CHAPTER I

INTRODUCTION

1.1 Background

It is quite an unavoidable fact that our lives are intertwined with data. Wherever a person lives, whichever company a person works under, whatever the profession, one certain constant is that everyone will come across data one way or another. According to the Cambridge dictionary, data is defined as information, facts or numbers that are collected to be examined and used to help decision making [1]. Companies, organizations and governments have come to a point of understanding the importance of harvesting data to the fullest. Processing, collecting and reporting data is essential in making the best decision in a given situation. Analysis of data will reveal trends and patterns, such as most sold items in the market or peak taxi hour in a region. Further processing the data, one might be able to predict future values of certain factors given the relation of the datasets. However, in many cases, a perfectly complete dataset might not be as easy to obtain as one might think. Even in a well-designed and controlled study, the occurrence of missing data is not improbable. With that in mind, missing data can reduce the statistical power, which refers to the probability that the test will reject the null hypothesis, therefore producing biased estimates and leading to inaccurate or invalid conclusions.

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest [2]. Conventional statistical methods and softwares will automatically assume that all the variables in a specified model are compulsory and must be measured for all cases. The simplest and most common strategy in dealing with absent values is to simply ignore cases which contains any missing data on any of the variables at hand by deleting them. This method is also known as listwise deletion (LD). Several research have shown the dangers of using this method. The most obvious drawback of this method is when given a dataset with high missingness value, listwise deletion method will choose to ignore any data with any amount of missingness, therefore deleting a large portion of the data, resulting in severe loss of statistical power. In regards of being reluctant towards erasing data that have been collected through spending great amount of time, money and effort, missing data imputation have attracted quite an attention in recent decades in the area of statistics and business [3].

The best possible method of handling missing data problem is to prevent the problem in the first place by well-planning the study and collecting the data very carefully. Nevertheless, non-responses and erroneous responses seem impossible to avoid in surveys, even when a lot of attention is devoted to data collection. Other causes such as human error, equipment error, and lack resources are commonly faced. An answer for missing data problems comes in the form of imputation methods.

The method of substituting an estimated value for the missing or inconsistent data field is called Data Imputation [4]. The missing data imputation emerges in many areas that interacts with data sets. Medical documentation, survey, census, and data modelling are just some examples of the areas that the problem of imputation of missing data may occur. Techniques on how to treat missing values with estimations have surfaced since 1977 [5]. However in recent years, due to the rise of popularity in machine learning methods, the popularity in research towards data imputation have also risen significantly. This is due to how machine learning and neural network systems learn, by using datasets to train itself, creating a model for a specific use case. Because these methods creates a model from the dataset for the dataset, it obviously faces a big issue in facing missing data which hinders these methods to learn from the respective datasets, hence the rise in popularity. Interestingly, analysts have shifted towards using machine learning methods to impute datasets that will be used as training data for another machine learning project. Modern problems require modern solutions, using machine learning to impute missing dataset for machine learning purposes.

Several strategies inspired in statistics and machine learning have been developed to address this problem. Literature that discusses this topic reveals that the efficacy of the proposed methods strongly depends on the problem domain. The number of cases, number of variables and pattern of missingness are just a few examples of what effects the efficacy of the method. Thus, there is yet a clear indication that favours one method over the others. Therefore, comes a need for an analysis in methods of approaching an imputation problem in a given domain.

Since imputation methods had grown interest in the statistician community, several researches comparing the use of statistical imputation has been made regarding certain topics. A simple but popular approach, yet well-known to produce biased estimates, is to substitute means for missing values which is referred to as mean imputation. Although still problematic, linear regression based imputation methods are often indefinitely much better. A more complex method of imputation which researches has shifted their focus towards to is multiple

imputation, which is a method that implements single imputation multiple times to produce the optimal result. Multiple imputation methods has been around in one form or another for at least three decades but has only recently become fully developed and incorporated into widely-available and easily-used softwares.

On the other end, artificial Neural Networks (ANN) are appearing as useful alternatives to traditional statistical modelling techniques in many scientific disciplines. Imputation methods inspired by machine learning are based on the construction of a predictive model to estimate absent values from the information available in the data set. Well-known learning algorithms such as multi-layer perceptron (MLP), K-Nearest-Neighbours (KNN), Self-Organizing Maps (SOM) and Decision Tree (DT) construction algorithms have been commonly used as imputation methods in different problem domains in the scientific field.

One such dataset that faces missing data problems are life expectancy data. Life expectancy is the average number of years at birth a person could expect to live if current mortality trends were to continue for the rest of that person's life [6]. Analysis of life expectancy often results in being used to appraise the overall health of a given population. Documentation of life expectancy data and the trends in health of several populations have been recorded by both national governments and united organizations such as World Health Organization by the United Nations. However, missingness is still a reoccurring problem in these types of datasets due to a diverse problem. Governments may disclose information, fail to comprehensively collect data or is focusing on other matters. It is then quite understandable that a biased conclusion from the data will result in a serious issue since it correlates strongly with life and death in a given region. In the past few decades, the trend in life expectancy of humans has been a slow and steady increase. Despite widespread knowledge to reduce the severity of problems, observed trends in life expectancy drops, such as high obesity in the United States, has continued to worse. Health and life expectancy of current and future generations are threatened to diminish if no action is taken.

