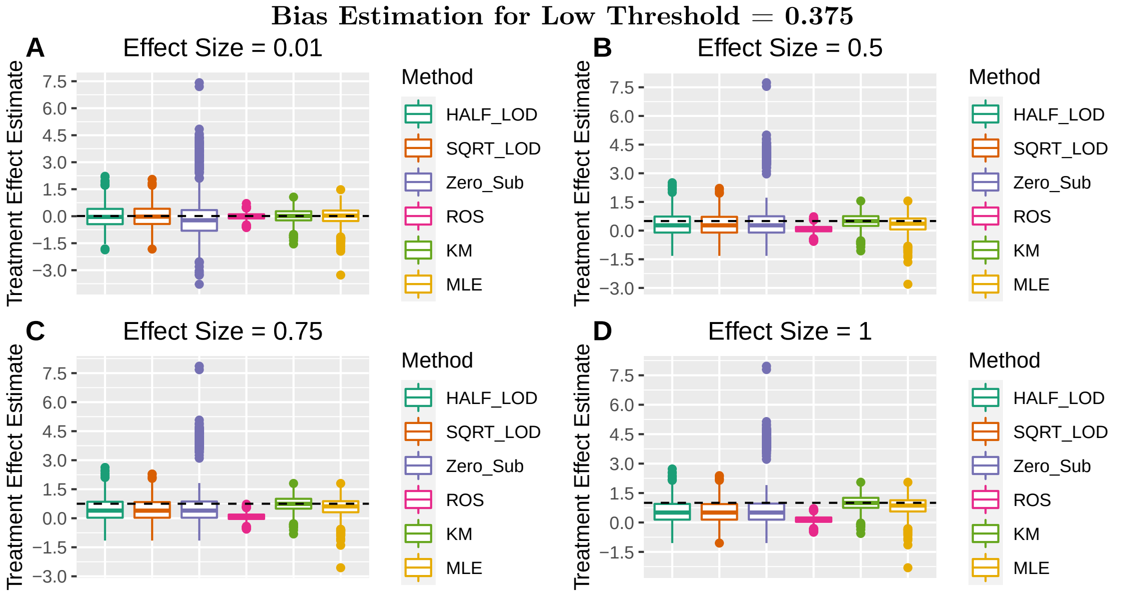


Fig. 3.4: Comparison of non-detect methods' variability across effect sizes for low threshold = 0.375. Treatment effect estimates from each of the methods considered are shown, from each of 1,000 simulations. The true treatment effect size is shown with a horizontal dashed reference line. Method name abbreviations are summarized in Table 3.2.3.

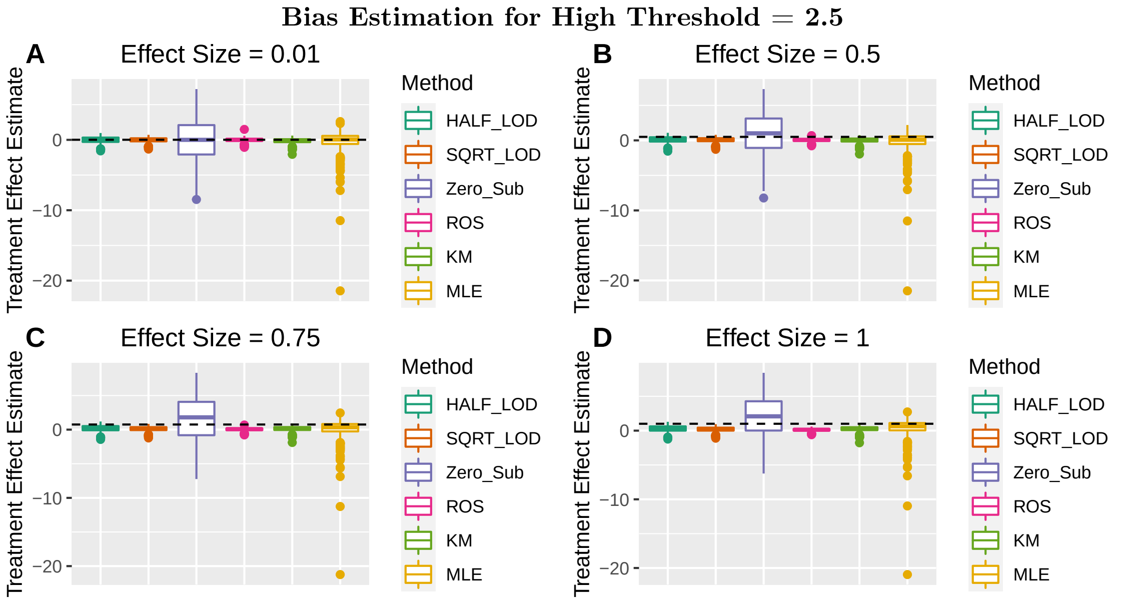


Fig. 3.5: Comparison of non-detect methods' variability across effect sizes for high threshold = 2.5. Treatment effect estimates from each of the methods considered are shown, from each of 1,000 simulations. The true treatment effect size is shown with a horizontal dashed reference line. Method name abbreviations are summarized in Table 3.2.3.

this study, the power significantly improved across all methods at an average non-detect rate below 70% (i.e at effect size 3). This could imply that all the methods considered for this simulation study cannot handle extreme non-detect percentages (greater than 70%). In conclusion, the results of this simulation study justified that substitution with LOD/2 was appropriate for this type of design.

3.4 Limitations and Future Work

Multiple imputation (MI) methods yield valid and robust parameter estimates and explicit imputed values for variables that can be analyzed as outcomes or predictors (Chen et al., 2011). It is therefore deemed as the "gold standard" technique by most researchers to handle non-detects. Multiple Imputation, however, failed to work with the design of the simulated data in this study because there were not enough predictors in the data to be used to build a model for the imputation as required. Even though, scRNA imputation techniques were being considered as part of the analysis in this study, there were not enough samples (groups) to be considered to apply these techniques to. Due to the above, the results obtained from this study are not generalizable to other designs but are limited to the specific nested design that was used in this study. Another key limitation in this study is that the nested design used did not account for bee colony effect.

A possible future direction is to explore more complex designs. No repeated measures structure was accounted for in this simulation study, primarily for the sake for simplicity because in the original study there were many colonies with single measurement times reported and several measurement times were sparsely repeated. Rather than attempting to model this kind of complex repeated measures structure, this initial analysis and simulation focused on the simple nested design. Subsequent study can be done to account for repeated measures structure. In addition, the simple nested design used in this study can be improved to contain more predictors so that Multiple Imputation can be implemented and assessed with the other non-detect methods mentioned in this study.

An extension of this study, could comprise of simulating each dataframe to contain multiple censoring levels instead of having a single censoring per dataframe as done in this study. This can help evaluate how well substitution, imputation or non-imputation methods for non-detects are able to handle multiple censoring in data.

The simulation study in this paper can be further explored by introducing bee colony effect into the nested design since the nested design used in this study did not account for bee colony effect. Preliminary work is currently being done on this subject to extend this study for publication.

Lastly, the Type I error and Power scenarios for this simulation study can further be assessed by simulating data at varying sample sizes (e.g., 50, 100, 200, 500, etc.). This would be potential to assess how well these non-detect methods perform on data of any magnitude (large or small).

CHAPTER 4

A RESAMPLING STRATEGY TO EFFICIENTLY HANDLE IMBALANCED DATA

Abstract

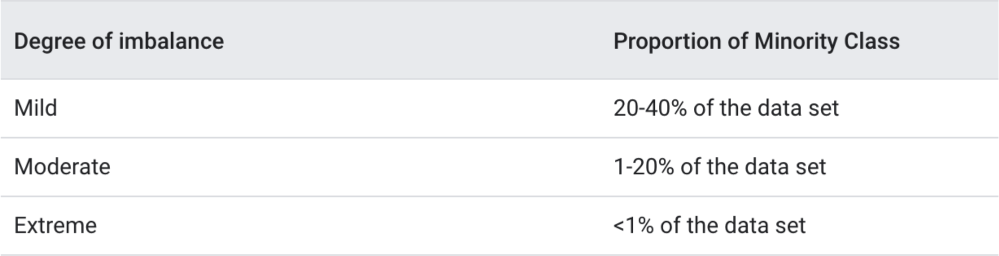
Imbalanced data is an enigma that is inherent in a lot of real-world datasets such as medical diagnosis, fraud detection, etc. Particularly in the field of machine learning and data mining, classification of imbalanced data is a challenging task. This is because a significant number of standard classifiers assume a balanced class distribution which results in prediction bias towards the majority class. The ultimate challenge associated with imbalanced classification is to increase the sensitivity of a classifier towards the minority class. Diverse solutions have been proposed over the years to address this challenge. Prominent amongst them is the use of random undersampling to balance class distribution. As simple as this technique may seem, random undersampling presents two inherent challenges: 1. How much to sample from the majority class? and 2. How to compensate for information lost through sampling from the majority class? These fundamental questions are addressed in this paper by devising a strategy aimed at resolving these issues whilst improving prediction performance towards the minority class. The proposed strategy combined random undersampling with different weighting schemes using standard machine learning classifiers like random forest and logistic regression to predict the minority class at different degree of imbalance. This paper also focuses on choosing an optimal percentage to sample from the majority class based on the degree of class imbalance (class distribution) by evaluating different probability values. The different weighting schemes used were 1) Inverse of Number of Samples(INS) (Singh, 2020), 2) Inverse of Square Root of Number of Samples (ISNS) (Singh, 2020), 3) Inverse of Downsampling factor (IDS), and 4) Upweighting of Downsampled Class (UDS) (GoogleDevelopers, 2021). A case of no weighting was considered as a baseline strategy. The ultimate strategy was selected based on a simulation study, evaluating popular performance metrics like F1-Score, AUCPR, AUC, and precision. The effectiveness of the proposed strategy was then evaluated through application on real datasets. The strategy proposed to deal with a moderate imbalanced classification was random undersampling with a downsampling probability at 0.43 using random forest classifier and a UDS weighting scheme. In a mild imbalanced classification, multiple weighting methods (INS, ISNS, UDS) as well as no weighting approach were found to be effective at an optimal probability of about 0.45 in the proposed strategy. The recommended strategy for extreme balance classification did not yield high performance metrics in its application to a real dataset (Credit card) as compared to the mild and moderate imbalance scenarios and as such was inconclusive.

