KWAME NKRUMAH UNIVERSITY OF SCIENCE AND

TECHNOLOGY, KUMASI

A COMPARISON OF MULTIPLE IMPUTATION TECHNIQUE WITH LINEAR INTERPOLATION METHOD FOR TIME SERIES DATA

BY

EMMANUEL ACHEAMPONG

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MSC. APPLIED STATISTICS

SEPTEMBER, 2019

Declaration

I hereby declare that this submission is my own work towards the award of the MSc. Applied Statistics degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Emmanuel Acheampong (PG4411715) Certified by:	..... Signature	..... Date
Nana Kena Frempong (PhD) (Supervisor)	..... Signature	..... Date
Certified by: Prof Mrs. Atinuke Olusola Adebanji (Head of Department)	..... Signature	..... Date

Dedication

I dedicate this thesis to Almighty God; my creator, strong pillar, source of inspiration, wisdom, knowledge and understanding. He has been my source of strength throughout this course and on His wings have I soared. I also dedicate this work to my family. A special feeling of gratitude to my beloved wife, Diana Acheampong and my loving parents, Philip Num and Mary Acheampong whose words of encouragement and push for tenacity ring in my ears. God bless you.

Abstract

This thesis evaluates the performances of Multiple Imputation Technique (MIT) and Linear Interpolation methods for the estimation of missing values in a time series data (CO_2 emissions data under the Fuel combustion sub-category of the Energy sector. Under this sub-category, data of two codes namely; i) Energy industries and ii) Manufacturing Industries and Construction were used).

The performances of both methods were then compared using two notable indicators; the Mean Absolute Error (MAE) and the Mean-Square Error (RMSE). This thesis highlights some advantages and limitations of each method compared with the other, thereby providing suggestions on which method to be used under prevailing conditions.

Acknowledgements

First and foremost, praises and thanks to God Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor, Nana Kena Frempong (PhD) for giving me the opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honour to work and study under his guidance. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, empathy, and great sense of humour. I am extending my heartfelt thanks to all his support staff who assisted in one way or the other. I am deeply humbled by their immense support.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my beloved wife and in-laws for their love, understanding, prayers and continuing support to complete this research work.

Also I express my thanks to my brothers and sisters for their support and valuable prayers. I also express my thanks to my MSc. Applied Statistics colleagues, lecturers for their support. I thank the management of IDL, KNUST for their support to do this work.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Contents

List of Abbreviation

UNFCCC ..... United Nations Framework Convention on Climate Change

1/CP.16 .....Conference of the Parties at its Sixteenth Session

NIR .....National Inventory Report IPCC .....

Intergovernmental Panel on Climate Change

CO_2 ..... Carbon Dioxide CH_4

.....Methane

N_2O .....Nitrous Oxide PFCs

..... Perfluorocarbons

IPPU ..... Industrial Processes and Productive Use AFOLU

..... Agriculture, Forestry and Other Land Use

MIT .....Multiple Imputation Technique

GHG .....Greenhouse Gas

MCAR ..... Missing Completely at Random MAR

..... Missing at Random

CO .....Carbon Monoxide UNEP

.....United Nations Environment Programme UN

.....United Nations FAR

..... First Assessment Report

SAR ..... Second Assessment Report

TAR ..... Third Assessment Report AR4 .....The Fourth Assessment Report AR5 .....The Fifth Assessment Report

SR ..... Special Reports AR6 ..... The Sixth Assessment Report

EF .....Emission Factor NEU ..... Non-Energy Use

NOAA .....National Oceanic and Atmospheric Administration

GWP ..... Global Warming Potential

PPM .....Part Per Million EMEP .....European Monitoring and Evaluation Program

EEA ..... European Economic Area SEM .....Structural Equation Modeling

E-M ..... Expectation - Maximization MNAR .....Missing Not at Random

MAE .....Mean Absolute Error RMSE .....Root Mean Square Error EPA ..... Environmental Protection Agency Mt ..... Million tonnes $MtCO_2e$ ..... Million tonnes Carbon Dioxide Equivalent

Min .....Minimum Max ..... Maximum NA ..... Not Applicable

LI .....Linear Interpolation MAPE .....Mean Absolute Percentage Error

MAD ..... Mean Absolute Deviation

MSD .....Mean Square Deviation

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

This chapter discusses the background and problem of the study. Areas being researched will be reviewed using prevailing data surrounding the issue, preceding researches on the issue, and applicable history on the issue. It also discusses rationale of the project, objectives and methodology for undertaking the research.

1.1 Background

The only best solution to the problem of missingness in data is not to have any. So in the life cycle of research projects, it is vital to put great effort into lessening the incidence of missing data. Statistical modifications can never be a substitute for sloppy research (Paul D. Allision, 2001)

One of the obligations of Ghana (i.e. being part of the United Nations Framework Convention on Climate Change (UNFCCC)), is to develop, publish and regularly update the national communication, as well as its national emission inventories. This is captured under the first paragraph of Article 12 of the 14th pact and it states "Parties are mandated to communicate to the conference of the parties, a national inventory of anthropogenic emissions by sources and removals by sinks of all greenhouse gases not controlled by the Montreal Protocol, to the extent it capacities permit, using comparable methodologies to be promoted and agreed upon by the Conference of the Parties"

In furtherance to the above mentioned article of the Convention which deals with reporting requirements, paragraphs 60 (a-c) of decisions 1/CP.16 introduced to improve reporting regime, which parties are to prepare and submit every four years a national communication and Periodic update report to the Conference of

Parties.

Hence, Ghana prepared a National Inventory Report (NIR) which captures updated versions of Greenhouse Gases (GHG) emissions estimate of 1990-2012 from four major economic sectors. The compilation of the NIR was done using the Intergovernmental Panel on Climate Change (IPCC) 2006 guidelines.

The scope of the inventory covers sources of greenhouse gas emissions, which is caused by anthropogenic activities for direct greenhouse gases, such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O) and Perfluorocarbons (PFCs) and their removal sinks. The greenhouse gases inventory is conducted for main productive sectors that support Ghana's economic development. The emission/removal levels of the various economic sectors largely hinges on: deployment levels of environmentally sound technology; sector mitigation policy drive; and how much sustainability underpin sector productivity.

The emissions/removals occurrences as a result of economic activities in Ghana have been categorized under four sectors defined by the 2006 IPCC guidelines. The sectors include; Industrial Processes and Product use (IPPU), Energy, Waste, Agriculture Forestry and Other Land Use (AFOLU). The key economic activities that contribute to the release (capture) of greenhouse gases into (from) the atmosphere are represented by these sectors.

Under the Energy sector, following factors influences the levels of carbon emissions from different energy activities; volumes of fuels consumption, rate of operations, technological types, and conditions of the environment. The energy sector activities are classified into combustion and fugitive sources, according to the 2006 IPCC guideline. The activity types and the processes through which carbon emissions are produced are highly considered, in the classifications of the activities.

The combustion sub-sector consist of both stationery and mobile sub-sectors. The stationery combustion sub-sector was the focus of attention for this study. The

stationery combustion carbon emissions mainly are emanated from point source operation in power plants, industrial boilers, refinery plants, standby generators, household, commercial cooking devices etc. The stationery combustion sources are disaggregated into the following IPCC codes; Energy industries, Manufacturing industries and Construction and Other sectors.

This study focuses on providing a credible alternative approach in addressing gaps found in the time series data collected to measure CO_2 emissions under the Fuel combustion sub-category of the Energy sector. Under this sub-category, data of the following two codes were collected: i) Energy industries and ii) Manufacturing Industries and Construction were chosen for the research. The diagram below indicates CO_2 emissions emanating from Energy Industry; and Manufacturing and Construction activities under the Energy Sector.

Figure 1.1: CO_2 emissions from Energy Sector

According to the second Ghana NIR, "gaps were identified in data obtained, regional and international sourced data were used to fill the missingness. In the occasion that data was not obtainable in the regional and international sources, four main statistical approaches namely, the Overlap, Surrogate data, Interpolation and trend extrapolation in tandem with the IPCC good practice guidance were employed to create the missing data." These approaches are otherwise termed as conventional methods in dealing with missingness of data.

This study, therefore, focuses on Multiple Imputation Technique (MIT) as a more plausible alternative method to dealing with gaps in the selected data. The MIT is well thought-out as "state of the art" missing data methods (Schafer and Graham, 2002) and are extensively recommended in the methodological literature (Schafer and Olsen, 1998; Allison, 2002; Enders, 2006). This approach is superior to conventional missing data techniques since they produce unbiased estimates with both MCAR and MAR data.

1.2 Problem Statement

IPCC Guidelines provide the procedural guidance to parties for their greenhouse gas emissions (GHG) and removals inventory annual reporting to the United Nations Framework Convention on Climate Change (UNFCCC). The approaches captured in the IPCC Guidelines vary in their complexity ranging from the simplest first Tier method, based on globally or regionally applicable default parameters, through second Tier methods based on the specific data of countries (Robert and Reuben, 2010)

In essence, the quality of data collected will inform the method (i.e. contained in the IPCC guidelines) to adopt for estimating the levels of emissions/removals. This implies that, when the processes of collecting the time series data is not representative enough, the resulting estimates are likely to suffer deficiencies and might not represent the true picture of the process. Ghana strictly adheres to the IPCC protocols and guidelines in gathering information on GHG emissions.

Data missingness are pervasive in quantitative studies. For the reason of its unescapable nature, it has been described by many as "one of the most central statistical and design problems in research" Azar (2002) . Notwithstanding the important nature of the problem, substantive researchers normally adopt old stand-in methods that have been cautioned in the procedural literature.