Researches regarding live expectancy, such as Isaac Sasson's analysis on trends of lifespan variation [7], James W. Shaw's analysis of the OECD Health Data [8] and Hanna van Solinge's 10-year panel study among older workers [9] has shown to face problems regarding missing values in its respective datasets. While some opted to use multiple imputation (MI), an optimal preferred imputation method is yet to be clear. Studies have been done comparing imputation methods in longitudinal studies [10], time-varying covariate data [11], individual participant data [12] and breast cancer data [13] with some using only statistical methods and

others also involving machine learning methods. This thesis aims to compare both statistical imputation methods and machine learning methods in life expectancy data to determine an optimal method for cases with its type of missingness pattern.

In this thesis, six well-known imputation methods, three in which are statistical method and the other three are machine learning methods, are used to impute absent values in the World Health Organization's and World Bank's Life Expectancy Data set, which contains 3104 case records of social, economic and health data that correlates with life expectancy of nations around the globe. A simulation of missing data will be done from the dataset by extracting non-missing data, simulating missingness and imputating simulated missing values using the six methods to conclude a performance comparison with known missing values. The complete missing dataset will then be imputed and the statistical characteristic changes after imputation will be compared throughout all six methods.

The statistical method used in this thesis are Mean Imputation, Hot-Deck Imputation and Multiple Imputation (MI). Mean imputation is chosen to represent the most basic form of imputation, which is customarily built in implemented in most mathematical softwares. Predictive Mean Matching (PMM) is a popular imputation category that was recently suggested in 2010 [14]. Hot-Deck nearest neighbor imputation is chosen to represent the basic form of PMM based imputation. The Hmisc library's MI method is used to represent a category of MI which still falls in the basis of PMM and also stands for the MI extended method of the discussed Hot-deck [15].

The three machine learning methods used in this thesis are K-Nearest Neighbor (KNN), Self-Organizing Maps (SOM) and Multilayer Perceptron (MLP). Representing supervised machine learning, MLP is well known to be one of the most common, flexible and powerful machine learning models out there [13]. MLP regressors have been previously proven to perform adequetly in multiple dataset and cases but its performance heavily relies on the data and its missingness characteristic. Only been recently suggested by Kohonen, SOM, a unsupervised machine learning algorithm, has surprisingly shown adequete performance in imputing missing values in intelligent tutoring systems and a real-size transport survey database [13]. A machine learning method that falls in the PMM category, KNN has previously been used in dealing with incomplete data for DNA micro-arrays [13].

1.2 Problem Statement

The main problem that will be discussed on this thesis is determining which is the best method to impute missing data in the domain of life expectancy. With that said, three specific research problems are as follows.

1. Which statistical imputation method are best suited for life expectancy cases?
2. Which machine-learning imputation method are best suited for life expectancy cases?
3. Which method produces the least bias in imputating life expectancy values?

1.3 Objectives

The main objective of this thesis is to experimentally verify our method according to our problem statement. Listed below are the outline of our main objective in this thesis:

1. conclude the best statistical imputation method for life expectancy cases,
2. conclude the best machine-learning imputation method for life expectancy cases, and
3. conclude the best imputation method that yields the most accurate result in imputing life expectancy data.

1.4 Restrictions and Assumptions

Several key remarks concerning the restrictions of our problems to cap the complexity and focus of this thesis are as follows:

1. data records from pre-2000 era will not be calculated due to several crisis's that arose at the referred period,
2. terrorism and other unexpected force majeure crisis are not accounted to our dataset, and
3. methods used to account for the dataset uses different models for different periods in time but are uniformed for all region and checked comparatively.

1.5 Benefits

The benefits of this research can be categorized into practical and theoretical benefits, as follows.

1. Theoretical Benefit: Helps future researchers develop and compare various imputational methods, both statistical and machine learning in hopes that future research may be aided.
2. Practical Benefit: Helps organizations, researchers and readers with insight in choosing the best method to handle missing data based on the missingness pattern of the dataset. Choosing the appropriate method will maintain, or even improve, statistical power of the dataset thus allowing the further push of research and development regarding life longetivity.

1.6 Thesis Structure

The writing structure of this thesis is as detailed below.

1. Chapter I describes the background, problem statement, objectives, restrictions, and methodology.
2. Chapter II describes the statistical theories used to support this research. The theories used also serve a purpose of answering our problem statement followed by a detailed review of the statistical imputation methods used for the imputation of missing values.
3. Chapter III describes the machine learning theories used to support this research. The theories used also serve a purpose of answering our problem statement followed by a detailed review of the machine learning imputation methods used for the imputation of missing values.
4. Chapter IV describes the dataset that will be imputed with both the statistical and machine learning methods discussed in the previous chapter followed by a review of comparison experiment for the dataset
5. Chapter V describes the various experiments and results for the imputed dataset followed by an analysis of the performance of each method.
6. Chapter VI summarizes achieved results, draw conclusions and proposes directions for further research.