4.1 Introduction

One major consequence of missing data is imbalanced data. In the world of machine learning and data mining, imbalanced data is one of the fundamental problems associated with classification tasks. Technically speaking, an imbalanced dataset is any dataset that is characterized by uneven class distribution. The imbalanced data phenomenon is apparent in several real-life domains such as fraud detection, disease diagnosis, natural disaster, etc. This has therefore attracted a significant amount of attention from academia, industry, and government (Hasanin and Khoshgoftaar, 2018).

Typically, in imbalanced data, there are a large amount of observations or data for one class (referred to as the majority or negative class) and much fewer observations or limited data representation for one or the other classes (referred to as the minority, rare, or positive classes). The most common assumption, if the scope of data is not clearly stated in literature, is that imbalanced data is innately a binary (two-class) classification (Phung, 2020). This is not always true as imbalanced data may occur in every classification problem, including a multi-class (more than 2 classes) classification.

The degree or ratio of imbalance present in data is mainly gauged by the class ratio. For example, considering a binary classification problem, a ratio of 10:1 would mean that for every 1 example in one class, there are 10 examples in the other class. Orriols-Puig et al. (2009) defines the imbalance ratio (IR) as the ratio of the number of instances of the majority class to the number of instances of the minority class that are sampled to the system. According to GoogleDevelopers (2021), knowing the proportion of the minority can reveal the degree of imbalance in a data. They summarized the degree of imbalance into three levels: Mild, Moderate, and Extreme which is shown in Table 4.1 below. It should be noted that the minority class is not always rare in its own right (i.e., absolute rarity) but could just be of a lower proportion relative to the majority class (He and Garcia, 2009). Having absolute rarity in the proportion of the minority class means that regardless of how much data is increased, the minority class samples cannot increase. This leads to severe imbalance (rare instances imbalance). On the other hand, if the minority class remains outnumbered regardless of data increase, then a relative imbalance arises (Phung, 2020).



Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	<1% of the data set

Fig. 4.1: Imbalanced Data Levels

Another important aspect to understand when dealing with imbalanced classification is the domain of the problem, whether intrinsic or extrinsic. An intrinsic imbalance occurs when differences in classes is due to the nature of the dataspace, e.g. fraud detection, disease diagnosis, etc., whereas extrinsic imbalance is not directly related to the nature of the dataspace but is based on variable factors such as time, data collection, and storage (He and Garcia, 2009). Understanding the nature as well as the degree of imbalance is therefore very crucial, as they have significant consequences and can pose challenges to imbalanced data classification.

A dominant majority of traditional machine learning algorithms were designed based on the assumption of a balanced class distribution or equal misclassification costs (He and Garcia, 2009). So when there is class imbalance, the machine learning classifier tends to be more biased towards the majority class, causing bad classification of the minority class. This can lead to a deceptively high accuracy metric. In certain cases such as disease diagnosis studies, where the occurrence of false negatives is relatively costlier than false positives, a learner's prediction bias in favor of the majority class could have adverse consequences. For instance, if the minority class indicates a disease is present, then having a false negative, which means misclassifying patients as not having the disease when the disease is present, can be a serious error (Leevy et al., 2018). Brownlee (2020a) emphasized that a slight imbalance is often not a concern, and the problem can often be treated like a normal classification predictive modeling problem. However, a severe imbalance of the classes can be challenging to model and may require the use of specialized techniques. Some practitioners seem to disagree with the fact that the performance degradation in the machine learning algorithms to classify an imbalanced data problem should be mainly attributed to the class ratio (Batista et al., 2004).

The ultimate challenge yet to be resolved is being able to accurately classify the minority class to a large extent, especially in a severe imbalance case. The most obvious solution is to collect more data until the classes are balanced out. This approach can be expensive in terms of cost and time and may not always be feasible as the minority class may be rare (Hasanin and Khoshgoftaar, 2018). Considering the gravity of the imbalanced data problem, a number of solutions have been proposed over the years to deal with this issue. These proposals can be divided into three categories: the algorithm level, the data level, and the hybrid level. Data level approaches consist of resampling the training data in order to decrease the effect caused by the imbalance. The most popular resampling techniques used are random undersampling (Prusa et al., 2015), random oversampling with replacement (Kotsiantis et al., 2006), and generating synthetic samples using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). The algorithm solutions deal with either modifying existing classification algorithms such as decision trees (Truică and Leordeanu, 2017), random forests (del Río et al., 2014a), gradient boosting (Blagus and Lusa, 2017), etc., in order to adapt them to the imbalanced data or proposing specific strategies (Mahani and Ali, 2019). Hybrid level approaches like cost-sensitive learning combines the two previous options trying to minimize the misclassification costs, which are higher for the instances of the minority class (Hasanin and Khoshgoftaar, 2018).

The aim of this paper is to present an efficient strategy to handle any degree of imbalance based on the combination of random undersampling with different weighting schemes to minimize prediction bias towards the majority class using standard machine learning classifiers like random forest and logistic regression. The proposed strategy was implemented in three settings, considering a case of mild, moderate, and extreme imbalance. This study was first of all developed on a simulated data and the classifiers results evaluated based on aggregate performance metrics like F1-score, AUC, Precision-Recall AUC (AUCPR), and precision.

The best strategy was determined based on evaluating the different weighting strategies combined with the undersampled data to see which yields the best prediction across the aggregate performance metrics under extreme imbalance and relative imbalance. In addition, a case of no weighting was also considered as a baseline against the different weighting schemes. This will help assess whether simple random undersampling without weighting was enough to control prediction bias and further investigate whether class distribution does not affect a classifier's performance as asserted by some studies. The performance of the selected classifiers, random forest and logistic regression, was also checked in the process. The effectiveness of the winning strategy was evaluated through its application on real data with mild, moderate, and extreme imbalance. To make this study more intuitive, a binary classification problem was considered.

Another important area addressed in this paper is the optimal percentage to sample from the majority class in order to reduce the predictive modelling bias, given a relative imbalanced (i.e, mild and moderate) or extreme imbalanced data. Several attempts have been made by some research publications to address this issue directly or indirectly. Nonetheless, how much to sample from the majority class for random undersampling in order to improve prediction of the minority class remains a mystery and a difficult task. The work in Hasanin and Khoshgoftaar (2018) indirectly attempted this by investigating good ratios for undersampling big data without discarding too much of the majority class. Their work, however, was not conclusive about a particular ratio, instead they concluded that if the number of minority class labels was too low, increasing the minority class percentages from 0.1% to 1.0% can give a fair boost in the performance with regards to random forest, yet still retain a reasonable amount of information in relation to the original dataset. This paper focuses on choosing an optimal downsampling percentage based on the degree of class imbalance (class distribution) and also attempts to minimize information loss through weighting.

4.2 Methodology

This section briefly discusses resampling techniques, the different weighting schemes, performance metrics, datasets, classification methods, simulation set-up, and evaluation strategy employed for the various analyses.

4.2.1 Resampling Techniques

Resampling is a widely adopted technique that is used to address the issue of imbalance in data. The goal of sampling methods is to create a dataset that has a relatively balanced class distribution, so that traditional classifiers are better able to capture the decision boundary between the majority and the minority classes (Hoens and Chawla, 2013). Random undersampling (Prusa et al., 2015) involves randomly removing samples from the majority class until the majority and minority classes are balanced out. This does not necessarily mean achieving an exact 50:50 class distribution, but getting to a distribution that the classifier can handle. One major advantage of random undersampling is that it is simple to implement and less time consuming. Again, there are no artificially-created data points added to data. Hence, there is no chance of falsifying the data samples. This technique, however, has the risk of potentially losing important information for analysis due to the random removal of samples from the majority class (Phung, 2020).

To overcome this limitation of random undersampling, more sophisicated undersampling techniques have been developed. Popular of these include Condensed Nearest Neighbor Rule (CNN) (Thai-Nghe et al., 2010), Near Miss Undersampling (Mqadi et al., 2021) and Tomek Links method (Brownlee, 2020e).

Random oversampling, on the other hand, involves creating multiple duplicated data points by boostrap sampling of the minority class until the classes are balanced out. This can increase the training data size as well as computational time. Duplicating the minority class might make the model more prone to over-fitting.(Phung, 2020). Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) which combines oversampling and undersampling, oversamples the minority class by creating "synthetic" examples rather than by oversampling with replacement. SMOTE was not used in this study because of the falsification of data points for the minority class that will not be a true representative of the actual population and can result in biased predictions. In addition, oversampling was avoided due to the possible increase in training time.

Random undersampling was used as the baseline technique in the proposed strategy for this study because it does not involve heavy alteration of the data through data falsification. Again, a study by Weiss et al. (2007) revealed that undersampling does not only address the imbalanced problem but also makes processing more feasible when data are too big to handle.

4.2.2 Weighting and Downsampling

The strategy adopted in this study to combat the possible loss of useful information is to weight the majority and minority classes after random undersampling to preserve the feature of the original population and ensure the classification model used is still calibrated. This can potentially help to reduce bias in the predictive modelling process. Most machine learning algorithms do not take into account the skewed distribution of the classes and as such tend to produce biased results in favor of the majority class. The training algorithm can, however, be modified through weighting of the classes to influence the classification of the classes during the training phase. It is usually expected that the minority class is assigned a higher class weight than the majority class in order to penalize the misclassification made by the minority class (Kamaldeep, 2020).

In this paper, four different weighting schemes were evaluated which deal with upweighting and downweighting the minority class appropriately to capture the distribution of data and also account for possible loss of information (i.e., reduce loss of information in the majority class).

Two fundamental questions considered in order to implement the proposed strategy (i.e, random undersampling with weighting) succesfully were:

- What weighting technique will be the most representative of the true population in order to reduce prediction bias?
- What will be an optimal proportion to downsample the majority class by?

Four simple weighting schemes were considered to account for the contribution of the classes to the overall loss. The weighting schemes explored were Inverse of Number of Samples (INS) (Singh, 2020), Inverse of Square Root of Number of Samples (ISNS) (Singh, 2020), Inverse of Downsampling factor (IDS), and Upweighting of Downsampled Class (UDS) (GoogleDevelopers, 2021). ISNS, IDS, and ISNS weighting schemes assign a higher weight to the minority class as expected. On the other hand, UDS assigns a higher weight to the majority with the goal of minimizing information loss in that class. The weighting schemes used have been grouped based on the parameters they use: class sample size and downsampling probability.