A clear example is the approach the Technical Research Committee of EPA adopts in dealing with gaps in data when reporting on the National Inventory Report. The committee has adopted some statistical approaches namely; Interpolation and trend extrapolation and others; consistent with the IPCC good practice guidance to generate the missing data in accordance with the nature of missingness in data.

These approaches have been described by many as subjective. This is because, that the nature of the missingness will determine the approach to adopt. This creates the possibility of choosing an inappropriate approach in the event of the occurrence of missingness in a given data. With this research, we want to explore a more objective approach (the Multiple Imputation method) which is able to resolve all kinds of missingness in data.

1.3 Study Objectives

The overall goal of this study is to estimate the missingness in the annual time series data of CO_2 and CO gas emissions, collected over a period of 20years (that is, from 1990 to 2015) from the selected energy sector, using the multiple imputation method.

Specifically, the study seeks to adopt this more credible scientific approach to;

- Estimate the missingness of CO2 and CO gas emissions data from Manufacturing and construction industries under the Fuel combustion subcategory of the Energy sector using multiple imputation.
- Estimate the missingness of CO2 gas emissions data from Energy industries under the Fuel combustion sub-category of the Energy sector using multiple imputation method.
- Compare the results of the above estimates with the estimates of the conventional extrapolated techniques currently adopted by Ghana.

1.4 Justification

Since the current method employed in dealing with missing data has been described as subjective, a statistical tool that can impute the missing values so that maximum amount of information is restored while keeping the data unbiased, has become more imperative to adopt. The Multiple Imputation Technique (MIT) will improve the estimation of CO_2 emission under the IPCC guidelines when there is missing data.

1.5 Limitation of Study

Generally, there was one main challenge that had the potential of negatively affecting the outcome of the research. This was the insufficient data from other Greenhouse gases sources such as CH_4, N_2O, NO_x etc. to compliment CO_2 emissions data (that is, CO was the only auxiliary variable obtained with sufficient information). Auxiliary variables are observed variables that are distinct from variables of interest in a given model (in this study, CO_2 is the only variable of interest). Their addition to models only to improve estimates, decrease error variance and thus, increase statistical power and precision of estimates (this pertains to analysis of variables with missing data).

1.6 Organizatoin of Thesis

The research consists of five chapters: The first chapter discusses the introduction which is made up of background of study, problem statement, objectives of the study, justification of the study and the study limitations. The second chapter provides an overview (a review) of other researchers works that are related to this study. It additionally discusses the various ideas in the research areas of missingness and possible mitigation methods.

Chapter Three discusses the methodical approach employed by the researcher to achieving the objective of the research. The chapter four discusses data presentation, analysis and discussion. The last Chapter, Chapter Five, also present the summary of the findings, together with the conclusion and recommendations.

CHAPTER 2

LITERATURE REVIEW

This section discusses the various study and other researches related to GHGs emissions, its time series missing data and the possible ways of dealing with it. Many data are found in fields such as the environmental studies, epidemiological studies, forestry studies and other fields of study. Once there is data gathering in such field of study, then there is the probability that there would be missing value within the dataset.

2.1 The IPCC Guidelines

In 1988, the United Nations Environment Programme (UNEP) and the World Meteorological Organisation set-up the Intergovernmental Panel on Climate Change (IPCC). It was subsequently endorsed by UN General Assembly in the same year. Its initial task was to prepare a comprehensive appraisal and recommendations regarding the state of knowledge of the science of climate change; the social and economic impact of climate change, and potential response strategies and elements for inclusion in a possible future international convention on climate, as outlined in UN General Assembly Resolution 43/53 of December 6, 1988.

Since then, five assessment cycles together with its Assessment reports have been delivered to the IPCC, the most comprehensive scientific report about climate change produced worldwide. It has also produced a range of reports, and technical papers, in response to request for information on specific scientific and technical matters from the United Nations Framework Convention on Climate Change (UNFCCC), governments and international organisations.

Since IPCC came into being, each Assessment report has fed directly into international climate policymaking. The First IPCC Assessment Report (FAR) which underlined the importance of climate change as a challenge with global consequences and requiring international cooperation, was produced in 1990. It played a pivotal role in the creation of UNFCCC, the main international pact to reduce global warming and cope with the impact of climate change.

The Second Assessment Report (SAR, 1995) provided vital material for governments to draw from in the lead-up to adoption of the Kyoto Protocol held in 1997. The Third Assessment Report (TAR, 2001) was geared towards the impacts of climate change and the necessity for adaptation. The Fourth Assessment (AR4, 2007) set-up the basis for a post-Kyoto agreement, concentrating on limiting warming to 2o C. Between 2013 and 2014, the Fifth Assessment Report (AR5) was completed. It provided the scientific input into the Paris Agreement.

Currently, the IPCC is in its Sixth Assessment cycle where it will come out with three Special Reports, a report on Methodology and the Sixth Assessment Report. The first of these special reports, Global warming of 1.5oC (SR), was requested by world governments under the Paris agreement. In May of 2019, the IPCC will complete the 2019 refinement (an update to the 2006 IPCC guidelines on National Greenhouse Gas inventories). The sixth Assessment report (AR6) is expected to be completed in 2022 for the first global stocktake in 2023.

2.2 Protocols for GHG emissions data collection

The IPCC Guidelines involve procedures for the estimation of greenhouse gas emissions and removals. Users are invigorated to go beyond these minimum default approaches where possible. IPCC systems use the following concepts:

1. Good Practice:

   This is to ensure the development of high-quality national greenhouse gas inventories, an assembly of procedural principals, actions and procedures as collectively defined in the guidelines. These definitions has achieved general acceptance amongst countries as the basis for inventory development.

2. Tiers:

   A tier is a level of procedural complexity. Three tiers are usually provided. The first tier (Tier 1) is the basic method, second tier, intermediate and third tier, most demanding in terms of complexity and data requirements. Second and third tiers are sometimes referred to as higher tier systems and are generally measured to be more precise.

3. Default data:

   The first tier approaches for all classes are intended to use readily available national or international statistics in blending with the provided default emission factors and additional parameters that are provided, and therefore should be achievable for all states (IPCC, 2006).

2.3 Gaps in Data

Gaps in data may arise due to; a new emission factor (EF) or method is applied for which past data are not available, new activity data become available, but not for historical years, there has been a modification on how the EF is developed or activity data are collected, or activity data cease to be available, a new source or sink category is added to the inventory, for which historical data are not available and errors are identified in historical data or calculations that cannot easily be corrected.

Splicing and gap-filling approaches, combining more than one technique or data series to form a complete time series, help lessen potential inconsistencies. It is good

practice to perform the splicing using more than one method before making a final choice and to document why a specific technique was chosen (IPCC, 2006).

Difficulties can ensue concerning how to differentiate between energy and nonenergy use of feedstocks in the Industrial Processes sector in their national GHG inventories, in the energy sector. Non-energy-related physical and chemical processes in production activities leading to the transformation of raw materials and emissions of greenhouse gases (e.g., decomposition reaction) are considered as Non-Energy Use of feedstocks (NEU) (IPCC, 2006).

It also comprises of feedstock in process reactions or stage processes that not only release heat but also act largely as reducing agents (e.g. metallurgical coke in the smelting of ores in metal production). Meanwhile, the energy sector accounted for the energy/heat required for initiating and/or sustaining chemical reactions kinetically and thermodynamically (IPCC, 2006).

2.4 The Greenhouse Effect

The greenhouse effect is mostly caused by the interaction of the sun's energy with greenhouse gases such as carbon dioxide, methane, nitrous oxide and fluorinated gases in the Earth's atmosphere. The ability of these gases to capture heat is what causes the greenhouse effect. Greenhouse gases consist of three or more atoms. This molecular structure makes it possible for these gases to trap heat in the atmosphere and then transfer it to the surface which further warms the Earth. This uninterrupted cycle of trapping heat clues to an overall increase in global temperatures. The procedure, which is very similar to the way a greenhouse works, is the main reason why the gases that can produce this outcome are collectively called as greenhouse gases. The prime forcing gases of the greenhouse effect are: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and fluorinated gases (Kweku et at., 2017).

2.5 Greenhouse Gases and Global Warming

Human activities mostly cause Greenhouse gases such as; carbon dioxide, methane, nitrous oxide, and halogenated compounds emissions to occur, although some do occur naturally. Infrared radiations are absorbed by greenhouse gases and heat trapped in the atmosphere, in so doing enhances the natural greenhouse effect defined as global warming. This natural occurrence, make life on earth possible through the atmospheric warming, without which life on earth would have been impossible with low temperatures (Kweku et at., 2017).

Michael Daley, an associate professor of Environmental Science at Lasell College said "Gas molecules in its substantial amount, can force the climate system. These type of gas molecules are called greenhouse gases." The net consequence is the steady heating of Earth's atmosphere and surface, and this process is called global warming (EPA, 2013).

These GHGs consists of; nitrous oxide (N_2O), water vapor, CO_2, methane, and other gases. The scorching of fossil fuels like coal, oil, and gasoline have significantly increased the concentration of greenhouse gases in the atmosphere, since the emergence of the Industrial Revolution in the early 1800s, precisely CO_2. Daley added that deforestation is the second largest anthropogenic basis of carbon dioxide to the atmosphere spanning between 6% and 17% (EPA, 2013).

