ADDIS ABABA UNIVERSITY
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Data Center Energy Inefficiency Root Cause and Sensitivity Analysis: The case of Ethio-Telecom Legehar Data Center

By: Zerihun Tesfaye

Advisor: Dr. –Ing. Dereje Hailemariam

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master Science in Telecommunications Engineering.

Addis Ababa, Ethiopia
November, 2021

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Data Center Energy Inefficiency Root Cause and Sensitivity Analysis: The case of Ethio-Telecom Legehar Data Center

Submitted by: Zerihun Tesfaye _____

Signature

Dr. –Ing. Dereje Hailemariam _____

Advisor Signature

_____ _____

Evaluator Signature

_____ _____

Evaluator Signature

Declaration

I declare that this thesis work "*Data Center Energy Inefficiency Root Cause and Sensitivity Analysis: The case of Ethio-Telecom Legehar Data Center"* is my original work, and has not been submitted for a degree at this or any other university, and all sources of materials used for the thesis have been fully acknowledged.

Zerihun Tesfaye _____

Student ID – GSR 5562/11 Signature

Place – Addis Ababa, Ethiopia

Date of Submission: _____

This thesis has been submitted for examination with my approval as a university advisor.

Dr.-Ing. Dereje Hailemariam _____

Advisor Name Signature

Dedication

I dedicate my study to my grandmother, father, and mother, may God's mercy be upon them, for their memories have given me the strength to continue.

Acknowledgment

First and foremost, I would like to praise and thank God, the Almighty, and the Virgin Mary, who have granted me countless blessings, knowledge, and opportunities so that I have finally been able to accomplish the thesis.

The success of this thesis depends largely on the support and guidance of many individuals. I would want to take this opportunity to thank everyone who has contributed to the successful completion of my thesis. First, my special and priceless thanks go to my advisor Dr. –Ing. Dereje Hailemariam for his endless support and kind encouragement during the entire course. His openness, guidance, and support taught me what kindness truly is. I'd want to extend my heartfelt gratitude to Eng. Yosef Mekonnen for his valuable, kind support and willingness to provide me the necessary input for this thesis and other courses throughout my MSC program. I would also like to thank Ethio-Telecom for giving me this golden opportunity to pursue my MSc. I would also like to thank the AAiT instructors who have been part of this program, especially Dr. Yalemzewd Negash and Dr. Surafel Lemma. Thank you very much for the important questions that you put forth and for the valuable comments on my thesis. Indeed, your comments were helpful in giving my thesis a good shape. I also want to thank my colleagues (CAAZ O & M and the P & E functional team) for helping me with their comments, discussions, and guidance throughout my thesis. Last but not least, I want to thank my family and express my very profound gratitude to my wife, Mrs.Yenenesh Ahmed. They not only assisted me in completing my MSC, but they have also been there to unconditionally support me. I can never fully explain the love I feel for them with words and will forever be grateful to them for everything they did for me.

Abstract

Data centers are the cornerstone of today's information age. As more people use information technology (IT), the volume of data processed and stored in the data centers rises. As a result of this growth, data center energy consumption has grown for both actual works (data processing and storage) and supporting infrastructure. A data center subsystem is classified as mission-critical and support infrastructure. The mission-critical parts are IT equipment (e.g. servers, routers, switches, and storage systems), whereas support infrastructure parts include mechanical and electrical components such as backup power supplies, Uninterrupted Power Supplies (UPS), Power Distribution Units (PDU), and cooling systems.

The Power Usage Effectiveness (PUE) metric is the global standard for data center energy efficiency, and it is defined as the ratio of total power delivered to the data center to actual IT equipment energy usage. The current overall average PUE value for the Ethio-Telecom Legehar data center is 2.34, indicating a considerable gap between the power supplied to the data center and the actual energy consumption of IT equipment. The PUE was calculated using data collected for 37-week (nine months) period from the Ethio-Telecom power and environmental monitoring system (NetEco).

This study addresses the problem of energy inefficiency by finding the fundamental cause of the problem using a combination of machine learning and Global Sensitivity Analysis (GSA) techniques. First, a machine learning technique was used to identify important features using the Random Forest Regression (RFR). Second, the Sobol-GSA technique is used to quantify the impact level of the selected features on PUE. Sobol-GSA consists of two scenarios: - main effect (first-order) indices, individual variables' contributions with PUE, and total order indices interactions between variables, i.e., the sum of first indices and higher indices. It was discovered that UPS efficiency and cooling systems are major factors in the energy efficiency problem.

Keywords: - Data center, PUE, Sobol-GSA, Random Forest Regression, NetEco.

Table of Contents

List of Figures

List of Tables

Acronyms

AC	Alternating Current
AH	Ampere-Hour
ASHRAE	American Society of Heating, Refrigerating, and Air-Conditioning Engineers
ATS	Automatic Transfer Switch
CCF	Cooling Capacity Factor
CDCs	Corporate Data Centers
DC	Direct Current
DT	Decision Trees
EESS	Energy Eat by Servers and Switches
EESF	Energy Eat and SLA Violation Factor
ET	Ethio-Telecom
EEU	Ethiopia Electric Utility
GSA	Global Sensitivity Analysis
HDD	Hard Disk Drive
HVAC	Heating Ventilation Air Conditioner
IT	Information Technology
IDCs	Internet Data Centers
KW	Kilowatt
KNN	K-Nearest Neighbor
KPI	Key Performance Indicator
LLD	Low Level Design
LSA	Local Sensitivity Analysis

ML	Machine Learning
MAE	Mean absolute error
NetEco	Network Ecosystem
NGN	New Generation Network
NREL	National Renewable Energy Laboratory
OOB	Out-of Bug
OPEX	Operation Expenses
PDU	Power Distribution Unit
P&E	Power And Environment
PF	Power Factor
PUE	Power Utilization Efficiency
RCA	Root Cause Analysis
RFR	Random Forest Regression
RH	Relative Humidity
RMSE	Root Mean Squared Error
Sobol-GSA	Sobol Global Sensitivity analysis
SSD	Solid State Drive
SVM	Support Vector Machine
UPS	Uninterruptable Power System
VA	Volt Ampere
VarImp	Variable Importance
VRLA	Valve Regulated Lead Acid

Chapter 1 - Introduction

1.1Background of the Study

Data center is the cornerstone of today's information age. As more people use Information Technology (IT), more data (e.g. photos, movies, financial transaction data, and so on), is generated leading to a rise in the amount of data processed, stored, and transported. As a result, data centers' energy usage has increased. A data center is a technological facility that houses both IT equipment and the infrastructure that supports it. Data centers are now an essential component for any company that uses IT as a means of communication, regardless of location [1, 2].