Weighting Schemes based on sample size

- **Inverse of Number of Samples (INS)** weights the samples as the inverse of the class frequency for the class they belong to.

$$W_{n,c} = \frac{1}{\text{Number of samples in class c}}$$

- **Inverse of Square Root of Number of Samples (ISNS)** weights the samples as the inverse of the Square Root of the class frequency for the class they belong to.

$$W_{n,c} = \frac{1}{\sqrt{\text{Number of samples in class c}}}$$

Weighting Schemes based on downsampling probability

- **Upweighting of Downsampled Class (UDS)** was a technique proposed by (GoogleDevelopers, 2021) as an effective way of handling imbalanced data. This strategy suggests weighting the downsampled class (majority class) using the same the factor by which it was downsampled by. For instance, if the majority class was downsampled by a factor of 20, the same value of 20 should be used to weight the majority class. In this paper, a modified version of the upweighting proposed was used by inverting the downsampling probability and a weight of 1 assigned to the minority class since the entire sample for the minority class was used to train the model.

$$W_{\text{majority}} = \frac{1}{\text{downsampling probability}}$$

$$W_{\text{minority}} = 1$$

- **Inverse of Downsampling factor (IDS)** weights the majority class with the inverse of the downsampling probability multiplied by 100%. This approach downweights the majority class. The minority class is assigned a weight of 1 just like the UDS weighting scheme. This is a new scheme to assess whether downweighting the majority class by the downsampling probability is powerful enough to penalize the misclassfication towards the minority in the training algorithm.

$$W_{\text{majority}} = \frac{1}{\text{downsampling probability} * 100\%}$$

$$W_{\text{minority}} = 1$$

4.2.3 Classification Models

Random forest (Breiman, 2001) and logistic regression were used as classifiers for this study. All classifier parameter values were used in their default state without tuning. For instance, the default probability threshold of 0.5 was used for logistic regression whilst the default ntree value of 500 was used in random forest classification.

- **Random Forest Classifier**

  Generally speaking, ensemble methods have been researched to demonstrate good behavior when confronted with imbalanced data (Galar et al., 2011), and it is believed that using one of them as a basis for comparison should not bias the results regarding the minority class (del Río et al., 2014b). Random forest which is a well-known decision tree ensemble method, has proven to be no exception to this claim in several studies. An extensive study by (Fernández-Delgado et al., 2014) which compared the application of about 179 classifiers arising from 17 families to real-world classification problems concluded that random forest is most likely to be the best classifier. It is also believed that combining random sub-sampling with random forest may overcome the imbalance problem (Hasanin and Khoshgoftaar, 2018). This makes random forest a classifier worth considering for this study.

- **Logistic Regression Classifier**

  Logistic regression is a well established classification algorithm which remains a reference benchmark in many domains like consumer credit risk, due to the regulatory requirement of interpretability (Li et al., 2022). In spite of this, several studies have found logistic regression not to support imbalanced data directly but instead requires modification to its training algorithm in order to take into account the imbalanced distribution (Brownlee, 2020b). One way this study seeks to address this issue is to evaluate the performance of a logistic regression model on a randomly undersampled dataset with different class weights schemes. This study considers whether this will make it comparable in performance to using a random forest classifier.

4.2.4 Datasets

The data used in this study for analysis were both synthetic and real-world data. To simplify the analysis of study results, all datasets contain only two classes (i.e., binary). Features of the simulated and real data have been summarized in Tables **??** and **??** below.

- **Simulated Data**

  The twoClassSim function from the caret package in R (Kuhn, 2008) was mainly used to simulate data at varying class imbalances (i.e., mild, moderate, and extreme imbalance). A total of 2000 observations with a 75:25 data partition were simulated under each case of class imbalance except for extreme imbalance which had 25,000 observations with an 80:20 data partition (i.e., 8000 observations in training set and 2000 observations in test set). Additionally, important variables were added by setting the linearVars argument in the twoClassSim function to 10 for mild and moderate imbalance and for extreme imbalance set linearVars to 14. The dimension of simulated data were 2000 observations by 16 variables for the mild and moderate imbalance and 25,000 observations with 20 variables for the extreme imbalance. The target variable was made of up two classes: Class 1 (majority) and Class 2 (minority) with Class 2 being the class of interest in this study.

  By default, the twoClassSim function generates a balanced class distribution of about 50:50. The intercept argument in the twoClassSim function was therefore used to control the overall level of class imbalance. To simulate the different degrees of imbalance, the intercept was set to -7 for mild imbalance which yielded a minority class proportion of about 36%. Moderate imbalance was introduced with an intercept of -15, yielding a minority class proportion of about 10% and then extreme imbalance was simulated with an intercept of -25 resulting in approximately 1.7% minority class proportion.

- **Real Data**

  After conducting a simulation study, the effectiveness of the recommended strategy in real-life with the problem of class imbalance was assessed. Three real-world datasets were selected, each representing the degree of class imbalance that was simulated in the study. In choosing the real data, the distributional characteristics were considered so that they were comparable to the simulated data in this study. The descriptions of all the real data are as follows:

  - The first real dataset is a credit card fraud detection dataset from Kaggle (Yadav, 2020). This dataset contains about 25,000 observations and 18 variables. The target variable is defined as 0 when a transaction is fraud and 1 when a transaction is not fraud. The class distribution is about 422 observations (1.7%) for fraud transactions (minority class) and about 24712 observations (98.3%) for no fraud (majority class). This shows that the credit card dataset is heavily imbalanced with a class ratio of about 60:1.

  - The imbalanced binary thyroid gland data is the next real dataset that was considered in this study. This dataset is available under the Imbalance package in R. The thyroid data contains 215 observations with 6 variables. The target variable is made up of two classes: positive as hyperthyroidism, negative as non hyperthyroidism. There are about 35 observations in the minority (positive) class corresponding to about 16%. Hence the thyroid data can classified as being moderately imbalanced.

  - The third real dataset is a breast cancer database obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg (Wolberg and Mangasarian, 1990). This dataset is available in the UCI Machine Learning repository. This dataset is also accessible through the imbalance package in R (Cordn et al., 2018). The original Wisconsin breast cancer data in the UCI respository contains 10 variables and about 699 observations. The target variable has two possible classes -- benign (2) and malignant (4). Out of the 699 observations, there are 16 missing values. The class distribution for the target variable was relatively imbalanced with benign having 458 observations (65.5%) and malignant with 241 observations (34.5%). The Wisconsin data used for this study is accessed through the Imbalance package in R which omits the missing values. Hence, the data used had 683 observations with 10 variables with 444 negative observations and 239 positive observations. The class ratio for the Wisconsin data shows a mild imbalance closer to a 9:5.

The class distributions for the real data may not be exact as the simulated data but are comparable as they fall within the same type of class imbalance.

Table 4.2.4 summarizes all simulated and real datasets that were used in this study.

	Simulated Data			Real Data			
	Total Sample Size	**Num of Variables**	**Pct Minority Class**	**Total Sample Size**	**Num of Variables**	**Pct Minority Class**	**Name**
Mild	2000	16	0.363	683	10	0.35	Wisconsin
Moderate	2000	16	0.103	215	6	0.16	Thyroid
Extreme	25,000	20	0.016	25,134	20	0.017	Credit Card

4.2.5 Simulation Setup

Random undersampling was performed on the training data using the ROSE (Random Over-Sampling Examples) R package (Lunardon et al., 2014) which uses a bootstrap-based technique which aids the task of binary classification in the presence of rare classes. The ovun.sample function was used to implement the undersampling by specifying the method as "under". The probability to downsample from the majority class was specified as a sequence of twenty probabilities ranging from 0.45 to 1 for the mild imbalance training data, 0.15 to 1 for the moderate imbalance, and 0.05 to 1 for the extreme imbalance data.

The minimum downsampling probability specified was dependent on the degree of imbalance and sample size in the minority class. The minimum probability specified could not be unreasonably smaller than the actual proportion of minority class examples in the original population. Hence the downsampling probability, p, in the ovun.sample function was tuned until a reasonable downsampling probability was obtained for each degree of imbalance. Each class imbalance was evaluated separately across the four different weighting schemes using logistic regression and random forest classifiers to predict the minority class.

4.2.6 Evaluation Strategy

Choosing the right evaluation metric is also another challenge in imbalanced data classification as most standard evaluation metrics assume a balanced class distribution. According to Brownlee (2021), a classifier is only as good as the metric used to evaluate it. He emphasizes that one is likely to choose a poor model or be misled by the expected performance of a model, if the wrong metric is chosen to evaluate a model. Accuracy is a widely acknowledged evaluation metric to assess a classifier's performance. However, in the presence of imbalanced data, accuracy can be misleading.

There are a ton of other evaluation metrics for classification and there are a number of varied views on which metrics are ideal for imbalanced data. In lieu of accuracy, F1-score, precision, AUC, and precision-Recall AUC (AUCPR) were considered as alternative metrics to assess model performance in this study. These performance metrics were selected because they are popular metrics for imbalanced classification. In addition to this, these metrics put some focus on the minority class, which is the class of interest in this study.

Precision summarizes the fraction of examples assigned the positive class that belong to the positive class. A high precision value is said to be indicative of a low false positive rate.

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.1}$$

F1-score, also known as F-measure, is a popular metric for imbalanced classification. This is because unlike the Accuracy metric which works best if false positives and false negatives have similar cost, F1-score is useful when the cost of false positives and false negatives are very different. The F1-score is defined as the weighted average of precision and Recall. Recall summarizes how well the positive class was predicted and is the same calculation as sensitivity (Joshi, 2016).

$$\text{F1-score} = \frac{(2 * precision * Recall)}{(precision + Recall)} \tag{4.2}$$

Another commonly used measure of classification performance is the Area under the Receiver Operating Characteristics (ROC) curve **(AUC)**. AUC is a single score calculated from the ROC plot to assess a model's predictive ability. The ROC curve is a diagnostic plot for summarizing the behavior of a model by calculating the false positive rate (x-axis) and true positive rate (y-axis) for a set of predictions by the model under different thresholds. A no skill classifier will have an AUC score of 0.5, whereas a perfect classifier will have a score of 1.0 (Brownlee, 2021). Although AUC is a preferred metric over Accuracy in an imbalanced classification, it has its own drawbacks. Brownlee (2020d) pointed out that for imbalanced classification with a severe skew and few examples of the minority class, the AUC score can be misleading. This is because a small number of correct or incorrect predictions can result in a large change in the AUC score. Regardless of this, the AUC score was analysed across the different class imbalance to assess its impact on model performance in this study.