Production and consumption of fossil fuels, bush burning, waste from incineration processes, use of various chemicals agriculture and other industrial activities are instances of human activities that have increased the concentration of greenhouse gases (GHG), particularly CO_2, CH_4, and N_2O in the atmosphere making them harmful.

This upsurge in greenhouse concentration has prompted environmental change and an unnatural weather change impact, which is inspiring worldwide endeavors, for example, the Kyoto Convention, consenting to of Paris agreement on environmental change and different activities to control the negative results of the greenhouse

effect. The role of a greenhouse gas to global warming is usually articulated by its global warming potential (GWP) which empowers the correlation of global warming effect of the gas and that of a reference gas, especially carbon dioxide.

Since the start of the Industrial Revolution, atmospheric CO_2 intensities have increased by more than 40% from around 280 parts per million (ppm) in the 1800s to 400 ppm. The last time Earth's atmospheric CO_2 levels reached 400 ppm was between 5 and 3 million years ago during the Pliocene Epoch, according to the University of California, the San Diego Scripps Institutions of Oceanography.

The greenhouse effect, together with the growing levels of greenhouse gasses and the resulting global warming, is expected to have profound consequences, according to the scientific consensus. If global warming goes unhindered, it will cause significant climate change, rising sea levels, increasing ocean acidification, life-threatening weather events and other severe natural events.

2.6 Missingness of Time Series Data

Missing data are ubiquitous throughout the social, behavioural, and medical sciences. For decades, researchers have relied on a variety of ad hoc techniques that attempt to "fix" the data by discarding incomplete cases or by filling in the missing values. Sadly, most of these techniques require a relatively strict assumption of the cause of missing data and are prone to significant bias. Methodological literature has increasingly favored these methods (Little Rubin, 2002), and they are still widely used in published research articles (Bodner, 2006; Peugh and Enders, 2004).

The ubiquitous nature of missing data is further explained by (Allison, 2001) when he remarked that, missing data are ubiquitous in psychological research. By missing data, it implies, data that are missing for some (but not all) variables and for some (but not all) cases. If data on a variable is missing for all cases, this variable is said to be latent or not observed. On the other hand, if data on all variables is missing in

some cases, we have what is called unit non- response rather than item non-response. Methods for latent variables or unit nonresponse will not be dealt here, although some of the methods we will consider can be adapted to those situations.

For nearly a century, methodologists have been studying missing data problems, but major breakthroughs occurred in the 1970s with the advent of maximum probability estimation routines and multiple imputations (Beale & Little, 1975; Dempster, Laird, & Rubin, 1977; Rubin, 1978b; Rubin, 1987). At the same time, a theoretical framework for missing data problems was outlined by (Rubin, 1976), which is still widely used today. Over the past 30 years, the methodological literature has received considerable attention to maximum probability and multiple imputations, and researchers generally regard these approaches as the current "state of the art " (Schafer & Graham, 2002).

Regarding traditional approaches, maximum probability and multiple imputations are theoretically attractive, as they require weaker assumptions about the cause of missing data. In practice, this means that these methods produce estimates of parameters with less bias and greater power. Researchers have taken the maximum likelihood and multiple imputations relatively slowly and still rely heavily on traditional missing data handling techniques (Bodner, 2006; Peugh and Enders, 2004).

Partly because of the lack of software options, this hesitancy may be due to the fact that the maximum probability and multiple imputations were not widely available in statistical packages until the late 1990s. The technical nature of the missing data literature is probably another major barrier to the widespread use of these techniques. The primary objective of this study is therefore to analyze missing data with a particular emphasis on multiple imputations.

(Allison, 2001) further revealed that conventional statistical methods and software assume that for all cases, all variables in a given model are measured. For virtually all statistical software, the default method is simply to delete cases with missing data on

the variables of interest, a method known as list deletion or complete case analysis. The most obvious disadvantage of list deletion is that a large fraction of the sample is often deleted, leading to a serious loss of statistical power. This is one of the reasons why missing data is problematic in his opinion.

King et al., 2001) showed that using listwise deletion under the most optimistic of assumptions, with the average amount of missingness apparent in political science articles, the estimates of a standard error are farther from the truth than failure to control variables with missingness. The strange assumptions that would make the listwise deletions better than multiple imputations are that we know enough about what our observed data generated not to trust them to impute the missing data, but we still trust the data sufficiently to use it for our subsequent analyses. It has long been recognized that a sound, more principled approach is desirable and considerable efforts in this direction have been made. Much of it comes from the seminal work in (Little and Rubin, 1987) (a second enlarged edition was published in 2002). The monograph (Schafer, 1987) describes a methodology for dealing with missing data in cross- sectional data in considerable detail and (Rubin, 1996) provides a useful overview of ideas on multiple imputations and their impact on statistical practice.

However, the literature on missing data in (multiple) time series is scarce. The lack of time series data is considered conceptual in (Little and Rubin, 2002) and can be handled in the same way as cross-sectional data. The problem is, however, that they are both harder and more pressing. Because there is an additional level of complexity when dealing with multivariate time series, it is harder to consider both contemporary and lagged relationships between components when imputing a missing data point. More pressing, because the use of complete data records alone is no longer feasible.

With cross-sectional data, the discarding of records with completely random data (MCAR) has no other effect than the reduction of the available sample. In a time

series, every record is unique: it would leave us with a series of holes, unusable for many purposes. In the last 15 years, timeseries modelling has been more widely accepted and is now a well-established tool in the applied statistician's kit. Several theoretical breakthroughs such as the smoother simulation[(Harvey et al.,2004), (De Jong, 1995), (Durbin and Koopman, 2002) and Markov Chain Monte Carlo (see for example (Gamerman, 1997)), In addition to the constantly increasing computing power on the desktop, our chances of dealing with missing data in multivariate time series have improved.

In statistics, missing data or missing values occur when no data value for the observation variable is stored. This is mainly due to manual data entry procedures, errors in equipment and incorrect measurements.

2.7 Issues with data availability

According to European Monitoring and Evaluation/ European Economic Area (EMEP/EEA) emission inventory guidebook (2009), under the complete and consistent time series, data availability must be determined for each year. It will be difficult to recalculate previous estimates using a higher level method or to develop estimates for new categories if data are missing for one or more years.

Periodic data: Statistics on natural resources or the environment, such as national forest inventories, waste statistics and agricultural statistics, may not cover the whole land on an annual basis. Alternatively, they can be taken at intervals such as every fifth or tenth year or region by region, which means that national level estimates can only be obtained directly once the inventory has been discharged in each region (EMEP/EEA, 2009).

When data are available less often than annually, several problems arise. Firstly, estimates must be updated whenever new data are available and the years between the available data must be recalculated. The second problem is the production of

inventories years after the last available data point and the availability of new data. New estimates should in this case be extrapolated on the basis of available data and then recalculated when new data are available (EMEP/EEA, 2009).

Changes and gaps in data availability: a change in data availability or data gap is different from regularly available data because it is unlikely that there will be an opportunity to recalculate the estimate using better data at a later date. In some cases, countries will improve their ability to collect information over time, so that higher levels of information can be used in recent years but not in previous years. This is especially relevant for categories in which direct sampling and measurement programs can be implemented, since these new data may not be indicative of the conditions in recent years (EMEP/EEA, 2009).

Some countries may find that the availability of certain data decreases as a result of changes in government priorities, economic restructuring or limited resources over time. Some countries with economies in transition may no longer collect certain data sets that could be used in the base year, or these data sets may contain different definitions, classifications and collection levels if available (EMEP/EEA, 2009).

2.8 The concept of missing data

Many researchers have faced the challenges of handling of Missing data in the field of research. According to the author (Graham, 2009), the approach that most researchers used to handle data in the twentieth century was based on the assumption that most of the dataset were complete with no missing data value(s) in it. It is recently that most statistician and mathematicians decided to find ways of dealing with the missing observation in the field of research.

Previous studies reveal that missing data analysis began to gain attention somewhere in the 1987, even though some authors have written few articles on it before but it was not popular by then [Dempster et al. 1977, Heckman 1979, Rubin 1976]. In the

year 1987, what happened was the introduction of two main or important publication of books in the world of statistics by (Little and Rubin, 1987) with the title of the book statistical analysis with missing data with it second edition been published in the year 2002.

In the same year 1987, Rubin published his book, of the title multiple imputation for nonresponse in survey. These publications by the authors coupled with the development of an application or missing data analysis software laid down the foundation and analysis of missing data for the upcoming years even up to now. Studies have shown that in the year 1987, there were two publications of important articles; firstly, the missing data analysis using Structural Equation Modelling (SEM) software authors (Allison,1987 , Muth'en et al.,1987).

The SEM is one of the multivariate Statistical technique for the analysis of data and is capable to handle data that have some challenge of nonresponse or missing observation in it. The author Tanner & Wong (1987) also came out with an article on data analysis technique which could be used for the multiple imputation (MI) software in the near future. All these were some of the occurrence that took place when it comes to the area of missing data analysis.

Approach	Applicability	Notes
Overlap	This requires that both the formerly used and new approaches to be applied must be available for a minimum of one year, preferably more.	• Most reliable when the overlap between two or more sets of annual estimates can be assessed. • If the trends observed using the formerly used and new approaches are inconsistent, this approach is not good practice.
Surrogate Data	Emission factors, activity data or other estimation parameters used in the new technique are strongly interrelated with other familiar and more readily-available indicative data.	• Many indicative data sets (singly or in combination) should be tested in order to determine the most strongly correlated. • Should not be done for long periods
Interpolation	Data required for recalculation using the new technique are available for sporadic years during the time series.	• Estimates can be linearly interpolated for the periods when the new method cannot be applied. • The method is not applicable in the case of large annual fluctuations.
Extrapolate Data	Data for the new technique are not taken annually and are not available at the commencement or end of the time series.	• Most reliable if the trend over time is constant. • Should not be used if the trend is changing (in this case, the surrogate method may be more suitable). • Should not be done for long periods
Other Techniques	The normal alternatives are not valid when technical circumstances are changing during the time series (e.g., due to the introduction of mitigation technology).	• Document customized approaches thoroughly. • Compare results with standard techniques.