The telecommunication industry contains fixed telephone, broadband Internet, and wireless applications. The telecom infrastructure requires the interactions between different reliable systems in data centers, which are divided into two categories: *"mission-critical" and "support infrastructure"*. The Mission-critical part is IT equipment like servers, storage, and network devices that are responsible for data delivery and computing. Depending on the size and capacity of the data center, a facility with many rows of server racks is a common IT equipment configuration. Each row has several racks or cabinets and each rack houses a number of servers[3]. The support infrastructure parts are also sub-divided into electrical and mechanical parts which are devices responsible for cooling and power delivery for a data center, such as electrical systems (e.g., uninterruptible power supply (UPS), power distribution unit (PDU), and mechanical systems (e.g., cooling systems)[4].

In the data center, mains electricity supplied by utility companies such as Ethiopia Electric Utility (EEU), can be quite high voltage, with "medium voltage" of 15KV-33KV and a high voltage above 66KV. EEU provides 15 KV for data centers in Ethio-Telecom. The high voltages are reduced to low levels suitable for electrical devices using a step-down transformer. Because electricity providers cannot guarantee power supply at all times and in all places, data center owners and/or service providers must have a backup diesel generator (or other renewable energy sources) to provide a backup power source for supporting data center equipment. When the power is interrupted, the Automatic Transfer Switch (ATS) will turn ON the backup generators, which will supply the IT equipment. However, from the moment the generator is turned ON to the time it delivers full power, it takes a little while. UPS helps bridge the time gap by acting as a backup

power supply. When power is disconnected from the system for maintenance, the UPS has a manual bypass switch that allows it to continue to run without interruption. The power will eventually be connected to a power distribution unit (PDU), which will then be connected to a series of power supply units on each IT device[3].

Data centers consume a lot of energy, which raises energy costs and has an effect on the environment due to indirect carbon emissions. Some of the reasons for data center energy inefficiency are dominated by support infrastructure equipment and redundant configuration to ensure availability and reliability [5].

Data center service providers have a great desire to reduce operation expense (OPEX) by improving the day-to-day activities and equipment efficiency [3]. However, the challenge is to identify the areas that should be improved because data center operations comprise IT devices, servers that are connected internally and externally via communication equipment to store, transport, and retrieve digital information, as well as power supply and cooling systems. To know the area of improvement in the data centers, it is necessary to analyze the working principle of data centers, particularly power delivery systems. That is an end-to-end analysis of the power delivery chain from the main power (utility) to the power supply unit at the server [6].

The first step in identifying a data center energy efficiency problem is to assess the efficiency level using acceptable Key Performance Indicators (KPI), such as Power Usage Effectiveness (PUE). PUE metric is used in this thesis to assess the data center's total energy consumption performance. Historical data of energy features of overall IT equipment and supporting infrastructure, which include PDU, UPS, cooling systems, and accessory load like energy consumption of lighting, surveillance security, door access, etc., are used for the PUE analysis. PUE is an industry-standard metric defined as the ratio of the total power entering into a data center to the amount of power used by IT equipment [6]. In addition to the above-mentioned, the PUE computation in the research location takes into account the accessory load, which comprises battery and UPS room and main power distribution room air conditioners other than the server room cooling system. The results revealed that the data center under consideration has a total average PUE of 2.34, indicating that it is inefficient in terms of energy usage.

Secondly, feature selection based on Random Forest Regression (RFR) – VarImp algorithm is used to identify significant and non-significant features. This can help us to explain and focus on certain

components of the data center energy efficiency problem. Therefore, the features at the top of the list are insights as to the most significant contributors to the energy efficiency problem in data centers. However, there is a lack of information on the magnitude of features individual contribution and combined interactions of features with PUE. As a result, Sobol-Global Sensitivity Analysis (Sobol-GSA) overcomes this problem by focusing on the essential features that have been selected. The Sobol-GSA analysis consists of two scenarios to see which feature causes the largest changes in PUE. Main effect (first-order) indices and total order indices, individual variables' contributions with PUE varying simultaneously, and interactions between variables, respectively. This thesis major focus is on combining Root Cause Analysis (RCA), and GSA to better understand the data center's energy efficiency captured via PUE. The data was obtained from the Ethio-Telecom Legehar data center.

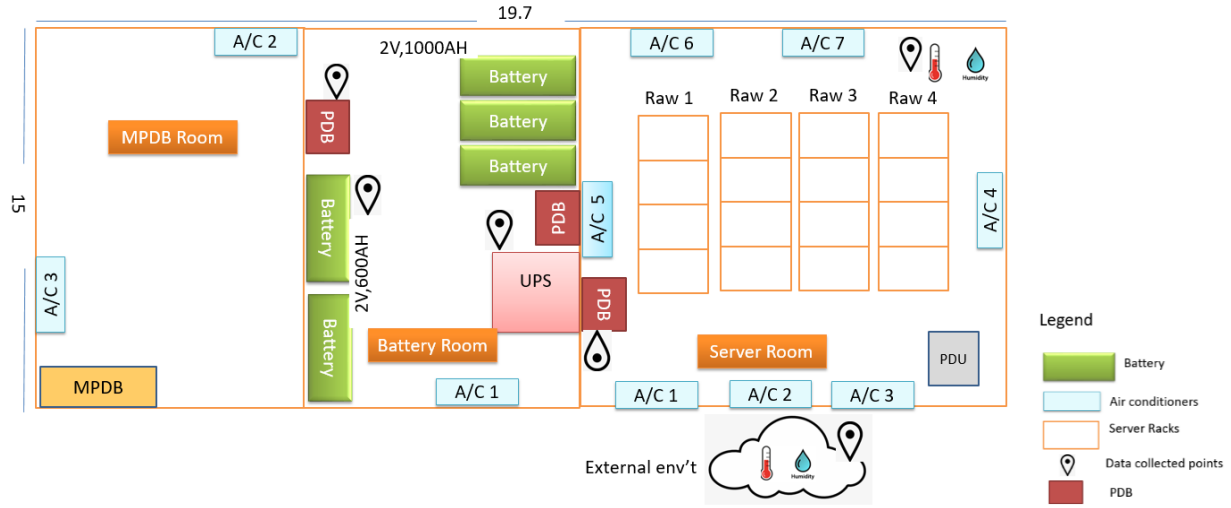

Figure 1-1 Legehar data center room Layout

The data center's gross floor area was 300 m^2. Figure 1-1 shows a simplified representation of a data center main components, which includes space for IT equipment, air conditioning equipment, UPS units, batteries, and Air circulation is maintained through cold and hot aisles (corridors). The three primary rooms were the server room (a combination of IT equipment, an air conditioner, and open space), the UPS and battery room, and the main power distribution room.