Area under the Precision-Recall (PR) Curve (AUCPR) is considered an alternative to AUC in imbalanced classification. The PR curve is very similar to the ROC curve but replaces the False Positive Rate by precision. The focus of the PR curve on the minority class makes it an effective diagnostic for imbalanced binary classification models. The AUCPR score is therefore recommended for highly skewed domains to have a better view of model performance (Brownlee, 2020d). Unlike AUC which uses a baseline of 0.5 to evaluate a classifier's performance, with AUCPR the baseline is the fraction of positives (relative to the total number of examples) (Brownlee, 2020d). This means that the baseline for AUCPR could be lower than 0.5 depending on the class distribution.

A higher value for all selected performance metrics is preferred in this study. The following steps were taken to evaluate the results from the random forest and logistic regression models in the simulation study.

- Perform random undersampling on simulated training data (mild, moderate, or extreme imbalance) using sequence of twenty downsampling probabilities on the majority (negative) class.
- Fit random forest and logistic regression model on undersampled simulated train data using all default parameters.
- Apply different weighting schemes to classifiers in model fitting. A case of no weighting is considered to fit model for prediction.
- Predict the class of interest (minority class) using fitted model on simulated test data.
- Run simulations 1000 times to ensure results are reproducible, and average performance metrics across 1000 simulation runs.

The averaged results obtained for the performance metrics were generated across the sequence of twenty downsampling probabilities. The classifier with the best prediction results (i.e., higher performance metrics on average) was then applied to real data using the corresponding downsampling probability and weighting scheme. The above steps were then repeated at a single simulation run for the real dataset. All simulations were run through the high performance computing platform at the University of Utah. The run time was much faster for logistic regression compared to random forest. The maximum run time for logistic regression was about six hours whilst the maximum run time for random forest was approximately fourteen hours.

4.3 Results and Discussion

Simulated data results

Results for logistic regression were not reported in this paper as random forest significantly outperformed it. Logistic regression generally yielded lower evaluation metrics across the different levels of class imbalance.

The overall performance of random forest was however noticeably high across all degrees of imbalance. The results in Figure 4.2 clearly show that IDS weighting is not appropriate for relative imbalanced data using the random forest classifier. The IDS scheme with random forest consistently yielded average F1-scores right around 0.5 and precision values below 0.5 across all downsampling probabilities as shown in Figures4.2A and 4.2D. This could be indicative of poor prediction of the minority class. The average AUCPR and AUC scores were also significantly lower compared to using the other weighting schemes and even a no weighting approach.The trade-off between F1-score and precision is again noticed between the downsampling probabilities 0.45 to about 0.5 (see 4.2A and 4.2D). The performance of not weighting the classes in the random forest model was impressively similar to the performance of the other weighting schemes (INS and ISNS). This can be seen at a downsampling probability of 0.45 in Figure4.2 which appears to be the optimal for the relative imbalance scenario. Considering this, multiple weighting methods (i.e., INS, ISNS, No weighting) could be considered when dealing with mild imbalance in order to achieve better classification.

From Figures 4.3A and 4.3D, IDS weighting with random forest yielded very poor F1 and precision values across downsampling probabilities, 0.375 to 1. IDS weighting showed generally high performance across all evaluation metrics at a downsampling probability of 0.15. The performance of random forest with weighting schemes INS, ISNS and noweighting (Nowgt) looks quite comparable in Figures 4.3A, 4.3B, and 4.3C. UDS weighting with random forest has the highest AUCPR and precision across a significant range of the downsampling probabilities (see Figures 4.3C and 4.3D). Comparing Figures 4.3A to 4.3D, there appear to be a trade-off between the Average precision and F1-score for the UDS weighting scheme. Figure 4.3 shows that optimality in random forest performance can be achieved at a downsampling probability of about 0.43 using the UDS weighting scheme. This yields an average F1-score and precision close to a 0.7 and an average AUCPR of 0.9 which is significantly greater than the baseline of 0.1 minority class rate.

With extreme imbalance, IDS weighting scheme with a downsampling probability of 0.05 was found to be the most ideal technique combination to achieve optimal classification results using random forest. From Figures 4.4A and 4.4D, the average F1-scores and precision using IDS weighting scheme is significantly high (close to 0.75) at a 0.05 downsampling probability. However, there is a drastic decline in F1-scores and precision beyond the 0.05 probability mark which is shown by the dashed reference line. AUC and AUCPR values are significantly high (close to 1) using IDS and a 0.05 downsampling probability. There is a trade-off between F1-score and precision at 0.05 probability when using UDS weighting scheme (see Figures 4.4A and 4.4D). Again, there is a similar prediction performance amongst no weighting, INS, and ISNS weighting schemes as shown in 4.4.

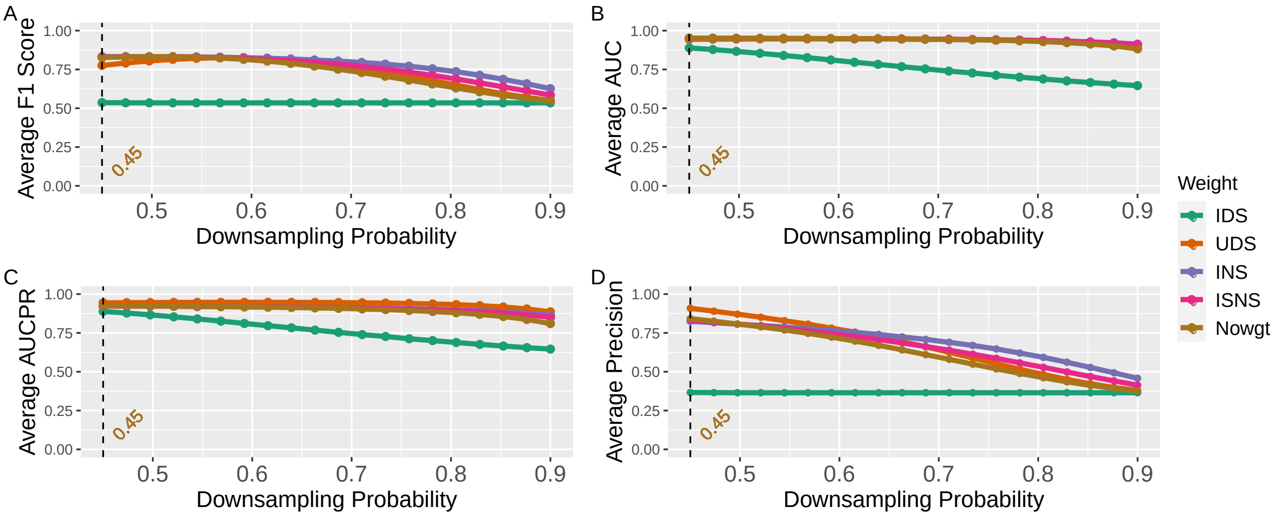


Fig. 4.2: Average Performance metrics across different weighting schemes using random forest classifier on mildly imbalance simulated data

Concluding from Figures 4.2, 4.3, and 4.4, UDS weighting with a downsampling probability of 0.43 yielded more optimal results in the moderate imbalance situation. On the other hand, for extreme imbalance, IDS weighting with 0.05 downsampling probability was selected. The performance metric values obtained when IDS weighting scheme was applied to mild and moderate imbalance were much lower. Hence it can be inferred that IDS is not an ideal weighting strategy to use when dealing with relative imbalance in data. Multiple weighting techniques (such as no weighting, INS, ISNS and even UDS) with downsampling probability of 0.45 can be considered when dealing with mild imbalance in order to achieve optimal results.

Real data results

Table 4.1 summarizes the results of the real data after the application of selected weighting techniques with random forest at specific downsampling probability in the simulation study. Each

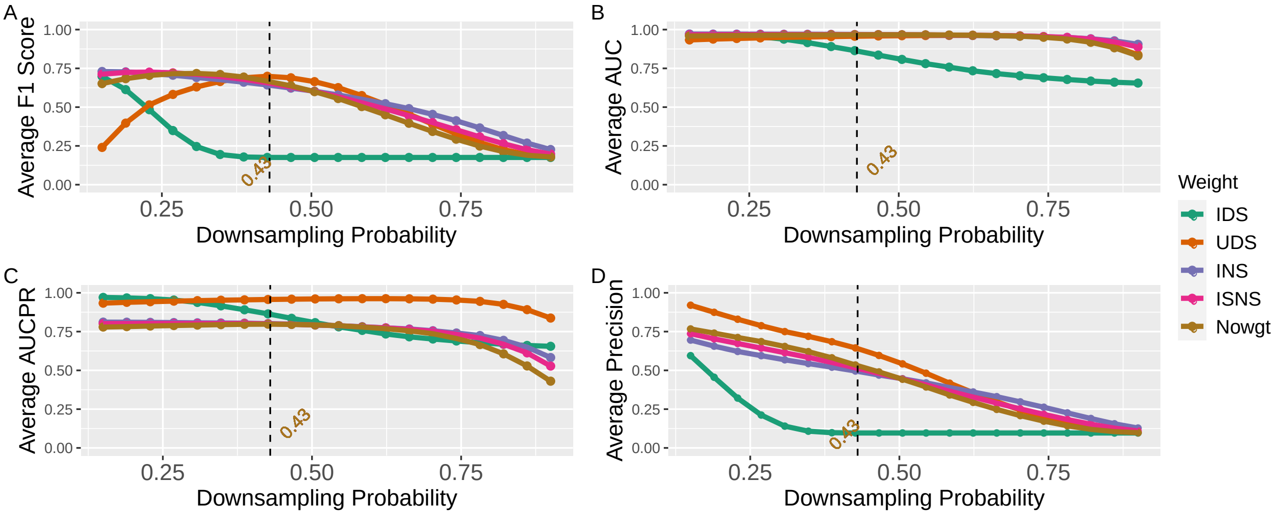


Fig. 4.3: Average Performance metrics across different weighting schemes using random forest classifier on moderately imbalanced simulated data