Figure 2.1: Approach to Addressing Data Gaps

CHAPTER 3

METHODOLOGY

This chapter discusses te adopted approach to dealing with missing and its associated limitations.

3.1 Patterns of the missing observation

Missing dataset can have different format or take different patterns, according to the (SAS Institute, 2005), provided two ways that missing observation in a given data can

take. First, missing data is monotone which occurs when there is a clear pattern showing within and among the missing data values. With this approach it could be early reorder the individual values to get a well-ordered dataset. The second approach is when the missing data values are indeed missing arbitrarily and there is no way to reorganize or reorder the individual values to achieve a certain kind of pattern within the dataset. The table below is the diagrammatic representation of the two main types of the missing observation pattern.

Table 3.1: Monotonic Missing Observation

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	√	√	√	√	√
2	√	√	√	√	–
3	√	√	√	–	–
4	√	√	–	–	–
5	√	–	–	–	–
6	–	–	–	–	–

Table 3.2: Missing Arbitrarily Observation

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	–	–	√	√	–
2	√	√	–	√	–
3	–	–	√	√	√
4	–	√	–	–	–
5	√	–	–	–	–
6	√	–	√	–	√

3.2 Reasons for missing observations

There are several reasons for the occurrences of missing observation within dataset, these may be due to factors such as the respondents used for the study was not able to complete the information required from him or her in full. It may also be due to the fact that the researcher who was organizing the data collection failed to take

record of some of the information provided by the respondents. It may be due to the fact that respondents withdraw from the studies before it completion, those who do the data entry or transcriptions or coded of the responses could make mistakes in handling the data. All these are some of the causes of missing observations in a given dataset as indicated by Cohen and Cohen (1983). The authors indicated that when there are missing observation or data is on the part of the dependent indicators, then such subject may be remove from the analysis of the data, however they made it known that when it is on the independent variable(s) then it could be more of important to maintain by estimating the percentage of the data that is considered missing before proceeding to conduct the analysis of the data. This means that in some instance the missing values cannot just be removed or replaced with the values but must be properly be examined before handling it.

3.3 Procedures for handling missing data in the dataset

3.3.1 Traditional Procedures

Listwise deletion (or Complete Case Analysis)

This procedure is one of the easiest and simplest approaches that one can adopt, when the data analyst released that there is missing observation within the dataset, this is done by completely or deliberately removing or excluding such an observation from the dataset or from the analysis, in most of the statistical packages the options are there to exclude missing observation from the analysis, this was indicated by (Briggs et al.,2003). One advantage about it is the convenient using such method or approach since it does not required any scientific calculations. It does not take much time of the analyst and does not involve much thinking as to how the missing observation or data could be replaced.

The approach has got some level of limitation especially when there are a lot of such missing observation within the data set. For example, suppose that a researcher is interested in measuring factors that influences organizational performance and decided to engage 1000 employers and employees of some selected business establishment with 20 variables understudy. Suppose that in each variable there is possibility about say 5% missing in each case, then the resultant variables without missing observation would be far lower than the variables with the missing observation.

It means in such a case there will be approximately 360 individuals who would have no issue of missing information and about 640 individuals or respondents having issue with missing data, which is problematic and a worrying situation. In such situation, it would have been appropriate when the data is missing completely at random (MCAR) (Nakai & Weiming, 2011).

Imputation Technique or Procedure

This is mostly done when a given dataset have missing observation in it. It is done by finding a substitute or a reasonable value to fill in the missing observation before carrying on with the analysis. In most cases certain assumptions are made before carrying on with such replacement, some consider normality of the dataset and others basic assumption underling the gathered dataset. Below are the main imputation techniques that are used when there is missing observation.

- Marginal mean imputation

  This is mostly done by computing the average of the series say A using the non-missing values and using it to impute missing observation within the series of A. In most cases it leads to biases within the dataset especially when the dataset is non-normal and it affects the variances and covariance

- Conditional mean imputation:

This is done or applied especially when one uses complete data set with no missing observation to predict series of variables with missing observation component in the dataset. Generally, the use of the above mention approaches could lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error (Allison, 2001).

3.3.2 Advanced or Mordern Methods

There are several methods that are more advance in handling the missing observations in a given dataset. Among the few one's are the multiple imputation under the normal model.

The EM - algorithm

The EM algorithm is capable of computing or estimating the means and covariance matrix which is used to drive consistency and accurate parameters of interest. The method is based on the expectation step and the maximization approach, which is done by continuously repeating the processes many times till maximum likelihood result or estimates is derived. The method requires the use of large dataset and that the missing observation in the series is also random (MAR), this algorithm could be performed with the use of the SAS (Graham,
2009).

The method is scientific technique that drive it result after several iterations to produce maximum likelihood results. At the E-step, the iteration, cases are read in, one after the one. When a value is inputted, the algorithm calculates the sum, sum of squares and the cross-products of the values, however when there is a missing value in the dataset, the algorithm is capable of suggesting a best guess value to replace the missing value, and this best value to be replaced would be based on the regression approach of the imputation procedures as described before. When there

is one value missing in the dataset, then the quantity is incremented and when there are two values missing then there is quantity incremented and there is an addition or introduction of a correction factor or component.

In the M-step of the algorithm, there is the computation of the variance, covariances and the means of the series. these calculations are based on the sums, sum of square and the sums of the cross products .the use of the covariance matrix , there is new regression model or equation derived and they are used to update the best values for the missing values during the E-step of the iteration processes. When the iteration converged then the covariance matrix stops.

Good uses of the EM algorithm

The EM algorithm is capable of estimating excellent parameter for a given dataset, however there is lack of standard errors associated with its estimation, which makes it a bad estimator for hypotheses testing. Many analyses do not require the usage of the standard errors in its computing, which means that the EM algorithm is very useful for data analysis.

The technique is capable of estimating the means, the standard deviation and sometimes it is able to compute the correlation matrix. All these are some of the estimates that the EM algorithm is able to compute and also the technique is useful when it comes to the calculations or estimating of coefficient alpha which does not require the estimating of the standard errors (Graham et al,2003). The EM algorithm is the basis for many exploratory techniques such as the factor analysis with missing values in the dataset. This could easily be done using the SAS or the SPSS software.

Multiple Imputation

Multiple imputation (MI) is a two-stage approach where missing values are imputed a number of times using a statistical model based on the available data and then inference is combined across the completed datasets. This approach is becoming increasingly popular for handling missing data. Lee and Simpson (2014) and Wiley

(2002) affirmed that missing data occur frequently in survey and longitudinal research. Incomplete data are problematic, particularly in the presence of substantial absent information or systematic nonresponse patterns.

Listwise deletion and mean imputation are the most common techniques to reconcile missing data. However, more recent techniques like multiple imputation (MI) may improve parameter estimates, standard errors, and test statistics. Recent theoretical and computational advances, most notably multiple imputation (MI) methods, enable the researcher to use the existing data to generate, or impute, values approximating the real value, while preserving the uncertainty of the missing values (Schafer, 1997).

The method is used to solve some of the challenges the conventional methods face, by introducing an additional form of error based on the variation in the parameter estimates across the imputations. It is capable of replacing the missing observations with two or more values representing a distribution of possibilities (Allison, 2001). According to Schafer (1997), the multiple imputation technique is used to handle missing observation with the aim of achieving valid and reliable statistical inferences rather than the just the recreate the missing observations or replacing the missing values with the closest values as in the case of the traditional methods.

Maximum Likelihood

The tool is considered to be one of the best modern missing data analysis tools that most researchers used or have recommended for the estimating of the missing values with a given dataset. The technique assumes multivariate normality assumption, it works by factoring all the data points either missing or nonmissing to be able to compute the estimate for the parameters required. Due to its complexity, it is difficulty to compute it manual, however software has been developed to enhance its estimation.

The calculations use what is known as the loglikelihood function to find the standardized distance between the data points and the parameters of interest such as the mean or the other parameters of interest. This is one of the technique that could be used in computing the variance-covariance matrix of a given variables based on the dataset available. The obtained covariance matrix is then used to estimates or compute the regression model (Schafer, 1997). When it comes to simplicity the maximum likelihood is simpler to use as it does not require much decision in computing it estimates such as the number of dataset to use, the number of iterations to be performed, the distribution to use as compared with the multiple imputation. It only requires the data analyst person to just state the model to be formulated by indicating the ML (SAS, 2005)

3.4 Mechanisms of Missing Data

In conducting a research, if missing value problems are encountered, the first thing that ought to be done is to examine the natural pattern of the missing values. According to Little and Rubin (1987), missing patterns generally fall into one of three types, i) Missing Completely at Random, ii) Missing at Random and iii) Missing Not at Random. To identify which category the data fall into, the percent of missing values could be forecasted for each demographic category.

Any treatment of missingness must start out with the query of why it occurred in the foremost. These missingness could occur for simple and innocuous reasons, such as a group of people having an automobile accident and not being able to appear for testing. Missing is more a nuisance than a problem to overcome in such a scenario. Data could, however, be missing on the basis of either the potential score of the participant on the dependent variable (Y) or any of the independent variables (XI). The reasons for lack of data play an important role in the handling of these data.