1.2 Statement of Problem

The telecommunications industry has a large network infrastructure and is constantly expanding it in size and capacity with huge investments. Ethio-Telecom currently operates 15 data centers

across the country (regions and Addis Ababa) to process and store data for subscribers, share applications, and transmitting information. Until now, no trend and gap analysis of these data centers' energy efficiency has been conducted. Trend and gap analysis methods are used to assess the overall status of the data center, allowing users to see which parts are doing well and which parts need to be improved.

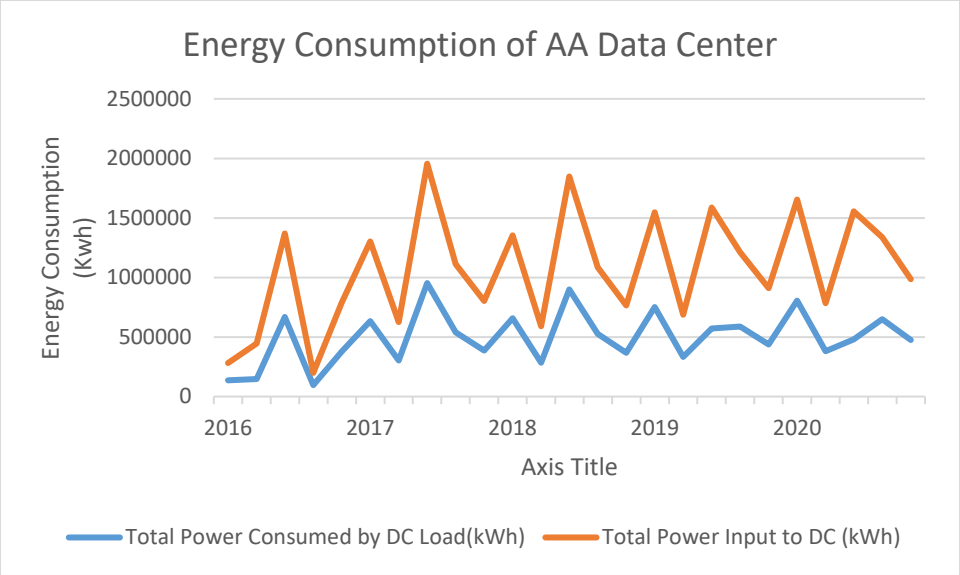

Figure 1-2 Power Consumption in Data centers Source: [NetEco monitoring tool].

The main problem motivating this study is data obtained from the NetEco Power and Environment Monitoring System over the previous five years (from 2016 to 2020). Figure 1-2 shows the total energy delivered to data centers versus the energy consumed by IT equipment. The data indicates that there is an energy loss between IT equipment and the main power distribution that grows year after year. The difference between the input power to the data center and the power consumed by IT equipment is 153,100 KWh, according to the 2019 EEU medium voltage industry tariff, which equates to a loss of roughly Birr, a local currency of 122,602.00 [7]. As a result, this study was motivated to understand more about the factors that contribute to data center energy losses and how to improve data center energy efficiency.

1.3 Objective of Study

1.3.1 General Objective

The main objective of this thesis is to investigate a root cause of the data center energy efficiency problem based on RFR VarImp feature selection and a GSA approach by using the historical data of the Ethio-Telecom Legehar datacenter.

1.3.2 Specific Objectives

In line with the general objective, the specific objectives to be addressed are the following:

- To undertake a literature survey on data center energy efficiency, RCA, GSA.
- To perform historical data collection of data center energy consumption in Legehar
- Compute PUE in order to understand the energy efficiency level.
- To select algorithms and techniques used for RCA and GSA suitable for data center energy dataset.
- Identifying the major factors for data center energy inefficiency.
- To recommend ET to improve its data center on the basis of the findings by considering performance target PUE value below 2.0.

1.4 Literature Review

This subtopic will include some of the literature on data center energy efficiency, RCA, and GSA. Previous research on data center efficiency has been studied from different perspectives, including data center environmental impact, energy cost reduction, and data center performance improvement.

The authors in [8] provide information on the major energy-consuming equipment and its factors in data center. According to the authors, cooling and ventilation system consumes about 40% of the total energy and the two factors for this energy consumptions are airflow management and climate condition (data center location). The authors' proposed solution for data center energy efficiency is ventilation and airflow management, with vertically placed server racks having better efficiency than horizontally placed servers by providing efficient heat transfer with lower airflow and contributing to data center energy savings. The author concludes Data center location selection is also an important factor for ensuring the reliable functioning of data centers in order to improve energy efficiency.

The research described in[9] proposed two new data center efficiency metrics based on the energy usage of IT equipment, including servers and switches. "Energy Eat by Servers and Switches (EESS) and Energy Eat and SLA Violation Factor (EESF)" are the names of the two metrics. These metrics are used to assess the quality of data centers in order to make more effective use of energy. Because data centers contain a large number and variety of components, a standard and diversified set of metrics is necessary in addition to the current measurement PUE to assess data center performance based on the kind of equipment used, such as IT equipment energy metrics, support infrastructure energy metrics, and so on. However, some common features must be present. The name of the measure, for example, should be logical and self-explanatory. Scalability is critical, which means that it should be able to adapt to technological and environmental changes, accurate, and that it should be general enough to give data-driven decisions.