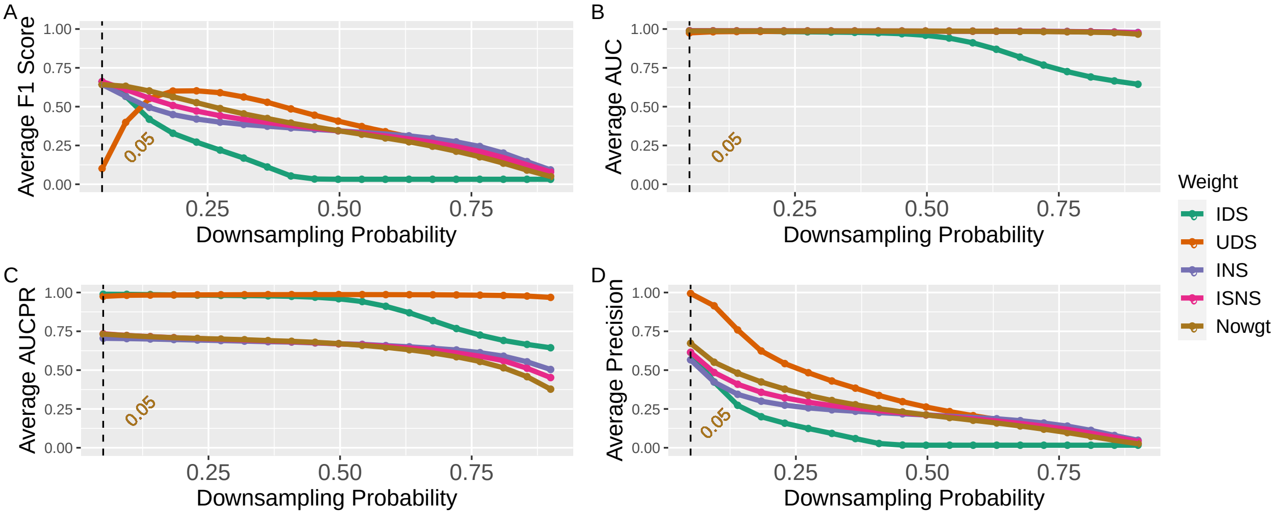


Fig. 4.4: Average Performance metrics across different weighting schemes using random forest classifier on extremely imbalanced simulated data

real dataset is supposed to have similar distributional characteristics as the simulated data in order for the recommended strategy to be effective and applicable.

The strategy recommended for extreme imbalance (i.e., downsampling with 0.05 and weighting with the inverse of Downsampling factor (IDS) strategy in a random forest model) exhibited not so high performance metrics like the mild and moderate imbalance cases when applied to the credit card data (see last row of Table 4.1). All evaluation metrics except for AUC had significantly low results. This confirms literature that AUC is a misleading metric to rely on in extreme imbalance classification. Given the heavy imbalance in data, the high AUC could mean that the classifier detected more true negatives than false negatives, hence making it unreliable. We can however assess the effectiveness of the extreme imbalance strategy used based on the AUCPR. Comparing the AUCPR value (0.191) for credit card data to the baseline (minority rate) of 0.017, the AUCPR is quite high which is an improvement (see Table 4.1).

On the other hand, the results for Wisconsin breast cancer dataset classification in Table 4.1 confirmed that random undersampling with a downsampling probability of 0.45 and INS weighting scheme in a random forest model is an effective strategy to predict relative imbalance data when faced with a similar class distribution of about 3:2. The precision and F1-score were significantly high (greater than 90%). Similar results were obtained when ISNS or no weighting was used. This means that using INS, ISNS, or no weighting can help to correctly classify the minority class (malignant cases) to a large extent. The proposed strategy for moderate imbalance data (i.e UDS with a 0.43 downsampling probability in a random forest model) was highly effective when applied to the real Thyroid data. A perfect precision score (i.e., 1) was achieved. This means that none of the minority class samples were incorrectly which of main interest in this study. The other performance; F1-score, AUC, and AUCPR were also significantly high (see second row in Table 4.1).

Table 4.1: Summary of Real Data Classification Results

Data	**Degree of Imbalance**	**Weighting Scheme**	**Downsampling Probability**	**Precision**	**F1-Score**	**AUC**	**AUCPR**
Wisconsin	Mild	INS	0.45	0.952	0.975	0.994	0.988
Thyroid	Moderate	UDS	0.43	1.0	0.857	1.0	1.0
Credit Card	Extreme	IDS	0.05	0.345	0.291	0.743	0.191

4.4 Limitations and Future Work

The performance of the proposed strategy to address extreme imbalance classification in this study should be further investigated by tuning the hyperparameters in the random forest model to see how well it can improve extreme imbalance classification.

An extension of this study should explore an optimal probability threshold that needs to be set for extreme or relative imbalance that can help improve the performance of logistic regression across the different weighting techniques and downsampling probability. This is because the default probability (0.5) used may not accurately reflect the class imbalance even though class weights were assigned in the logistic regression model.

Lastly, a comparative analysis can be done to compare popular heuristic techniques of random undersampling like Condensed Nearest Neighbor Rule (CNN) (Thai-Nghe et al., 2010), Near Miss Undersampling (Mqadi et al., 2021) and Tomek Links method (Brownlee, 2020e) to the basic random undersampling by incorporating the different weighting schemes mentioned in this study (i.e., IDS, UDS, ISNS, INS) into an ensemble classifier like random forest or gradient boosting at different degrees of class imbalance.

CHAPTER 5
CONCLUSION

Missing data remains a major challenge that researchers need to deal with in almost every field. The impact of missing data, if not carefully handled, can be detrimental to any statistical analysis. Overlooking missing data can result in loss of information, bias the estimation of study parameters, reduce statistical power, and affect generalizability of findings. It is established in this dissertation that the choice of analysis depends on the type of data used in a study. It is therefore crucial that researchers pay close and particular attention when dealing with missing data.

This dissertation generally addresses statistical challenges presented by different missing data examples in different research domains. Possible solutions to effectively handle some of these challenges were also discussed in this through simulation studies and application to real datasets. The challenges posed by missing data were recognized to be associated mainly with the type of missingness as well as the degree of data missingness. The ultimate challenge that seems to run through this dissertation was to know the appropriate technique to select to deal with the missing data phenomenon. The different missing data scenarios considered are the dropout issue in single-cell RNA sequencing analysis, non-detects (otherwise referred to as below detection limit), and imbalanced data, which is an effect of missing data. These missing data examples are organized as stand-alone papers which are captured in Chapters 2, 3, and 4, respectively.

The first paper in Chapter 2 was a comparative study analysing popular imputation and differential test methods in single-cell RNA (scRNA) sequencing. Dropout zeros, which are designated as the missing values in single-cell RNA (scRNA) data, are usually difficult to distinguish from the true zeros (zeros arising from biological factors). This leads to biased downstream analysis results in scRNA sequencing. In an attempt to find a technique that will be best for scRNA data analysis, a simulation study was conducted evaluating the average Type I error rate and power from four popular scRNA sequencing imputation methods (MAGIC, SAVER, DrImpute, and scImpute) combined with three differential expression (DE) test methods (DESingle, MAST, and Seurat). Additionally, the ongoing controversy of whether imputation is a necessary step in scRNA sequencing analysis was addressed by considering a case of not imputing the scRNA data.

Based on the results of the study, there was no clear winner. Instead, the best method was selected based on varied reasons. First, MAGIC paired with DESingle or MAST was found to be appropriate if a researcher was primarily interested in knowing whether genes are differentially expressed or not (i.e, obtaining a fairly higher power and lower false positive). Though MAGIC had a poor bias estimation of the magnitude difference, it did the best job at detecting the ground truth simulated by yielding the lowest Type I error rate and highest power consistently across all DE methods. On the other hand, if a researcher is primarily interested in less biased estimates, DrImpute would be an ideal imputation method to consider. This is because DrImpute was the next best imputation method to yield a fairly higher power with DESingle and MAST. Imputation was found to be a crucial step in scRNA analysis as the estimated power results obtained from not imputing scRNA data were much lower than that of using imputation methods in the simulation study. An extensive analysis can be conducted as a follow-up to this study, incorporating dimensionality reduction techniques as well as more advanced imputation and differential testing methods. One important finding in this study was that the various imputation methods failed to detect smaller magnitude differences when combined with the differential methods. An additional study will therefore be required to investigate this interesting relationship.

The second paper of this dissertation (Chapter 3) investigates how to handle non-detects through simulation studies. More specifically, the paper considers whether it will be appropriate to address the issue of non-detects using a traditional substitution approach, imputation, or a non-imputation based approach in a simple nested design. Seven existing non-detect techniques namely, Zero substitution, Substitution with half Limit of Detection (LOD/2), Substitution with LOD/$\sqrt{2}$, Multiple Imputation (MI), Regression on Order Statistics (ROS) (Imputation approach), Maximum Likelihood Estimation (MLE), and Kaplan-Meier (KM) were applied to simulated data at varied censoring levels (thresholds) and effect sizes. The performance of each non-detect method was then evaluated based on the average Type I error rate and the power.

This study strongly demonstrated that substitution with LOD/2 was an appropriate technique for a nested design but this raises further fundamental questions about the structure of the nested design used and also the complexity of the design. The study design was a simple nested design consisting of 100 observations with Treatment (Treated or Control) as a fixed factor and bee colony (Group) as a Random factor nested within Treatment. The simulated data used in this study assumed a no colony effect. To comprehend the impacts of the findings brought out in this study, a follow-up study can be conducted, introducing colony effect into the nested design.

Again, the simple nature of this study design was not applicable to a Multiple Imputation (MI) as it expects a multivariate distribution. Because of this, more complex designs, like repeated measures design, could be investigated to assess how these non-detect methods compare to MI at varying censoring levels and effect sizes. The sample size for the simulated data could be varied as an extended version of this study. This will help researchers to understand how well these non-detect methods perform at any magnitude of data (large or small). Furthermore, multiple censoring levels could be analysed. It must be noted that the findings from this study are only applicable when working with designs similar to what was used in this study. Zero substitution, which is a common technique to handle non-detects, was found to produce highly biased estimates, and so will not be recommended. A technique which effectively accounts for zero-inflation (e.g., Zero-Inflated Poisson regression) may be considered if a researcher wants to use zero substitution to deal with non-detects. Overall, the power significantly improved across all the non-detect methods at an average non-detect rate below 70%. This may imply that at extreme non-detect levels (beyond 70%) these non-detect method may not produce the best results.