3.4.1 Missing Completely at Random (MCAR)

Rubin (1976) defined a clear taxonomy of missing data, which became the standard for any discussion on this subject. This taxonomy depends on why data is missing. If for any of the variables the fact that data is missing does not depend on any values or potential values, the data is said to be completely random (MCAR). The case of the careless motorist, who does not appear for testing due to an accident, who has not done anything with the study, is a case in point.

Pickles (2005) expressed the condition somewhat differently by saying that the probability of missing is a constant for MCAR. Any observation of a variable is equally likely to be missing. If you have missing data, this is the ideal case because the treatment of the existing data does not result in partiality in the estimated parameters. It can contribute to a power loss that is often not a serious problem in census work, although it can certainly be found in experimental studies, but it does not contribute to partial parameter estimates.

Little (1998) provided a statistical examination of the assumption of MCAR. His MCAR test is a square measurement test. A substantial value shows that data are not MCAR. This test is offered in the SPSS Missing Values Analysis (MVA), which is not part of the basic system, and should be performed whenever MCAR is in doubt. SAS also includes this PROC MI test.

3.4.2 Missing at Random (MAR)

Data is missing at random (MAR) if the likelihood of missing data on a variable (Y) does not depend on its own value after checking for other design variables. Allison (2001) is an example of' missing' in income data that depends on marital status. Unmarried couples may have less chance of reporting income than married couples.

Unmarried couples probably have lower incomes than married couples, and at first it would appear that income shortages are linked to the value of income. However, the

data would still be MAR if the conditional likelihood of missing were unrelated to the value of income in each marital class. The real question here is whether the value of the dependent variable determines the likelihood of reporting or whether there is another variable (X) where the likelihood of missing Y depends on the levels of X. In other words, data is MAR if p(Y missing , X)= p(Y missing).

3.4.3 Missing Not at Random (MNAR)

The data shall be classified as missing not at random(MNAR) if either of the two above classifications is not met. Therefore, if the data is not at least MAR, it is not missing at random. When MNAR data is presumably a model that lies behind failure. If we knew this model, we could obtain appropriate estimators of the parameters in the model that underlies our information. For example, if people with low incomes are actually more reluctant to report their incomes than people with higher incomes, we can probably write an equation (a model) that predicts income-based missingness. Sadly, we seldom know what the missing model is, so it's hard to know how to proceed. Furthermore, incorporating a missing model is often a very difficult task and can be specific to each application.

See Dunning and Freedman 's (2008) article for a useful example of a 5missingness model. Also note Dunning and Freeman's interesting example of a situation where data on the independent variable is missing due to their score. This example shows that such data can seriously distort the correlation between the two variables, but can have little effect on the coefficient of regression.

This study focuses on Missing at random (MAR) assumption. With this assumption, the missing data mechanism is said to be ignorable, which essentially implies that there is no need to model the missing data mechanism as part of the estimation process.

3.5 Data

A secondary data was employed for the research. This data was obtained from the National greenhouse gas inventory of anthropogenic emissions by sources and removals by sinks of all greenhouse gases not controlled by the Montreal Protocol and greenhouse gas precursors. Data on CO_2 (Gg) and CO emissions under the Energy and Manufacturing Industries and Construction sub-section were selected for the research. It is a time series data taken and collated by the Environmental Protection Agency annually from 1990 to 2012.

3.5.1 Description of the data

The National Greenhouse Inventory was carried out in accordance with the methodologies contained in the IPCC guidelines for 2006. The guidelines provide step by step guidance on how to consistently apply the methodology and the underlying assumption. The guidelines also provided guidance on how to estimate GHG emissions with AD and EF and documentation, archiving and reporting.

Activity data refers to the measurement of the intensity and/or frequency of use and/or the number of specific activities leading to the generation of emissions at various stages of use or production. Similarly, the emission factor is a measure of the rate at which the level, intensity, frequency of use or production leads to specific emissions of GHG under certain conditions. The product of the activity data and the emission factor therefore indicates the total GHG emissions of a specific activity.

$$E(Gg) = AD * EF$$

Where;

E = Emissions

AD = Activity Data

EF = Emission Factor

3.6 Multiple Imputation Method

Multiple imputation offers a useful strategy to deal with missing values in data sets. Rubin ’s (1987) multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to be imputed instead of filling in a single value for each missing value. These multiple imputed data sets are then analyzed using standard comprehensive data procedures and the results of these analyses are combined. The process of combining results from different imputed data sets is essentially the same regardless of which complete data analysis is used. This leads to valid statistical conclusions that correctly reflect the uncertainty due to missing values. Yang. Yang, (2001).

According to Sori-Bori (2013), the multiple imputation method mainly involves 3 phases namely; imputation phase, data analyses phase and results pooling phase as expanded below:

1. Run an imputation model to create imputed data sets defined by the selected variables. In other words, the missing values are filled to generate m complete data sets in m times. m=20 is considered sufficiently good. Diverse algorithms for the imputation phase have been proposed. Notable examples are the E-M algorithm and data augmentation procedures etc. In this study, the EM algorithm approach will be adopted.
2. The m complete data sets are analyzed by using standard procedures
3. Parameter estimates are combined from each imputed data set to obtain a final set of parameter estimates.

3.6.1 Expectation Maximization Algorithm (EM - Algorithm)

EM is a numerical algorithm for maximizing the likelihood of missing data models (Dempster et al., 1977). It is an iterative algorithm which cycles repeatedly in two

steps. In the expectation step, the expected log-likelihood value is taken over the variables with missing data by calculating the expectations using the current values of the parameter estimates. The expected log-like probability is maximized in the maximization step to obtain new parameter estimates. These two steps are repeated time after time until they converge.

Mathematically, the EM algorithm is described as;

Let, D be a complete set of data

D^{miss} represent missing data from the complete set of data D^{obs} represent observed data from the complete dataset, and α represent the parameter to be estimated.

$$D = \{D^{miss}\} \tag{3.1}$$

In any incomplete-data problem, the distribution of the complete data *D* can be factored as

$$P(D|\alpha) = P(D^{obs}|\alpha)P(D^{miss}|D^{obs},\alpha) \tag{3.2}$$

Viewing each term in (3.2) as a function of α, it follows that

$$l(\alpha|D) = l(\alpha|D^{obs}) + \log P(D^{miss}|D^{obs},\alpha) + c \tag{3.3}$$

where $l(\alpha|D) = \log P(D|\alpha)$ denotes the complete data loglikelihood, $l(\alpha|D^{obs}) = \log P(D|\alpha)$, the observed data loglikelihood and, c, an arbitrary constant.

Since D^{miss} is unknown, the second term on the right side of (3.3) can not be calculated, so instead we take the average of (3.3) over the predictive distribution $P(D^{miss}|D^{obs},\alpha)$, where $\alpha^{(t)}$ is a preliminary estimate of the unknown parameter. This averaging yields,

$$Q(\alpha|\alpha^{(t)}) = E_{\alpha^{(t)}}\left[l(\alpha; D)|D^{obs}\right] \tag{3.4}$$

where

$$Q(\alpha|\alpha^{(t)}) = \int l(\alpha|D)P(D_{miss}|D_{obs},\alpha_{(t)})dD_{miss} \tag{3.5}$$

A central result of Dempster, Laird, and rubin (1977) is that if $\alpha(t+1)$ is the value of α that miximizes $Q(\alpha\ \alpha^{(t)})$ then $\alpha^{(t+1)}$ is a better estimate than $\alpha^{(t)}$ in the sense that the data loglikelihood observed is at least as high as $\alpha(t)$,

$$l(\alpha_{(t+1)}|D_{obs}) \geq l(\alpha_{(t)}|D_{obs}) \tag{3.6}$$

Thus the EM algorithm consists of an E-step (Expectation step) followed by MStep (Maximization step) defined as follows:

E-step: Compute $Q(\alpha;\alpha^{(t)})$ where

$$Q(\alpha; \alpha^{(t)}) = E_{\alpha^{(t)}}\left[l(\alpha; D)|D^{obs}\right] \tag{3.7}$$

M-ste: Find $\alpha^{(t+1)}$ in α such that

$$Q(\alpha_{(t+1)};\alpha_{(t)}) \geq Q(\alpha|\alpha_{(t)}) \tag{3.8}$$

for all $\alpha \in \theta$

3.6.2 Linear Interpolation

In linear interpolation, a straight line is connected to two data points with a straight line and hence the interpolation function is given by;

$$g(x) = C_0 + C_1(x - x_0) \tag{3.9}$$

where x is the independent variable, $x_i(i = 0,1,2,\ldots)$ is a known value (i.e. of the independent variable), and C_i are unknown coefficients. Then from equation 3.1;

$$C_0 = g(x_0) \tag{3.10}$$

$$x = \frac{g(x) - g(x_1)}{x_1 - x_0} \tag{3.11}$$

In whic case $g = g(1)$. If three data points are available, interpolation is carried out using a quadratic polynomial. A convenient form for this estimation is

$$g_2(x) = C_0 + C_1(x - x_0) + C_2(x - x_0)(x - x_1) \tag{3.12}$$

3.7 Performance Indicators

To assess the Multiple Imputation (MI) and the Linear Interpolation (LI) methods, two performance indicators namely; mean absolute error, and root mean square error, were employed. The theoretical and observed data were compared with the best way to estimate missing values. The absolute mean error (MAE) is the average difference between the forecast and the actual data values.;

$$MAE = 1/n \sum_{n=1}^{n} |Pi - Oi|^2 \tag{3.13}$$

Where n is the number of imputations, Oi and Pi are the data points observed and imputed. MAE ranges from zero (0) to infinity and the MAE = 0 is perfectly fitted.