Milad and Darwish [4] analyze the effects of UPS technology and topology on data center energy efficiency. Because data centers contain both computed and non-computed equipment. All of these components require massive quantities of energy, causing a considerable rise in energy efficiency issues. UPS is one of the drivers of data center energy efficiency problems. To solve the energy efficiency problem caused by numerous equipment components and operations, the authors attempt to address how UPS redundancy level and technology enhance UPS system and data center efficiency.

The authors in [10], explain how Sobol-GSA has been effectively used to predict PUE. The method utilized in this study is based on Sobol's method for this analysis taking climate variables and energy system parameters as inputs. Total and first-order sensitivity indices of key modeling parameters, the results suggest that focusing on reducing key input parameters is most important for reducing uncertainties in PUE values. Climate variables and UPS efficiencies are the most important parameters.

According to Jim Gao[11], a data center is the interaction of various components, such as mechanical, electrical, and so on. The existing PUE formulae are incapable of describing complicated interdependencies, making PUE difficult to evaluate. It would be impossible to evaluate each and every feature combination to increase efficiency due to regular fluctuations in IT demand and weather conditions. Machine learning is a valuable technique for improving data center energy efficiency in order to address these problems. Furthermore, local sensitivity analysis

is used to determine the effects of specific variables by changing one variable at a time while maintaining all others constant. This also helps in examining optimum set points and assessing the impact of set point variations. The primary weakness of this study, however, is that it ignores the interaction impact of variables [11].

1.5 Research Methodology

The RCA technique is separated into two sections: RFR feature selection and GSA techniques. RFR feature selection is interested in identifying significant features, whereas GSA is more concerned with measuring the impact level of features against the PUE function depending on the features selected. Figure 1-3 shows typical RCA implementation procedures, which include determining input features, energy efficiency (PUE) models, run PUE models, perform Sobol-GSA, and presenting sensitivity analysis findings.

The data collection process was for 37 weeks, and 17 features are included in the data set in five-minute time-steps at the Legehar data center. Indoor and outdoor environment features (temperature and relative humidity), and energy-related features (UPS input and output, the energy consumption of air conditioners...) are among the data collected.

To obtain a better understanding about data center energy efficiency and its characteristics, a literature review was conducted using Google Scholar and references found in other literature (IEEE papers, journals, textbooks, public documents, and white papers), Ethio-Telecom's low-level design (LLD) document, operational procedure, and vendor manuals. In addition, an informal discussion with Ethio-Telecom and vendor experts was held to discuss the limitations of the present data center's performance as well as to review certain confidential documents.

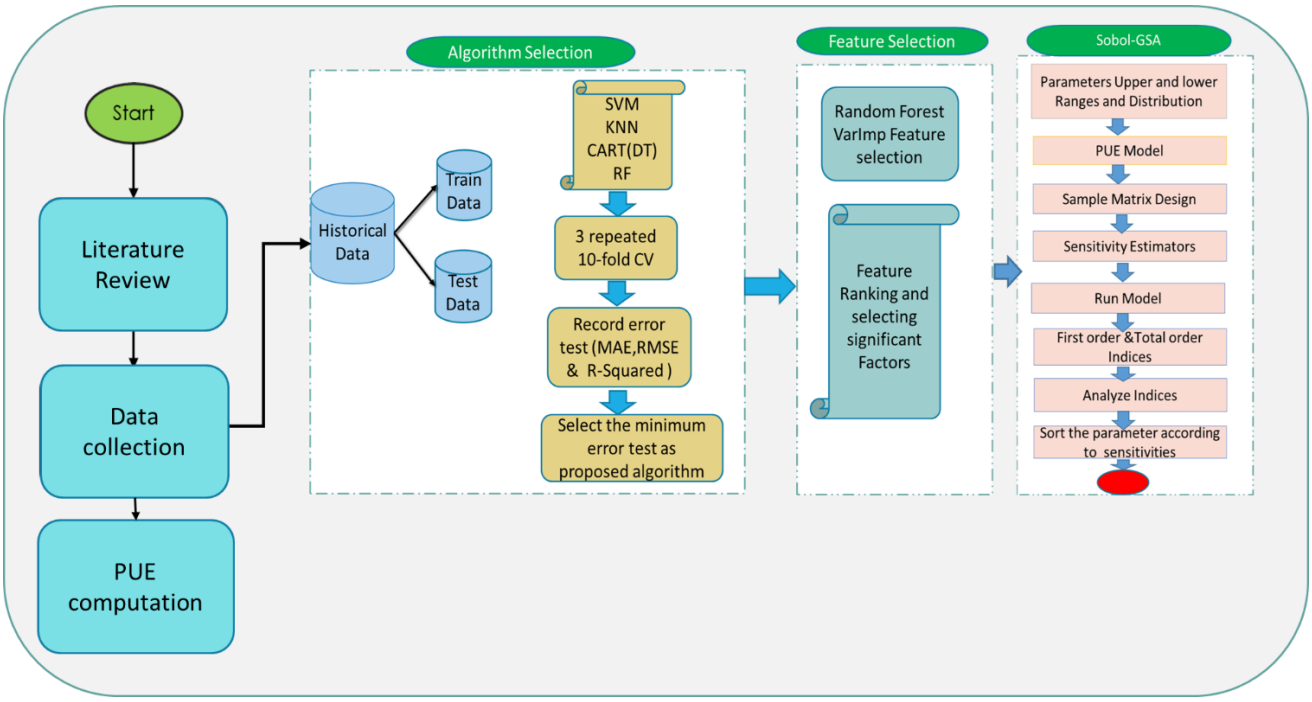

Figure 1-3 Methodology.

The two recommended tools are SALib in Python and the R-Studio packages. Both are open-source free software environments with a variety of sensitivity analysis and machine learning tools, but this study uses senSobol, caret, rpart.plot, Boruta, randomForest, corrplot, and ggplot2, are configured in R-Studio for statistical analysis, and simulated results.

1.6 Scope of the Research

The Legehar data center was visited as part of this thesis's fieldwork. And interviews with electrometrical technicians, data was collected from NetEco monitoring system, energy consumption analyzed by using PUE, and identifying energy efficiency problems using RFR and GSA techniques. Only covers the Legehar datacenter; it excludes regional and corporate data centers from consideration. However, the challenge is data collection process from NetEco monitoring system restrictions, the historical data is only collected on a weekly basis at a time with a five-minute sampling period. If the sample period is extended, we can get monthly data, but the sampling period will be in an hour because data centers are dynamic, there is a possibility that some data was lost in the process of gathering. As a result, the data obtained covers the period from May 2019 to February 2020, or nine months.