Some research works that use resampling techniques in imbalanced learning focus on using specific algorithms, paying little attention to the optimality of how much to sample and the need to compensate for the cost of sampling. These are two key areas that are addressed in the third paper (Chapter 4) in this dissertation. A major problem that persists in imbalanced classification is the prediction bias of a classifier towards the majority class. This is because most standard classifiers were designed to handle balanced class distribution. An efficient strategy was therefore presented to mitigate the prediction bias problem by combining random undersampling with weighting of standard classifiers like random forest and logistic regression. Four different weighting schemes were used: Inverse of Number of Samples(INS), Inverse of Square Root of Number of Samples (ISNS), (Inverse of Downsampling factor (IDS), and Upweighting of Downsampled Class (UDS). A case of no weighting was also considered.

By comparative analysis of simulated data (considering relative and extreme imbalance), it was shown that using random forest as a classifier outperformed using logistic regression based on the aggregate results of the evaluation metrics used in this study (i.e., F1-score, precision, AUC, and Precision-Recall AUC (AUCPR)). The proposed strategy was unique to the degree of imbalance (whether extreme or relative). When dealing with relative imbalance, the recommended strategy was to randomly undersample data with a downsampling probability of about 0.45 using the Inverse of

Number of Samples(INS) weighting scheme and random forest classifier. The results obtained when this strategy was applied to a real dataset of a similar distributional characteristic were impressively high. The precision, AUCPR, and F1-score were significantly high, indicating an almost perfect classification of the minority class (above 0.9). The strategy that best fits extreme imbalanced classification was, however, inconclusive in this study. The strategy that was recommended (using a downsampling probability of 0.43 and UDS weighting scheme) based on the extreme imbalance simulation yielded poor results when it was applied to a real dataset with a more severe imbalance. The obvious question to here is whether the severity of imbalance in the simulated dataset representative of the real dataset. Another relevant question is whether the sample size could be a contributing factor to performance results. These questions would be worth exploring in in a follow-up study so that an appropriate strategy could be recommended to researchers to possibly address extreme imbalance classification.

The hyperparameters for the classifiers in this study were used in their default setting. A future consideration will be to analyse how tuning a classifier's hyperparameters can improve prediction of minority class in an imbalanced classification. Over the years, variants of undersampling have been developed. Prominent of these include Condensed Nearest Neighbor Rule (CNN) (Thai-Nghe et al., 2010), Near Miss Undersampling (Mqadi et al., 2021), and Tomek Links method (Brownlee, 2020e). A comparative analysis could be done comparing these techniques to the random undersampling strategy proposed in this study. Lastly, research could be conducted to investigate whether data partitioning matters in imbalanced classification. Taken together, the results from this study emphasize that random undersampling combined with weighting of classes (upweighting minority class) in an imbalanced classification is a useful tool to increase sensitivity of a classifier towards the minority class.

REFERENCES

Allison, P., 11 2012a. Why You Probably Need More Imputations Than You Think. Statistical Horizons(https://statisticalhorizons.com/more-imputations#:~:text=So%20if%2027%25%20of%20the,about%2030%20imputed%20data%20sets).

Allison, P., 09 2014. Sensitivity Analysis for Not Missing at Random. Statistical Horizons(https://statisticalhorizons.com/sensitivity-analysis).

Allison, P. D., 2012b. Handling Missing Data by Maximum Likelihood. In: SAS global forum. Vol. 2012. pp. 1038–21, (http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf).

Andrews, T. S., Hemberg, M., 11 2018a. False Signals Induced by Single-Cell Imputation. F1000Research 7, 1740–1740, (https://pubmed.ncbi.nlm.nih.gov/30906525), PMID: 30906525.

Andrews, T. S., Hemberg, M., 2018b. Identifying cell populations with scrnaseq. Molecular Aspects of Medicine 59, 114–122, the emerging field of single-cell analysis.
URL https://www.sciencedirect.com/science/article/pii/S0098299717300493

Barescut, J., Lariviere, D., Stocki, T., Wood, M., Beresford, N., Copplestone, D., 2011. Limit of Detection Values in Data Analysis: Do They Matter? Radioprotection 46 (6), S85–S90, (https://pdfs.semanticscholar.org/8617/7969fc512d7d340ec8788c2c102b522e3c27.pdf).

Batista, G. E., Prati, R. C., Monard, M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 6 (1), 20–29.

Bhadauria, R., 2019. ML | Expectation-Maximization Algorithm. (https://www.geeksforgeeks.org/ml-expectation-maximization-algorithm/).

Blagus, R., Lusa, L., 2017. Gradient boosting for high-dimensional prediction of rare events. Computational Statistics & Data Analysis 113, 19–37.
URL https://www.sciencedirect.com/science/article/pii/S0167947316301803

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Brownlee, J., 2020a. A Gentle Introduction to Imbalanced Classification. https://machinelearningmastery.com/what-is-imbalanced-classification/, updated on : January 14, 2020.

Brownlee, J., 2020b. Cost-sensitive logistic regression for imbalanced classification. https://machinelearningmastery.com/cost-sensitive-logistic-regression/.

Brownlee, J., 2020c. A Gentle introduction to Expectation-Maximization (EM Algorithm). (https://machinelearningmastery.com/expectation-maximization-em-algorithm).

Brownlee, J., 2020d. Roc curves and precision-recall curves for imbalanced classification. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/.

Brownlee, J., 2020e. Undersampling Algorithms for Imbalanced classification. https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/, posted on : January 20, 2020.

Brownlee, J., 2021. Tour of evaluation metrics for imbalanced classification. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., Stegle, O., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology 33 (2), 155–160.
URL https://doi.org/10.1038/nbt.3102

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chen, G., Ning, B., Shi, T., 04 2019. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in genetics 10, 317–317, (https://pubmed.ncbi.nlm.nih.gov/31024627), PMID: 31024627.

Chen, H., Quandt, S. A., Grzywacz, J. G., Arcury, T. A., 2011. A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values Below the Limit of Detection. Environmental Health Perspectives 119 (3), 351–356, (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059998/).

Company, G. . A. B. L. S., 2020. Single-Cell RNA Sequencing Frequently Asked Questions. (`https://web.genewiz.com/single-cell-faq`).

Cordn, I., Garca, S., Fernndez, A., Herrera, F., 2018. imbalance: Preprocessing algorithms for imbalanced datasets. R package verion 1 (2).

Ctech.com, 1994 - 2022. Handling non-detects. `https://www.ctech.com/studio_help/Content/file_format_details/handling_non_detects.htm`.

D. Hess, J., 2020. Missing Data: the Hidden Problem. (`https://www.bauer.uh.edu/jhess/documents/2.pdf`).

del Río, S., López, V., Benítez, J. M., Herrera, F., 2014a. On the use of mapreduce for imbalanced big data using random forest. Information Sciences 285, 112–137, processing and Mining Complex Data Streams.
URL `https://www.sciencedirect.com/science/article/pii/S0020025514003272`

del Río, S., López, V., Benítez, J. M., Herrera, F., 2014b. On the use of mapreduce for imbalanced big data using random forest. Information Sciences 285, 112–137, processing and Mining Complex Data Streams.
URL `https://www.sciencedirect.com/science/article/pii/S0020025514003272`

Drou, N., Gresham, D., Gunsalus, K., Singh Katari, M., Khalfan, M., Daisy Rowe, J., Twaddle, A., Yousef, A., 2022. Seurat part 3 – Data normalization and PCA. NYU Center For Genomics and Systems Biology, (`https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/seurat-part-3-data-normalization/`).

Enders, C. K., 2001. A Primer on Maximum Likelihood Algorithms Available for Use with Missing Data. Structural Equation Modeling 8 (1), 128–141, (`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.6288&rep=rep1&type=pdf`).

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., Theis, F. J., 2019. Single-cell rna-seq denoising using a deep count autoencoder. Nature communications 10 (1), 1–14.

Farnham, I. M., Singh, A. K., Stetzenbach, K. J., Johannesson, K. H., 2002. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemometrics and Intelligent Laboratory Systems 60 (1-2), 265–281.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research 15 (1), 3133–3181.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., Gottardo, R., 2015. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome Biology 16 (1), 278.
URL https://doi.org/10.1186/s13059-015-0844-5

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (4), 463–484.

Ganna, A., Lee, D., Ingelsson, E., Pawitan, Y., 09 2014. Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies. Briefings in bioinformatics 16.

Ganser, G. H., Hewett, P., 2010. An accurate substitution method for analyzing censored data. Journal of occupational and environmental hygiene 7 (4), 233–244.

Gebregziabher, M., 2019. Analysis of Zero-Inflated Count and Semi-Continuos Data. Workshop: DPHS Summer Institute 2019.

Gelman, A., Hill, J., 2006. Data Analysis using Regression and Multilevel/Hierarchical models. Cambridge university press, (http://www.stat.columbia.edu/~gelman/arm/missing.pdf).

Glen, S., 09 2015. EM Algorithm (Expectation-Maximization): Simple Definition. (https://www.statisticshowto.com/em-algorithm-expectation-maximization/).

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., Garry, D. J., 2018. DrImpute: Imputing Dropout Events in single cell RNA Sequencing Data. BMC Bioinformatics 19 (1), 220, (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2226-y).

GoogleDevelopers, 2021. Imbalanced Data. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.

Grace-Martin, K., 2020a. Em Imputation and Missing Data: Is Mean Imputation Really so Terrible? (https://www.theanalysisfactor.com/em-imputation-and-missing-data-is-mean-imputation-really-so-terrible/).

Grace-Martin, K., 2020b. Seven Ways to Make up Data: Common Methods to Imputing Missing Data. (https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/).

Grace-Martin, K., 2020c. Two Recommended Solutions for Missing Data: Multiple Imputation and Maximum Likelihood. (https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/).

Hafemeister, C., Satija, R., 2019. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome biology 20 (1), 1–15.

Harel, O., Perkins, N., Schisterman, E. F., 2014. The Use of Multiple Imputation for Data Subject to Limits of Detection. Sri Lankan journal of applied statistics 5 (4), 227–246, (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4838401/), PMID: 27110215.

Hasanin, T., Khoshgoftaar, T., 2018. The effects of random undersampling with simulated class imbalance for big data. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, pp. 70–79.

He, H., Garcia, E. A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21 (9), 1263–1284.

He, Z., Pan, Y., Shao, F., Wang, H., 2021. Identifying Differentially Expressed Genes of Zero Inflated Single Cell RNA Sequencing Data Using Mixed Model Score Tests. Frontiers in genetics 12, 616686–616686, (https://pubmed.ncbi.nlm.nih.gov/33613638).