The mean square erro is one of the most commonly used numerical prediction measurements. Its value is calculated using;

$$RMSE = \sqrt{(1/n \sum_{n=1}^{n} |Pi - Oi|^2)} \tag{3.14}$$

The smaller the RMSE value, the better the performance of the mode.

CHAPTER 4

DATA ANALYSIS AND SUMMARY OF RESULTS

This chapter presents the result obtained from the analysis of the data gathered. Data was obtained from the Environmental Protection Agency (EPA, 2015). The carbon dioxide and carbon monoxide emissions data obtained had presence of missing observations. The researcher therefore used advanced missing observations technique to impute these missing values; twenty (20) different imputed datasets were generated and then combined together before statistical techniques were used on the completed dataset. Linear interpolation and multiple imputation techniques were used to replace the missing values in the dataset.

4.1 Data Description

A time series data on carbon dioxide (CO_2) and carbon monoxide (CO) emissions from 1990 to 2012 were obtained from EPA. These emissions data on these gases are collected as part of the IPCC membership responsibility as a country. The incomplete dataset includes four (4) variables: years of emissions of CO_2 and CO, source of emissions either manufacturing or energy sectors and the CO2 and CO emissions with missingness. Table 4.1 shows the summary of the incomplete dataset. The CO_2 and CO in the dataset had 16 missing values out of 46 observations representing 34.90% and 3 out of 46 observations representing 6.6% respectively. The estimated average of CO2 is 779.3 million tonnes (Mt) and that of CO is 3.63 $MtC0_2e$

Table 4.1: Summary Statistics of CO_2 and CO emissions

Incomplete Data Set

	CO_2	CO
Mean	779.30	3.63
Standard Error	110.29	0.23
Median	875.5	3.30
Min	5.00	1.06
Max	2012.00	6.36
NA	16 (34.90 %)	3 (6.6 %)

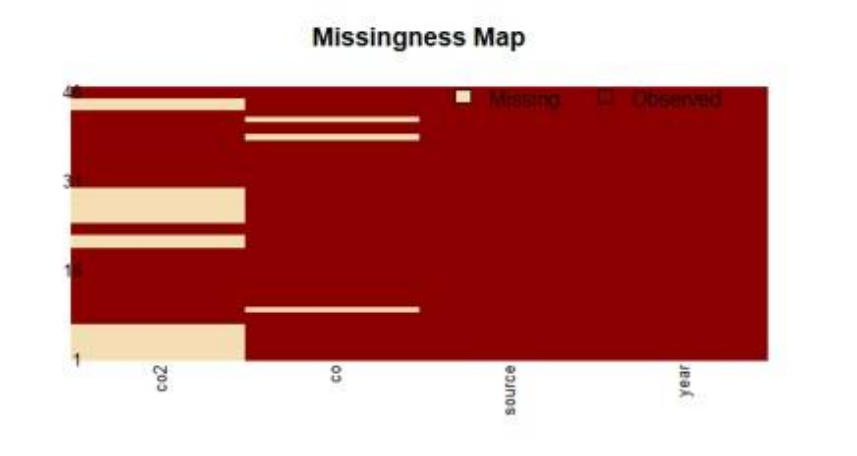

Figure 4.1: Missingness map of the variables

From Figure 4.1, it is observed that in the early 1990 to 1995 there were missing information on CO_2 emissions. Again, it was observed from 2007 to 2011 that, there were missing information on CO2 emissions. In the same figure, it was observed that, CO emissions' missingness occurred mostly between 1992 and 2000.

4.2 Imputation Analysis on CO_2 and CO

In performing multiple imputation process, the CO_2 and CO time series data were included in the analysis. This is to improve predictive power of the imputation method. A total of twenty(20) imputation datasets were generated for both CO_2 and CO emissions. Figure 4.2 shows the histogram of the CO2 emission of the 5th and the 15th imputations datasets. Since the multiple imputation method rely on multivariate normality assumption it is a good idea to look at the distribution of the imputed datasets.

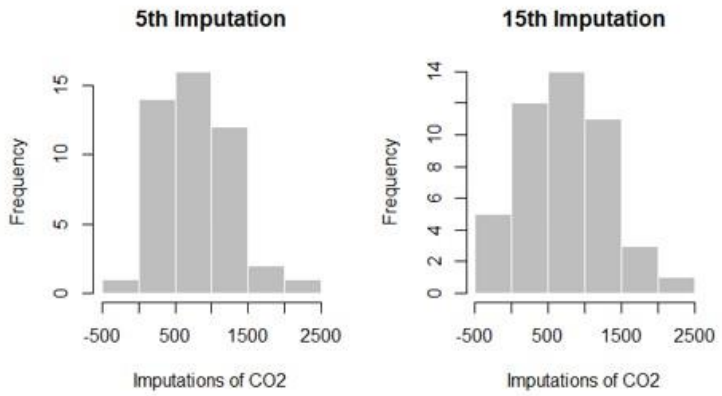

Figure 4.2: Histrogram of the CO_2 emission from the 5^{th} and 15^{th} imputed datasets.

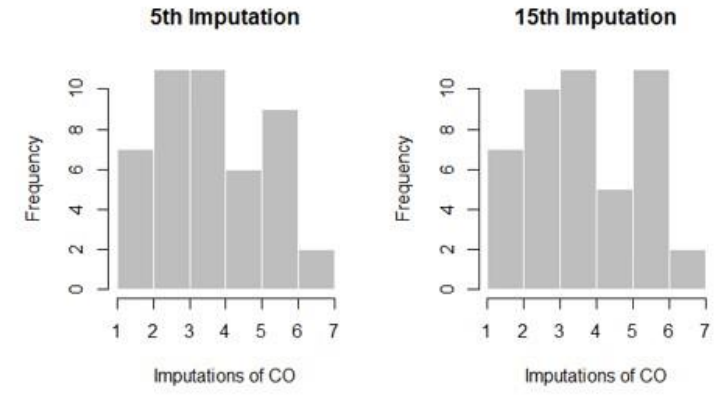

Figure 4.3: Histogram of the CO emission from the 5^{th} and $a5^{th}$ imputed datasets

The CO_2 and CO emissions were recorded annually, meaning in principle they are time series data. Knowing the observed values of observations close in time to any missing value may aid the imputation method. In the following results, it is assumed that both CO_2 and CO vary over linear time even though test for stationarity wasn't needed for the imputations. It is observed in Figure 4.4 that a much better prediction about the missing values of CO_2 emissions from manufacturing sector when incorporating linear time than when it is omitted. This conclusion is based on the 95% bands of the imputed values. The left panel in Figure 4.4 shows longer bands compared to the right panel in general.

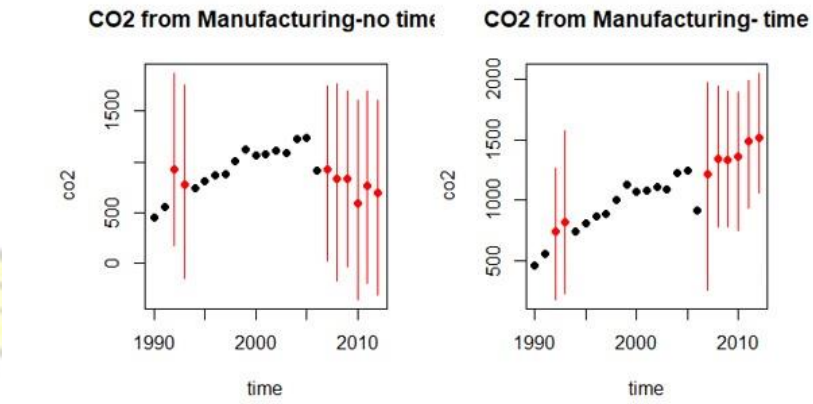

Figure 4.4: The increase in predictive power of CO_2 emissions from the manufacturing sector using linear time. The panel shows mean imputations with 95% bands in red

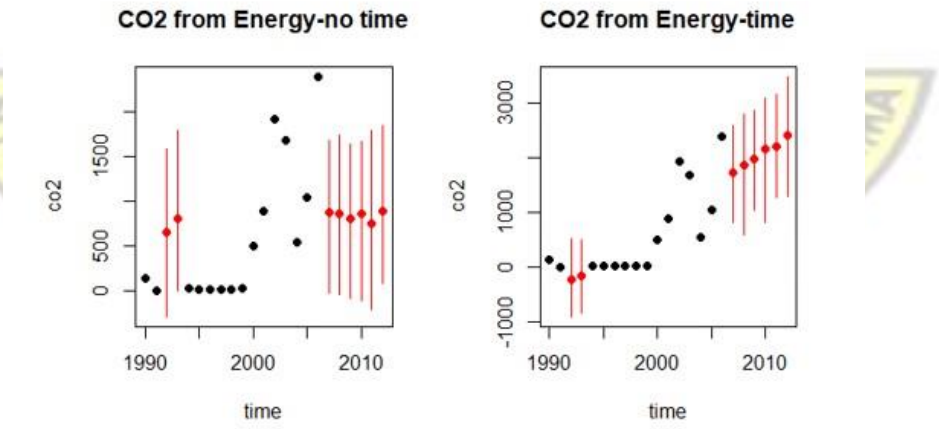

Figure 4.5: The increase in predictive power of CO_2 emissions from the Energy sector using linear time. The panel shows mean imputations with 95% bands in red.