1.7 Contributions of the Study

The main contributions of this thesis are:

- Ethio-Telecom saves money by lowering OPEX using energy saving, which decreases energy use due to inefficient equipment and installations, allowing Ethio-Telecom to be more cost-effective.
- Several earlier researches focused on thermodynamic (free cooling) approaches to energy efficiency analysis. This study contributes to the filling of gaps by providing a statistical analysis of a PUE-based environmental and energy-related model.
- Previous studies have mostly focused on "local sensitivity analysis (LSA)", i.e., only looking at a single feature at a time without considering the interaction between variables. But this thesis used GSA to include the interaction effects of features.
- Investigate and recommend operational solutions to Ethio-Telecom data center inefficient power consumption. As a result, this study may be used as a reference for future research and as input for Ethio-Telecom. The study's findings, when applied to the Legehar data center, were applicable to other Ethio-Telecom data centers in the region and in Addis Ababa to assure energy efficiency.

1.8 Organization of the Thesis

The thesis is structured into six chapters. Chapter one discusses the introduction, problem statement, literature review, and methodology used to handle the energy efficiency problem of data center. Chapter two of this thesis provides a general overview of data center operation, interconnection of equipment in a data center, power distribution architecture, and the power supply equipment. Chapter three describes a brief discussion of RCA and sensitivity analysis. Chapter four presents the research methodology that includes, PUE calculations and a brief discussion on the RFR and Sobol-GSA mathematical explanation. Chapter five presents the analysis results and discussion, and Chapter six concludes the research work and on future research work directions.

Chapter 2 - Data Center

This chapter provides an overview of the fundamental concepts behind data center systems and design, as well as the key components of a data center, the factors contributing to the energy efficiency problem, and data center management from an energy efficiency perspective.

2.1 Overview of Data Center Operation

As previously mentioned, the most common types of electronic equipment found in data centers are servers, storage, and communications devices. This equipment is known as "IT equipment" it processes, store, and transfer digital data. Power conversion and backup equipment, as well as environmental control equipment, are usually found in data centers to guarantee that IT equipment is kept at the correct temperature, humidity, and operating condition [12].

The primary energy efficiency problem causes include equipment efficiency, redundancy, and oversizing of equipment to ensure reliability of data center. Furthermore, energy efficiency has never been a major priority for IT equipment manufacturers, because IT equipment efficiency has grown twenty-five times in the previous decade while energy efficiency has increased just eight times. This implies that saving energy is not a primary concern for IT equipment manufacturers [13].

2.2 Types of Data Center

Based on applications, data centers are divided into two types: corporate and Internet data centers. Private companies and organizations operate corporate data centers [14]. Their major aim is to deliver data processing and Web-based services to their own companies, business partners, and customers. In-house IT departments generally provide data center support and maintenance services. Telecommunications firms and service providers, on the other hand, own and operate Internet data center. To provide IT services that may be accessed over the Internet. Data centers, On the other hand, can be classified into four categories based on availability (yearly uptime) and redundancy of electrical paths for power, IT equipment, and cooling systems [14]:

- **Tier 1**: is the most fundamental type of data center tier, having no redundant infrastructure parts. It is expected to be up and running 99.671% of the time (annual downtime of 28.8 hours).

- **Tier 2:** a data center contains redundant components but a single source of power and cooling supplies. It is expected to be operational 99.741 percent of the time (with 22 hours of downtime per year).
- **Tier 3**: IT equipment are supplied by dual, and independent sources and they are always up and running. It's supposed to work 99.982 percent of the time (with only 1.6 hours of downtime per year).
- **Tier 4:** All of the infrastructures in data centers, both computing, and non-computing, are redundant. It should be up and running 99.995% of the time (with a 26.3-minute annual downtime).

2.3 Energy Efficiency Standards and Metric

The ratio of total data center energy consumption to energy used by IT equipment is known as the PUE. The PUE metric is the global standard for analyzing data centers energy efficiency. It enables data center operators to assess their data center energy efficiency, compare it to that of other data centers, and develop plans to improve it. Total power includes energy used in useful work such as data storage and processing, as well as energy spent for "other" activities, such as lighting and air conditioning [15].

Table 2-1 Data center PUE rating Source [16]

PUE	Rating
PUE > 2.0	Inefficient
1.59-1.8	Average
1.2-1.4	Efficient
<1.2	Very Efficient (“state-of-the art”)

Table 2-1 shows the PUE standard for data centers. If all available energy is used for useful work, the PUE value of a data center is 1.0, which is the ideal value. According to a survey by Uptime Institute Network Members (a global user organization of big data center owners and operators) in 2007, the overall average PUE was 2.5, however, the first industry research in 2011 indicated that the PUE had improved to 1.98. However, further research in 2013 revealed additional

progress, with the average PUE improved to 1.65. The average PUE in 2020 was 1.59, indicating a little improvement since then. PUE values approaching 1.2 and 1.4, on the other hand, are considered "best practice"[17]. Moreover, National Renewable Energy Laboratory (NREL) also study, the average PUE value of a data center is 1.8, with efficient data centers having PUE values of 1.2 or below [18]

2.4 Data Center Power Supply System

The Main Power Distribution Board (MPDB) is a 380V three-phase power system board. It is an electric distribution system that splits electrical power into secondary circuits based on the rating of the loads attached to it. The system is protected by fuses and circuit breakers. MPDB receives power from a commercial source or a generator and distributes it to connected loads, such as IT equipment, via UPS and PDU, cooling system, and accessory loads. When a circuit breaker trips, it must be manually reset. However, when a fuse blows, a new one must be installed. The MPDB is connected to a number of sensors as Alternating Current (AC) parameters (three-phase voltage, phase current, and frequency) and communicates the readings to the NetEco monitoring system.

PDUs convert 380V three-phase power to 220V single-phase power before being used by IT equipment. IT equipment is powered by AC, which is converted to Direct Current (DC) by the IT device power supply unit, as shown in Figure 2-1. The IT equipment racks in a data center are arranged in a hot/cold aisle parallel configuration. Each row of equipment racks is usually equipped with its own PDU. The UPS, which is made up of batteries and power conversion devices, serves as a backup energy source for servers in the event of a power outage[19].