Helsel, D., 2010. Much ado about next to nothing: incorporating nondetects in science. Annals of occupational hygiene 54 (3), 257–262.

Helsel, D. R., 2005. More than obvious: better methods for interpreting nondetect data. Environmental science & technology 39 (20), 419A–423A.

Helsel, D. R., 2006. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65 (11), 2434–2439, environmental Chemistry.
URL https://www.sciencedirect.com/science/article/pii/S0045653506005157

Hoens, T. R., Chawla, N. V., 2013. Imbalanced datasets: from sampling to classifiers. Imbalanced learning: Foundations, algorithms, and applications, 43–59.

Hu, J., Shang, X., et al., 2020. SCC: An Accurate Modification Method for scRNA-Seq Dropouts Based on Mixture Model, current status: Under review.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J., Raj, A., Li, M., Zhang, N. R., 2018a. SAVER: Gene Expression Recovery for UMI-Based Single Cell RNA Sequencing. bioRxiv(https://www.biorxiv.org/content/early/2018/03/08/138677).

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., Zhang, N. R., 07 2018b. SAVER: Gene Expression Recovery for Single-Cell RNA Sequencing. Nature methods 15 (7), 539–542, (https://pubmed.ncbi.nlm.nih.gov/29941873), PMID: 29941873.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., Zhang, N. R., 2018c. Saver: gene expression recovery for single-cell rna sequencing. Nature methods 15 (7), 539–542.

ITRC, 2013. Groundwater Statistics and Monitoring Compliance, Statistical Tools For The Project Life Cycle. (https://www.itrcweb.org/gsmc-1/Content/Resources/GSMCPDF.pdf).

Ivanecky, S., 08 2020. What the Heck Does "Zero" Mean? (https://towardsdatascience.com/what-the-heck-does-zero-mean-8c5f42266dc6).

Jaakkola, M. K., Seyednasrollah, F., Mehmood, A., Elo, L. L., 07 2016. Comparison of methods to detect differentially expressed genes between single-cell populations. Briefings in Bioinformatics 18 (5), 735–743.
URL https://doi.org/10.1093/bib/bbw057

Jiang, R., Sun, T., Song, D., Li, J. J., 2022. Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome Biology 23 (1), 1–24.

Joshi, R., 2016. Accuracy, precision, recall f1 score: Interpretation of performance measures. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

Kamaldeep, S., 2020. How to improve class imbalance using class weights in machine learning. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/.

Kang, H., 2013. The Prevention and Handling of the Missing Data. Korean journal of anesthesiology 64, 402–406, ("https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/"), PMID: 23741561.

Kharchenko, P. V., Silberstein, L., Scadden, D. T., 2014. Bayesian approach to single-cell differential expression analysis. Nature methods 11 (7), 740–742.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al., 2006. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30 (1), 25–36.

Kuhn, M., 2008. Building predictive models in r using the caret package. Journal of Statistical Software, Articles 28 (5), 1–26.
URL https://www.jstatsoft.org/v028/i05

Kuivila, K. M., Judd, H., Hladik, M. L., Strange, J. P., 04 2021. Field-Level Exposure of Bumble Bees to Fungicides Applied to a Commercial Cherry Orchard. Journal of Economic Entomology 114 (3), 1065–1071.
URL https://doi.org/10.1093/jee/toab051

Kumar, N., 2 2019. Advantages and Disadvantages of KNN Algorithm in Machine Learning. (http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html).

Kwak, I.-Y., Gong, W., Koyano-Nakagawa, N., Garry, D. J., 2017. DrImpute: Imputing Dropout Events in Single Cell RNA Sequencing Data. bioRxiv(https://www.biorxiv.org/content/early/2017/08/28/181479).

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., Seliya, N., 2018. A survey on addressing high-class imbalance in big data. Journal of Big Data 5 (1), 42.
URL https://doi.org/10.1186/s40537-018-0151-6

Li, G., Yang, Y., Van Buren, E., Li, Y., 2019. Dropout Imputation and Batch Effect Correction for single-cell RNA Sequencing Data. Journal of Bio-X Research 2 (4), (https://journals.lww.com/jbioxresearch/Fulltext/2019/12000/Dropout_imputation_and_batch_effect_correction_for.4.aspx).

Li, W. V., Li, J. J., 2018. An Accurate and Robust Imputation Method scImpute for Single-Cell RNA-Seq Data. Nature communications 9 (1), 1–9, (http://jsb.ucla.edu/sites/default/files/scImpute.pdf).

Li, Y., Adams, N., Bellotti, T., 2022. A relabeling approach to handling the class imbalance problem for logistic regression. Journal of Computational and Graphical Statistics 31 (1), 241–253.

Lodder, P., 2013. To Impute or Not Impute: That's the question. Advising on research methods: Selected topics, 1–7(https://www.paultwin.com/wp-content/uploads/Lodder_1140873_Paper_Imputation.pdf).

Lun, A. T., McCarthy, D. J., Marioni, J. C., 2016. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Research 5.

Lunardon, N., Menardi, G., Torelli, N., 06 2014. Rose: a package for binary imbalanced learning. R Journal 6, 79–89.

Mahani, A., Ali, A. R. B., 2019. Classification problem in imbalanced datasets. In: Sadollah, A., Sinha, T. S. (Eds.), Recent Trends in Computational Intelligence. IntechOpen, Rijeka, Ch. 4.
URL https://doi.org/10.5772/intechopen.89603

Malarvizhi, M. R., Thanamani, A. S., 2012. K-Nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development 5 (1), 5–7, (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.925&rep=rep1&type=pdf).

McDavid, A., 2022. Interoptability between mast and singlecellexperiment-derived packages.

Miao, Z., Deng, K., Wang, X., Zhang, X., 04 2018. Desingle for detecting three types of differential expression in single-cell rna-seq data. Bioinformatics 34 (18), 3223–3224.
URL https://doi.org/10.1093/bioinformatics/bty332

Miao, Z., Li, J., Zhang, X., 2019. screcover: Discriminating true and false zeros in single-cell rna-seq data for imputation. bioRxiv, 665323.

Morton, M., Lion, K., 2016. Estimation with Values Below the Limit of Quantitation. (https://www.coresta.org/sites/default/files/abstracts/2016_ST15_Morton.pdf).

Moss, S., 06 2016. Expectation Maximization–To Manage Missing Data. sicotests(https://www.sicotests.com/psyarticle.asp?id=267).

Mqadi, N., Naicker, N., Adeliyi, T., 07 2021. Solving misclassification of the credit card imbalance problem using near miss. Mathematical Problems in Engineering 2021.

Newsom, J., 2020. Missing Data and Missing Data Estimation in SEM (Psy 523/623 Structural Equation Modeling, spring 2020). (http://web.pdx.edu/~newsomj/semclass/ho_missing.pdf).

Obadiah, Y., 2017. The Use of KNN for Missing Values. (https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c63).

Orriols-Puig, A., Bernado-Mansilla, E., Goldberg, D. E., Sastry, K., Lanzi, P. L., 2009. Facetwise analysis of xcs for problems with class imbalances. IEEE Transactions on Evolutionary Computation 13 (5), 1093–1119.

Phung, T. M., 2020. Imbalanced learning: sampling techniques. https://tungmphung.com/imbalanced-learning-sampling-techniques/, posted on : May 24, 2020.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., Napolitano, A., 2015. Using random undersampling to alleviate class imbalance on tweet sentiment data. In: 2015 IEEE International Conference on Information Reuse and Integration. pp. 197–202.

Qiu, P., 2020. Embracing the dropouts in single-cell rna-seq analysis. Nature communications 11 (1), 1–9.

Raghunathan, T. E., 2004a. What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. Annual Review of Public Health 25 (1), 99–117, ("https://doi.org/10.1146/annurev.publhealth.25.102802.124410"), PMID: 15015914.

Raghunathan, T. E., 2004b. What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. Annual Review of Public Health 25 (1), 99–117, ("https://doi.org/10.1146/annurev.publhealth.25.102802.124410"), PMID: 15015914.

Roy, B., 2019. All About Missing Data Handling. (https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d218).

Salgado, C. M., Azevedo, C., Proença, H., Vieira, S. M., 2016. Missing Data. Springer International Publishing, Ch. 13, pp. 143–162, (https://doi.org/10.1007/978-3-319-43742-2_13).

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 33, 495–502.
URL https://doi.org/10.1038/nbt.3192

Schechter, C., 05 2016. When Should Missing Data, in Numerical Variables, be Replaced by Zeros? (https://www.statalist.org/forums/forum/general-stata-discussion/general/1341703-when-should-missing-data-in-numerical-variables-be-replaced-by-zeros).

ScienceDirect, 2020. Limit of Detection. (https://www.sciencedirect.com/topics/nursing-and-health-professions/limit-of-detection).

Shi, Y., 2007. Gene Expression Microarray Missing Value Imputation and Its Effects in Downstream Data Analysis. Master's thesis, University of Alberta, (http://yishi.sjtu.edu.cn/Publications/yi07.pdf).

Shoari, N., Dubé, J.-S., 2018. Toward improved analysis of concentration data: embracing nondetects. Environmental toxicology and chemistry 37 (3), 643–656.

Shoop, S., 2015. Should We Ban the Use of "Last Observation Carried Forward" analysis in epidemiological studies. SM J Public Health Epidemiol 1 (1), 1004, (fulltext_smjphe-v1-1004.pdf).

Singh, K., 2020. Handling class imbalance by introducing sample weighting in the loss function. https://medium.com/gumgum-tech/handling-class-imbalance-by-introducing-sample-weighting-in-the-loss-function-3bdebd8203b4.

Soley-Bori, M., 5 2013. Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis. Tech. Rep. 4, Boston University School of Public Health, Department of Health Policy & Management, (https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf).

Solution, S., 2020. Handling Missing Data: Listwise versus Pairwise Deletion. (https://www.statisticssolutions.com/handling-missing-data-listwise-versus-pairwise-deletion).

Solution, S., 2020. Missing data: Listwise vs. pairwise. (https://www.statisticssolutions.com/missing-data-listwise-vs-pairwise/).

Statistical Consulting Group, U., 2020. Zero-inflated poisson regression | r data analysis examples. (https://stats.idre.ucla.edu/r/dae/zip/#:~:text=Zero%2Dinflated%20poisson%20regression%20is,zeros%20can%20be%20modeled%20independently).

Statistics, R., 2020. Fiml basic concepts. (https://www.real-statistics.com/handling-missing-data/full-information-maximum-likelihood-fiml/fiml-basic-concepts/).