It can be observed in Figure 4.5 that a much better prediction about the missing values of CO_2 emissions from energy sector when incorporating linear time than when it is omitted.

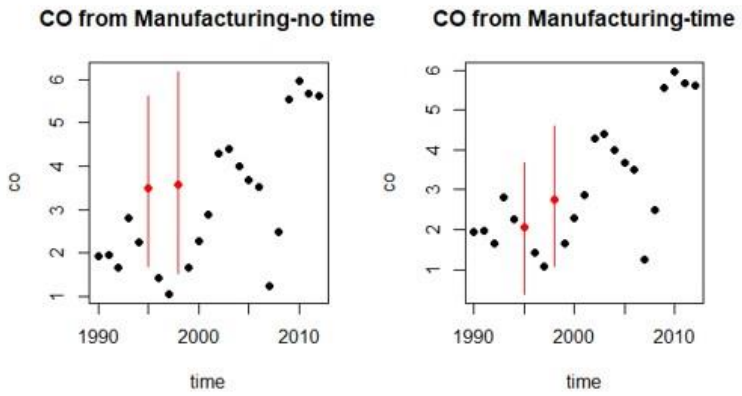

Figure 4.6: Predictive power of CO from manufacturing

It can be observed in Figure 4.6 that similar results in terms of prediction about the missing values of CO emissions from energy sector when the incomporation of linear time was compared with when it was omitted. This means that linear time do not have any effect on the predictive power of CO.

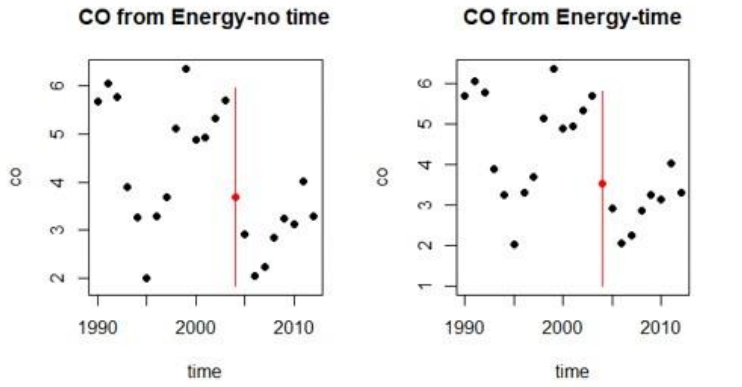

Figure 4.7: Predictive power of CO from Energy sector

Similarly, it can be observed in Figure 4.7 that similar results in terms of prediction about the missing values of CO emissions from energy sector when the incorporation

of linear time was compared with when it was omitted. This this confirms that linear time do not have any effect on the predictive power of CO.

4.3 DIAGNOSTICS

In other to inspect the imputations that are created, two diagnostic tools of graphical approach were used for this purpose.

4.3.1 Comparing Densities

The plots in Figure 4.8 shows the relative frequencies of the observed data with an overlay of the relative frequency of the imputed values of each CO_2 and CO emissions. We observe from Figure 4.8 that the imputed CO_2 emissions are quite similar to the observed CO_2 emissions but the imputation of the CO are quite different. This is possible because of the small number of missingness in the CO emission data.

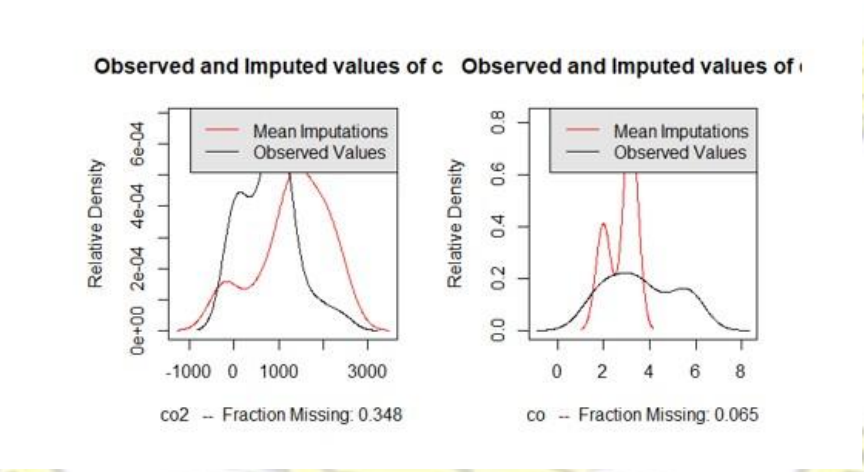

Figure 4.8: Observed and imputed values

4.3.2 Over - imputation

This is a technique implored to judge the fit of the imputation model. It is assumed that each observed value is treated as if missing. From Figure 4.9, for each observation the line indicates a perfect agreement if the imputed value is the same as the observed value. A 95% confidence interval are constructed where an observed

value would have been imputed had it been missing from the CO_2 emissions data. It can be observed that the line cuts almost all the confidence intervals, this is an indication that the imputed model is not over imputing.

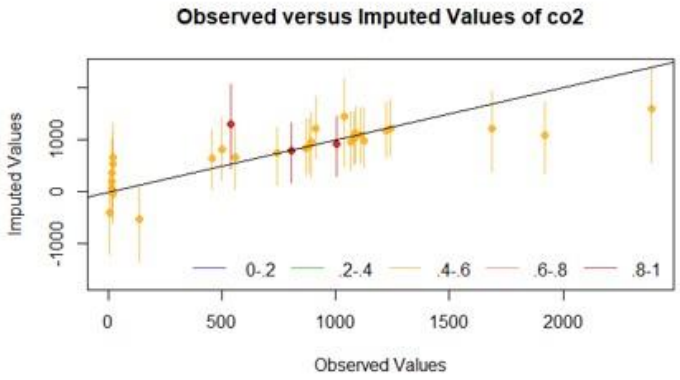

Figure 4.9: Observed versus imputed values of CO_2

4.3.3 Forecast for CO_2 for the next ten (10) years.

Measures of accuracy was used in the determination of the best model that fit the carbon-dioxide data as indicated in the table 4.4 below. From the results, the accuracy measures that has higher number of minimum estimates is considered to be appropriate to fit the data. The results show that, quadratic model has the maximum of the least values hence the best model to fit the CO2. for the manufacturing and energy industry data with an estimated equation of; $470.6 + 57.9t - [0.726t]^2$ and $-205 + 52.7t + [2.42t]^2$ respectively. The results in the table 4.5 below are the forecast values for the next 10years.

The figures below show the time plot analysis of the CO_2 data for both manufacturing and energy sector. From the figure 4.8, there is an indication that, in the case of the manufacturing sector, there is a gradual increase in the CO_2 in the initial stages and then decreases sharply between point 15 and 18, it takes off right after the 18 points as indicated on the x-axis and then gradually increases. Also, the Figure 4.9 provides information on the nature of the data for the energy sector. From the figure, some

YEAR	$CO_2(M)$	$CO_2(E)$		
1991	556	5	1.964	6.062
1992	656.99	-230.79	1.664	5.768
1993	773.88	-40.917	2.810	3.900
1994	743	22	2.259	3.255
1995	807	15	2.052	2.013
1996	867	15	1.423	2.300
1997	884	15	1.060	3.700
1998	1005	18	2.459	5.120
1999	1123	21	1.664	6.358
2000	1066	502	2.285	4.872
2001	1082	890	2.878	4.927
2002	1112	1919	4.280	5.320
2003	1088	1685	4.390	5.691
2004	1224	538	3.996	3.501
2005	1242	1039	3.677	2.918
2006	911	2392	3.514	2.048
2007	1213.53	1464.45	1.2456	2.235
2008	1193.800	1687.21	2.478	2.844
2009	1313.54	1775.37	5.543	3.245
2010	1488.33	1871.59	5.962	3.137
2011	1407.89	1857.39	5.615	3.295

Table 4.2: Imputed model for the emission dataset considering time effect

			CO (M)	CO (E)
1990	457	135	1.937	5.678

initial fluctuations and then maintains its steadiness before point 12. It then resumes massive fluctuations with an

Table 4.3: Summary Statistics for all the observations as indicated in Table 4.2

Variable	Mean	Median	Std. Dev	C.V	Skewness	Ex. Kurtosis
Multiple Imputation (M)	1029.00	1082.00	293.22	0.28	0.21	-0.65
Multiple Imputation (E)	858.95	538.00	892.55	1.04	0.29	-1.55
Linear Interpolation (M)	1008.80	1066.00	281.50	0.28	-0.12	-0.51
Linear Interpolation (E)	996.59	538.00	1023.80	1.03	0.31	-1.69

Table 4.4: Measure of accuracy in determination of the best model fit for $CO_2(M)$

Model	MAPE	MAD	MSD
Linear	8.5	76.6	10127.3
Quadratic	6.9	66.33	9315.11
Exponential growth	10.7	97	13587.7
S - curve	6.5	70.4	11850.2

Table 4.5: Forecast values for the model in the next ten years (10)

	$CO_2(E)$	$CO_2(M)$
YEAR	FORECAST (E)	FORECAST (M)
2013	2453.52	1442.21
2014	2624.74	1464.54
2015	2800.80	1485.41
2016	2981.69	1504.84
2017	3167.42	1522.81
2018	3357.99	1539.32
2019	3553.40	1554.39
2020	3753.64	1568.00
2021	3958.72	1580.16
2022	4168.64	1590.87

overall increase up to point 18. This proceeds with a gradual increase up to point 24.