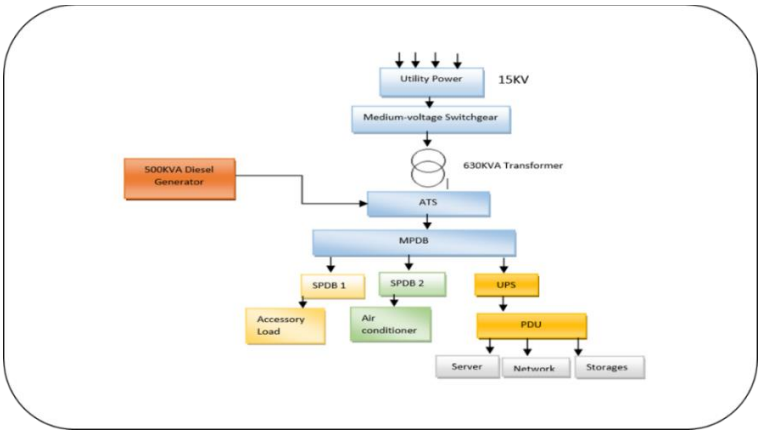

Figure 2-1 Power delivery path of a data center Source [6]

2.5 Data Center Components

A data center can range in size from a single room to a huge building. The most important components, on the other hand, are the same regardless of size and capacity; Figure 2-1 shows a list of common components as well as a power flow diagram for a data center.

2.5.1 IT Equipment

As previously discussed, IT equipment includes servers, storage, and network devices, which are also referred to as mission-critical loads. A large number of heterogeneous IT equipment operates in data centers. The IT equipment load in a data center is very dynamic, varying throughout the day based on user activity.

2.5.1.1 Servers

Servers are the core of a data center system that generates useful output, and they consume a large amount of energy in a data center while also generating heat it is a load on the cooling system. This is primarily due to equipped with a large amount of processing power and computing capabilities. The energy consumed by the server is represented as a function of the energy needed by the CPU, memory, fans, and I/O devices. The CPU (processor) has been the most power-hungry component of a server. Processor energy consumption increases with usage and the number of processor cores [20 , 21].

2.5.1.2 Storage (Memory)

A data center stores data both temporarily and permanently. For short-term processing is primary storage, whereas long-term storage is secondary storage. Secondary storage is divided into two categories: hard disk drives (HDD) and solid-state drives (SSD), often known as flash storage. An HDD retrieves data using mechanical platters and a moving read/write head, whereas an SSD stores data on memory chips that are easily accessible. The storage system is the second greatest power consumer in IT equipment [20].

2.5.1.3 Network

Network equipment is used to provide a connection between various external and internal IT devices, most commonly servers. There are several ports on this equipment, which are generally

switches and routers. The energy usage of network devices is a function of the number of ports network equipment has. When compared to other kinds of equipment, Network devices consume less energy than other types of IT equipment [20].

2.5.2 Components of UPS

UPS battery and a diesel generator are commonly used as backup power sources in data centers. UPS batteries will keep IT equipment powered in the interval between the utility power failure and the generator starting up.

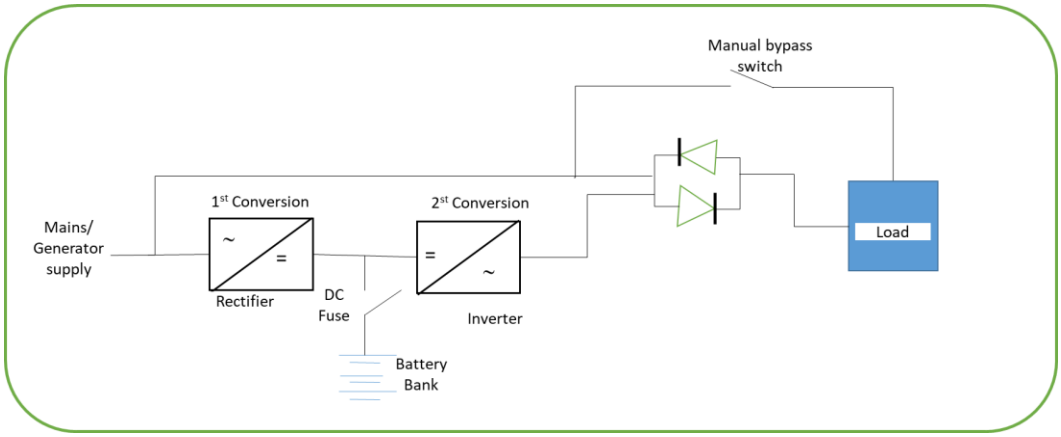

Figure 2-2 Schematic Diagram of double conversion UPS System Source [26].

Figure 2-2 shows a UPS with double conversions steps. In the first conversion from incoming AC power to DC for energy storage and then reconverting from DC back to AC powering the IT equipment via PDUs. During UPS repair, the manual by-pass mode is used to separate the path rectifier, inverters, and battery. A UPS configuration is usually designed with multiple redundant configurations for availability and reliability. The AC-DC-AC multiple power conversion steps result in power loss. UPS system measurement includes input and output voltage, current and frequency, output power Volt-Ampere (VA), Killo-Watt (KW), and Power-Factor (PF), battery voltage, battery current (charge/discharge), battery back-up time, and battery temperature [22].

The double conversion UPS system components:

1. A rectifier –AC to DC converter;
2. An inverter –DC to AC converter;
3. A static bypass switch- When the rectifier or inverter fails, a static bypass switch instantly connects the load to the mains power supply.
4. A manual bypass switch; also known as maintenance bypass.
5. A battery.

By the time of conversions, electrical energy was lost and converted to thermal energy. There are various factors that affect UPS efficiency such as low UPS load and technology type (Double conversion or flywheel) etc.