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., Carpenter, J. R., 2009. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. Bmj 338, (https://www.bmj.com/content/338/bmj.b2393).

Sun, S., Zhu, J., Ma, Y., Zhou, X., 2019. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. Genome biology 20 (1), 1–21.

Support, I., 4 2020. Pairwise vs. Listwise Deletion: What are they and when should i use them? (https://www.ibm.com/support/pages/pairwise-vs-listwise-deletion-what-are-they-and-when-should-i-use-them).

Technologies, L., 2012. Ercc rna spike-in control mixes. http://tools.thermofisher.com/content/sfs/manuals/cms_086340.pdf, Date: September 2012.

Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L., 2010. Cost-sensitive learning methods for imbalanced data. In: The 2010 International joint conference on neural networks (IJCNN). IEEE, pp. 1–8.

Thomas, H., 2006. Nondetects and data analysis: Statistics for censored environmental data.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., Rinn, J. L., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology 32 (4), 381–386.

Truică, C.-O., Leordeanu, C., 12 2017. Classication of an imbalanced data set using decision tree algorithms. University Politehnica of Bucharest Scientific Bulletin Series C - Electrical Engineering and Computer Science 79, 69–.

Unknown, 2020. EM Algorithm. Tech. rep., University of Cambridge, School of Clinical Medicine, MRC Bisotatistics Unit, (https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/EM_slides.pdf).

USEPA, J., 1998. Guidance for data quality assessment: Practical methods for data analysis.

Van Buuren, S., 2018. Flexible Imputation of Missing Data, 2nd Edition. CRC/Chapman & Hall, Boca Raton : FL, (https://stefvanbuuren.name/fimd/sec-MCAR.html).

van Dijk, D., Gigante, S., 11 2019. MAGIC - Markov Affinity-Based Graph Imputation of Cells. (https://cran.r-project.org/web/packages/Rmagic/Rmagic.pdf).

van Dijk, D., Nainys, J., Sharma, R., Kaithail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., Pe'er, D., 2017. MAGIC: A Diffusion-Based Imputation Method Reveals Gene-Gene Interactions in Single-Cell RNA-Sequencing Data. BioRxiv, 111591(https://www.biorxiv.org/content/10.1101/111591v1.full.pdf).

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al., 2018. Recovering Gene Interactions from Single-Cell Data using Data Diffusion. Cell 174 (3), 716–729, (http://www.sciencedirect.com/science/article/pii/S0092867418307244).

van Djik, D., Burkhardt, D., Vu, N., Krishnaswamy, S., 10 2018. Denoising and Imputing scRNA-Seq Data. (https://www.krishnaswamylab.org/blog/2018/10/28/denoising-noisy-gene-expression-in-scrna-seq).

Vidhya, A., 7 2020. KNNImputer: A Robust Way to Impute Missing Values (using Scikit-Learn). (https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/).

Wayne, 2011. Zero Inflated Models - "True Zero" vs. "Excess Zero". Cross Validated, (https://stats.stackexchange.com/q/13513).

Weiss, G. M., McCarthy, K., Zabar, B., 2007. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? Dmin 7 (35-41), 24.

Wolberg, W. H., Mangasarian, O. L., 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the national academy of sciences 87 (23), 9193–9196.

Yadav, S., 2020. Credit card fraud detection. https://www.kaggle.com/dark06thunder/credit-card-fraud-detection/notebook.

Yang, M. Q., Weissman, S. M., Yang, W., Zhang, J., Canaann, A., Guan, R., 2018. MISC: Missing Imputation for Single-Cell RNA Sequencing Data. BMC Systems Biology 12 (7), 114, (https://doi.org/10.1186/s12918-018-0638-y).

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., Smith, G. M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer Science & Business Media, (https://fukamilab.github.io/BIO202/04-C-zero-data.html).

[]

CURRICULUM VITAE

ROSE ADJEI

Address: 9301 Westbury Woods Dr, Apt E, Charlotte, NC, 28277 Phone: (928)-266-8023
GitHub: https://github.com/roseadjei04 Email: roseadjei04@gmail.com

Work Experience

Quantitative Analytics Specialist II Wells Fargo Bank - Risk Modelling Group July 2022 – Present
Financial Crimes Team

- Perform risk assessment and modelling to provide input and recommendations for financial crimes strategies and models.
- Conduct data reviews, review findings, determine risk level, and recommend fraud prevention strategies.
- Present findings from research to Senior Management and team members.

Graduate Research Assistant Utah State University - Math Department Jan 2022 – July 2022

- Met weekly with my advisor to discuss research progress and any challenges that arise.
- Performed data analysis using statistical packages.
- Conducted literature reviews.
- Wrote reports to summarize research results and prepare presentations.

Graduate Teaching Assistant Utah State University - Math Department August 2018 – December 2021

- Taught Trigonometry, College Algebra and Calculus Techniques at the undergraduate level.
- Tutored Bioinformatics at graduate level.

Wells Fargo Quantitative Analytics Summer Internship Program 2021 June - August 2021
MMB Risk Modelling Group (Remote)

- Did a research project on imbalanced data in Credit Risk modelling
- Built logistic regression models with Weight of Evidence (WOE) and applied models to credit risk data to address imbalanced data problem
- Presented findings from research to Senior Management and corporate risk team members.

Wells Fargo Quantitative Analytics Summer Internship Program 2020 July - August 2020
Financial Crimes Team (Remote)

- Assisted in building a reject inference model and applied it to a Financial Crimes Model
- Presented findings from Reject Inference to Senior Management

Summer Research Assistant Utah State University May – June 2020

Wells Fargo Quantitative Analytics Summer Internship Program 2019 June - August 2019
Consumer and Small Business Decision Support (San Francisco)

- Resolved three Model Risk Findings associated with Wells Fargo account management model submitted by the audit team.
- Created and filled in 7 new Risk Ranking Templates used to evaluate the impact of models on Wells Fargo Business.
- Wrote out PowerPoint to explain how Risk Ranking templates should be filled out.

Summer Instructor Northern Arizona University – Math Department June - July 2018
Taught Applied Statistics at the undergraduate level.

Graduate Team Support Member
Northern Arizona University– Peak Performance Mathematics Summer Bridge Program June - August 2017/18

- Oversaw a team of 7 undergraduate math coaches.
- Attended weekly meetings with program staff and Team Leaders to collaborate FAQs and challenges.
- Verified payroll for Math Coaches and aided in student recruitment

ROSE ADJEI

Address: 9301 Westbury Woods Dr, Apt E, Charlotte, NC, 28277 Phone: (928)-266-8023
GitHub: https://github.com/roseadjei04 Email: roseadjei04@gmail.com

Graduate Teaching Assistant Northern Arizona University – Math Department August 2016 - May 2018

- Taught Pre-Calculus Mathematics, Applied Statistics, Quantitative Reasoning at the undergraduate level

Teaching Assistant Kwame Nkrumah University of Sci.& Technology August 2015 - July 2016

- Tutored Financial Mathematics, Discrete Mathematics, Stochastic Processes, Probability and Statistics.
- Supervised undergraduate students with their academic research.

Domestic Operations and Retail Unit Intern Energy Bank Ghana Limited June - July 2015
Pension and Life Unit Intern Vanguard Life Assurance Company Limited June - July 2014

<u>Education</u>

Utah State University Graduation Date: June 2022
PhD in Statistics (GPA: 3.94/4.0)
Northern Arizona University Graduation Date: May 2018
Master of Science in Statistics (GPA: 3.58/4.0)
Kwame Nkrumah University of Sci. & Technology Graduation Date: July 2015
Bachelor of Science in Actuarial Science (Cumulative Weighted Average: 74.43% First Class Honors)

<u>Technical Skills</u>

- Intermediate proficiency in SAS, R, JMP, Latex, SPSS and proficient in Microsoft Office Suite
- High Performance Computing, GitHub

<u>Job-Related Skills</u>

- Good communication skills and able to work well in a team.
- Statistical Consulting and Data Analysis.
- Experienced with handling administrative issues.
- Trained in diversity, equality and managing bias in an educational and professional setting.

<u>Achievements and Honors</u>

Graduate Scholarship Award	Utah State University	August 2018 – July 2022
Placed Second in student presentation competition at Applied Statistics Conference, University of Florida		May 2021
Golden Key International Honor Society (GKIHS) Member		December 2017 - Present
Graduate Scholarship Award	Northern Arizona University	August 2016 - May 2018
Overall best female student	KNUST Actuarial graduating class	June 2015

<u>Research / Project Work</u>

NYC Housing Affordability with Shiny R	April 2020
Testing Missing High-Dimensional Count (RNA-Seq) data for Differential Expression	November 2019
Interval Censoring in Survival Analysis	April 2019
Time Series Models for Logan's Air Quality Index	December 2018
Applying Statistical Learning to Global Photosynthesis	December 2017
Call Option Modelling/Pricing	November 2016
Optimal Portfolio Choices on Investment in Ghana (Undergraduate Thesis)	May 2015

<u>Conference / Workshop</u>

Conference on Applied Statistics in Agriculture and Natural Resources,
Utah State University May 2022

ROSE ADJEI

Address: 9301 Westbury Woods Dr, Apt E, Charlotte, NC, 28277 Phone: (928)-266-8023
GitHub: https://github.com/roseadjei04 Email: roseadjei04@gmail.com

Fall Student Research Symposium 2020 - Evaluator December 2021
Evaluated undergraduate student research presentations during symposium.

JSM Virtual Conference – Speed Presenter August 2021
Comparative Analysis of Statistical Methods for Single-Cell RNA Sequencing Data

Conference on Applied Statistics in Agriculture and Natural Resources,
University of Florida May 2021

JSM Diversity Mentorship Program (DMP) 2020 participant August 2020

Medical University of South Carolina (MUSC) Summer Institute 2019 May 2019
Analysis of Zero Inflated Count and Semi-continuous Data Workshop

<u>Certifications & Training</u>
SAS Certified Base Programmer for SAS 9 May 2019

<u>Associations & Extracurricular Activities</u>

American Statistical Association (ASA)	- Member -	October 2018- present
Association of Women in Mathematics (AWM-Utah State University)	- President -	September 2018 - May 2019
International Association of Black Actuaries (IABA)	- Member-	January 2017- present
Graduate Students Association of Ghana (GRASAG)- USA	- Member-	September 2016 - present