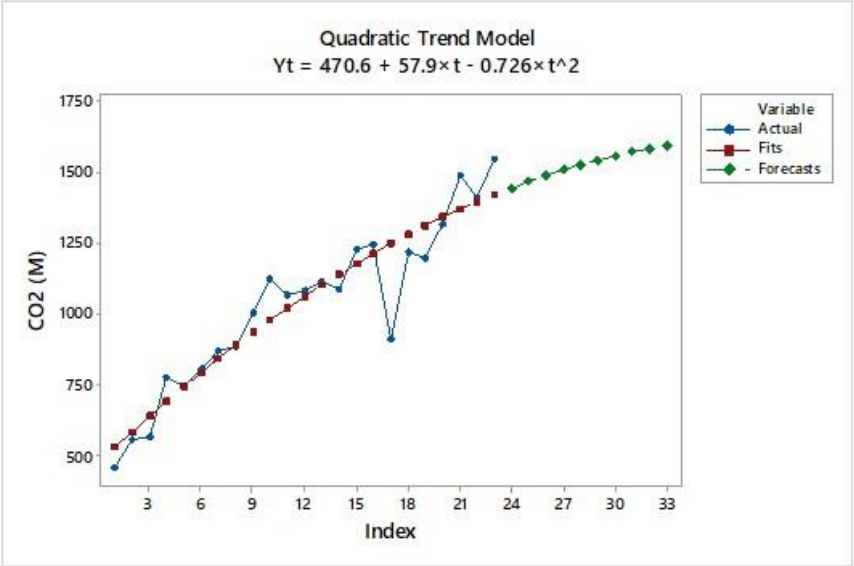

Figure 4.10: Trend Analysis plot of $CO_2(M)$

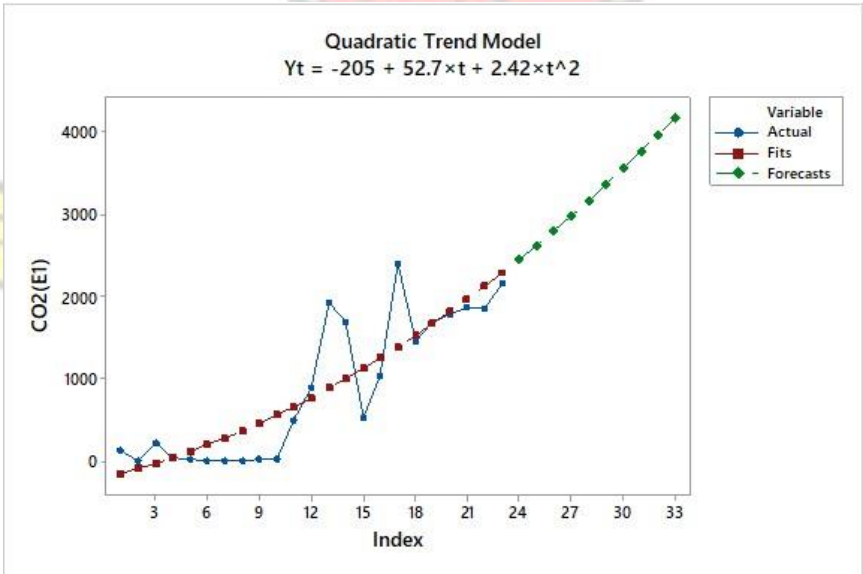

Figure 4.11: Trend Analysis plot for $CO_2(E)$

4.4 Imputation Performance Analysis

In this study, the performance of the various missing data imputation methods were compared under two (2) categories of data namely; the CO_2 emission data under i) the Energy Industries and ii) Manufacturing and Construction Table 4.6: Performance analysis using RMSE and MAE

	RMSE	MAE
Multiple Interpolation (E)	495.0757	298.1944
Linear Interpolation (E)	510.8089	336. 0303
Multiple Imputation (M)	96.298	63.835
Linear Interpolation (M)	96.551	58.685

Industries. For the Energy Industries data (E), it was clear from table 4.6 that, the Multiple Imputation method outperformed the Linear Interpolation technique. As indicated in the previous chapter, a lower RMSE means a more accurate a prediction to the actual value. With regards to the MAE, zero (0) means a perfect prediction. The RMSE and MAE values of MI (E) were; 495.0757 and 298.1944 respectively as against 510.8089 and 366.0303 respectively under LI (E). However, for the Manufacturing and Construction Industries data (M), a mixed results ensued. Whilst MI continued to outperform LI under RMSE, the LI did better as compared with MI under the MAE as shown on Table 4.6. In general, the MI was robust against LI as it had better prediction of accuracy when the two performance indicators were applied.

4.4.1 Summary Statistics for only the replaced observation

The results in Table 4.7 below shows the summary statistics for the missing observations in the dataset. It was observed that, the estimated mean value and standard deviation for the multiple imputation under the manufacturing and construction industries (M) is 1137.70 ± 340.34 and the linear interpolation having an estimated value of 1068 ± 318. This clearly shows that, the mean of the linear interpolation is quite smaller as compared with multiple imputation method. Also, the energy industries data (E) has an estimated mean and a standard deviation of 1197.80 ± 922.73 for the multiple imputation method as compared Table 4.7: Summary Statistics for only missing observation

Variable	Mean	Median	Min	Max	Std. Dev	C.V	Skewness	Ex. Kurtosis
MI(M)	1136.70	1213.50	565.99	1488.30	340.34	0.30	-0.75	-0.90

MI(E)	1197.8	1687.20	-230.90	1871.60	922.73	0.77	-0.89	-1.10
$LI\_(E)$	1644.2	2264.70	10.67	2360. 20	1114.70	0.68	-0.94	-1.10
LI_M	1068.10	1127.30	618.33	1451.80	318.72	0.30	-0.37	-1.27

with that of linear interpolation (1644.20 ± 1114.70). This is an indication that, the multiple imputation technique has the least error or deviation associated with its estimates as compared with the conventional method.

CHAPTER 5

SUMMARY FINDINGS, CONCLUSIONS AND RECOMMENDATION

This chapter presents the final chapter of this study. It seeks to provide a summary of findings emerging from the preceding chapter. The findings are summarized under the study objectives. Appropriate recommendations were made based on the findings. This is to assist researchers have a wider scope of models to adopt, when selecting appropriate model to solving issues of missingness. The content of this chapter comprises of; summary of findings, suggested recommendations and a final conclusion for the research.

5.1 Summary of Findings

From the results obtained in the previous chapter, out of the total of 46 observations, carbon-dioxide (CO_2) and carbon-monoxide (CO) in the dataset had 16 (34.90%) and 3 (6.6%) missing values respectively. The estimated average of CO2 is 779.3 million tonnes (Mt) and that of CO is 3.63 $MtC0_{2e}$. Normality test was performed on the CO_2 and CO and it was realized that in each case, they conformed to the normal distribution assumption as shown in Figure 4.2 and 4.3 respectively.

The diagnostic test using graphical and over impute approaches were performed on the generated dataset and the results showed that, the imputed CO_2 emissions are quite similar to the observed CO_2 emissions whilst the imputation of the CO are quite different. This is possible because of the small number of missingness in the CO emission data. In the case of the over-impute approach, it can be observed that the line cuts almost all the confidence intervals, this is an indication that the imputed model is not over imputing.

In Table 4.2, a complete set of data (observed and the imputed values) was presented, all the missing values have been generated and replaced. After the missing values were replaced, an appropriate statistical model, was used to estimate the future values for CO_2. This model was selected based on measures of accuracy as indicated in Table 4.3.

To assess the Multiple Imputation (MI) and the Linear Interpolation (LI) methods, two performance indicators namely; mean absolute error, and root mean square error, were employed. The theoretical and observed data were compared to select the best method for estimating missing values. These performance tests, confirmed in general that the MI as a more robust model to dealing with missingness in time series data as shown in Table 4.6.

5.2 Conclusions

The study was premised on three main objectives. The first objective which sought to estimate the missingness of Manufacturing industries and construction data under the Fuel combustion sub-category of the Energy sector using multiple imputation was duly achieved.

The second objective also sought to estimate the missingness of Energy industries data under the Fuel combustion sub-category of the Energy sector using multiple imputation method. After applying the adopted techniques, the estimated CO_2 values for the missingness in the Energy sector dataset, were generated as indicated in Table 4.2 of the preceding chapter. Hence the objective has been achieved.

The final objective discusses the comparison of Multiple Imputation and Linear Interpolation methods to estimate missing values. This study is carried out to prove that, usage of the Multiple Imputation technique will greatly enhance the statistical performance of the data. The yearly CO_2 emissions data under the Energy Industries and Manufacturing and Construction Industries sub-sectors from 1990 to 2012 was

used to compare the performance of the methods. The simulated missing values were used. Two performance indicators were calculated in order to select the best method of replacing missing values. They are the mean absolute error (MAE) and the root mean square error (RMSE). From these performance indicators, the best method was found to be the Multiple Imputation.

5.3 Recommendation

Based on the results and conclusions made, the study recommends the following;

First, the multiple imputation technique provides a more accurate results as compared with the traditional method of replacing missing observations in a dataset.

Second, further studies is recommended to be done to determine the required sample size that, enhances the performance of the multiple imputation technique.

Third, again further studies is recommended to be done to affirm the robustness of the MI using additional measures of accuracy.

Last, the technique must be applied in other field of study, to determine the technique effectiveness.