2.5.3 UPS Battery

The battery is a major element of the UPS power system, which is connected in a series-parallel configuration. The series connection increases the voltage level, but the parallel connection increases the battery's current producing capacity Ampere-Hour (AH). The number of connected batteries, commonly referred to as battery banks, is determined by several factors, including the battery cell voltage, which can be either 12V or 2V, the system voltage required (380V or 220V), the actual loads required, and how long the battery is expected to power the IT equipment. It is also dependent on the reliability of mains supply stability and the backup diesel generator status. UPS batteries are classified as Valve Regulated Lead Acid Battery (VRLA) or sealed batteries, as well as wet or flooded-cell batteries. VRLA batteries are the best solution for supplying uninterruptible power since they require less maintenance and have a longer lifespan. Flooded-cell batteries, on the other hand, have shorter lifespans and must be maintained more frequently. All of the batteries in the Ethio-Telecom data center are VRLA batteries [23].

2.5.4 Mechanical Infrastructure

One of the modern data center criteria is raised floors and dropped ceilings, i.e., a down flow air conditioner and a hot aisle/cold aisle (hot and cold corridors) arrangement between rows to make the data center energy efficient by preventing air recirculation. [24].

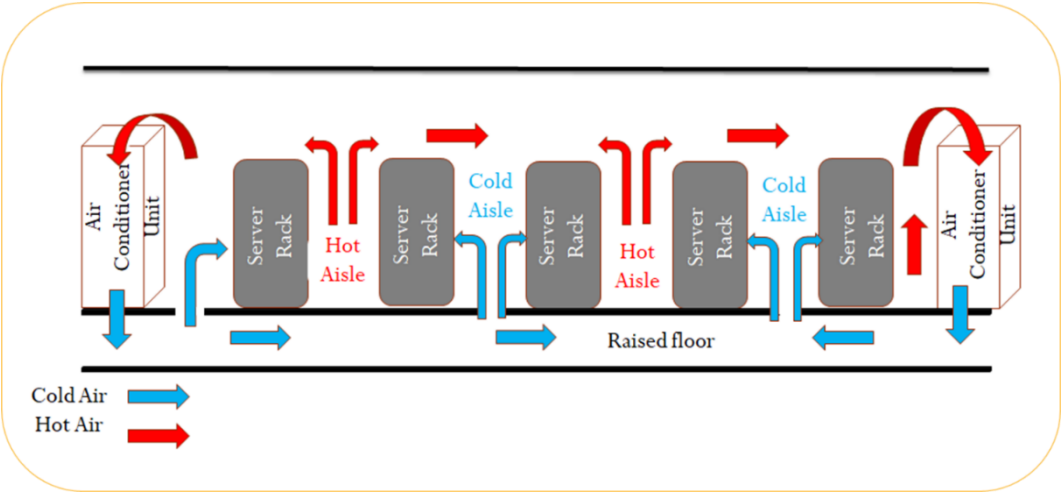

Figure 2-3 Hot /Cold aisle and raised floor layout Source [25]

As figure 2-3 depicts, an air conditioner in a data center, including supply and return air circulation, both the area beneath the floor and the space above the ceiling may be utilized for cabling, and the space beneath the floor is also used to provide cold air to the IT equipment. Hot air from the room is collected in the area above the ceiling [26]. According to the "American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE)" a data center should be kept at a temperature of 18 to 27 degrees Celsius (°C) or 64 to 81 degrees Fahrenheit (°F), with relative humidity (RH) of 40 to 55 percent [27].

2.5.5 Accessory Load

In a data center, accessory load is part of the support infrastructure. Power for lighting, door access control, video surveillance systems, and fire and safety systems are generally included in the accessory load, but in our case, the battery room, power distribution room, and UPS room air conditioners are also included.

Chapter 3 - Root Cause and Sensitivity Analysis Techniques

This chapter will discuss why the RCA and sensitivity analysis were chosen as techniques for this thesis. The benefits and drawbacks of various root cause and sensitivity analysis techniques, as well as the importance of data center variables, are explained.

3.1 Overview of RCA

The RCA method is a problem-solving technique that aims to identify the root causes of problems and challenges. In other words, a root cause is an explanation of why something happened or determining what is causing it. It helps to solve the recurrence of a problem [28]. However, certain basic causes are more difficult to determine; as a result, cause analysis methods, such as machine learning-based analysis, are effective in finding the root of a problem [29]. This research utilized the RCA approach to figure out what was causing the data center's energy inefficient. When it comes to efficiency, the study focuses on the overall efficiency of data center components, for example, UPS efficiency, cooling system performance, etc. [30].

There are two types of RCA: data-driven RCA and non-data-driven RCA. Data-driven RCA is based on historical (recorded) data analysis and interpretation, such as machine learning algorithms, whereas non-data-driven RCA is based on qualitative analysis via a series of phases, for example, "five why's" methods refer to asking why five times[31]. The goal of this study is to provide a data-driven RCA that explains what happened, why it happened, and how to improve the problem.

Both data-driven and non-data-driven RCA have five phases in order to complete the RCA process[31].

1. Define the problem: Assessing the situation to focus on the real problem rather than its symptoms. SMART concepts are used in this case (Specific, Measurable, Achievable, Realistic, and Timely) For example, "Legehar data center energy efficiency problem".
2. Gather data/evidence: Getting actual information that helps to describe the problem as well as getting a detailed understanding of the problem such as environmental and energy data in our case.
3. Find root causes: Identify the probable causes that contributed to the problem.

4. Determine which causes to eliminate or modify in order to prevent the problem: Corrective intervention identifies and prioritizes the most likely underlying causes of the problem, as a temporary countermeasure may not fix the fundamental cause.
5. Confirm the solution: Develop solution recommendations that effectively prevent the problem.

3.2 Machine Learning for RCA

Machine learning algorithms are commonly classified as supervised, unsupervised, and reinforcement learning.

Supervised machine learning algorithms can apply what they've learnt in the past to new data and predict future occurrences using labeled data. The learning algorithm can also discover errors by comparing its output to the actual, intended output, allowing the model to be adjusted as needed. Unsupervised learning extracts patterns from unlabeled data and applies them to grouping and association problems. Reinforcement learning algorithms interact with their surroundings by taking actions and recognizing faults or rewards. Simple reward feedback is all that is necessary for the agent to determine which action is better [32].

In general, RCA identifies the particular cause of the problem in order to take appropriate action to prevent a reoccurrence of the problem. Within each part, there are different techniques, algorithms, and models to work for RCA such as Random Forest, Decision Trees, Support Vector Machines, and K-Nearest Neighbor, Neural Network, and many more. Random Forest has been the selected one.

3.2.1 Random Forest Regression

Random Forest Regression (RFR) is an algorithm that can be used both for a classification and regression technique. A set of decision trees based on the training data is used to create random forest models. The RFR method generates a forecast based on the average prediction of a group of trees, rather than getting the target value from a single tree. The trees themselves are built by fitting the training data to randomly selected sets of rows and columns. This approach is known as Bootstrap Aggregating (Bagging) randomly selecting parameters with replacement, and it reduces bias by building each tree on various sections of the input at random. The approach of RFR predictions avoids over fitting that can occur when single decision trees are used. In this research,

the tuning parameters for RFR m-try and N-tree were used for the variables selected at random and the number of trees respectively [30].

The number of trees in the RFR method is an essential parameter, and the more trees utilized, the less over-fitting will occur. However, there is a cost in terms of increased computing time. The RFR algorithm parameter defines Mean Squared Error and Mean Absolute Error, which can be used to determine the quality of splits in the trees.[33].

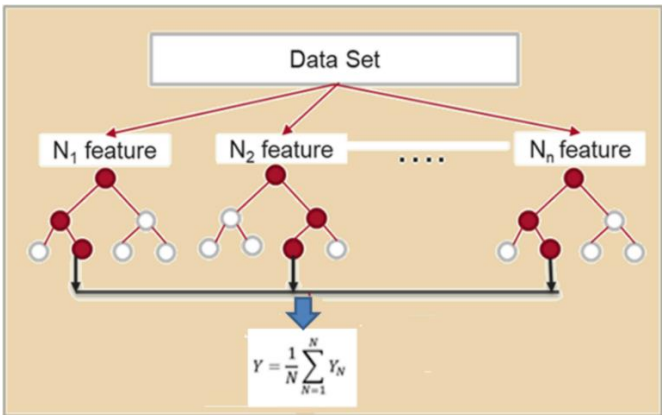

Figure 3-1 Random Forest Source [33]

As depicted in Figure 3-1, the following stages demonstrate how RFR is built using N-trees:

1. From the data, draw n_{try} bootstrap samples.
2. Build a tree for each of the bootstrap sample sets. Select m_{try} variables for splitting at each node of the tree at random for each tree (N = 1, 2... N). this tree's terminal nodes should contain at least n leaf cases.
3. Combined information from the N trees for new data prediction, according to steps (2) and (3).
4. Compute an out-of-bug (OOB) error rate based on the data not in the bootstrap sample.

3.3 Sensitivity Analysis

Sensitivity analysis is also known as “what-if analysis”. Sensitivity analysis is carried out within predefined boundaries that are determined by one or more input features. Can help to demonstrate how changes in a variable impact the outcome by establishing a collection of variables. It is a

valuable technique that provides decision-makers with more than just a solution to a problem. There are two different types of sensitivity analysis methods, LSA and GSA [34]:

1. LSA: - Sensitivity measurements are taken when one variable is changed while the others remain constant. The term "local" refers to the fact that the derivatives are computed at a single place ("one-at-a-time analysis"). This type of sensitivity analysis is useful for basic functions but ineffective for complex models.

Typical steps for LSA

    1. In energy efficiency analysis, for example, the PUE for a certain base case input value (for example, UPS Efficiency) for which the sensitivity is to be calculated is determined. Keep all of the model's other inputs constant.
    2. Then, while keeping the other inputs constant, calculate the output's value at a different input value (Cooling system).
    3. Calculate the difference in percentages between the output and the input.
    4. Divide the percentage change in output by the percentage change in input to obtain the sensitivity result.
    5. Repeat the sensitivity testing technique for another input while keeping the other inputs constant until you get the sensitivity figure for each of the inputs.
    6. Finally, the lower the sensitivity value, the less sensitive the output is to changes in the input, and vice versa.
2. GSA:-varies all the variables simultaneously, using a range of samples to analyze which model parameters have a small or large influence on the model output.

In general, sensitivity analysis tries to address the following key questions [35]:

1. Is there a single dominating factor, or are they caused by a number of factors? Determining how much uncertainty in model input parameters contributes to overall model output variability
2. Do the factors act independently or in combination with one another? Evaluate the input-output relationship.
3. Which factor is the most important? Identify the critical and influential factors that affect model outputs.

The fundamental drawback of the LSA technique is that interactions between input variables are ignored, whereas GSA is more concerned with the impacts of input interactions over the whole input space. As a result, the global approach is seen as more reliable. It helps to identify the primary factors influencing data center energy efficiency. The goal is to investigate the potential changes in energy usage for data centers in actual use and identify the important variables influencing PUE.

3.4 Types of GSA

There are many GSA techniques but the most commons are Morris, Fourier Amplitude Sensitivity Test (FAST), and Sobol.

3.4.1 Screening-based Method (Morris)

The Morris method is categorized as GSA because the baseline changes in each step and the final sensitivity measures are determined by averaging at different locations of the input space. Unlike other GSA approaches, which accept input values directly from distributions, this technique takes input factors as a discrete number. Morris technique may be used to generate two sensitivity indices. Standard deviation (σ) to measure the interaction with other factors or non-linear effects, and the average (μ) to quantify the primary influence of the input component on the output. When compared to other global sensitivity analyses, the major benefit of using Morris approach is the minimal computing cost. The disadvantage of this strategy is that it provides qualitative measurements by rating input elements, but it is unable to quantify the impacts of different factors on outputs. As a result, this approach does not allow for self-verification, which means users have no means of knowing how much of the overall variance of outputs has been included in the analysis [36].

3.4.2 Fourier Amplitude Sensitivity Test (FAST)

The FAST method is a widely utilized GSA technique. The variance of a model output is broken down into component variances given by distinct model parameters using a periodic sampling technique and a Fourier transformation. The FAST analysis is mostly limited to the estimation of partial variances supplied by the main effects of model parameters but does not allow for those contributed by particular interactions among factors[37].

3.4.3 Sobol

The Sobol technique is a variance-based method for decomposing the uncertainty of outputs in relation to their corresponding inputs. The first order (S_i) and Total order (S_{Ti}) effects are the two major sensitivity metrics utilized in this thesis. The S_i take into account the main effects of the output changes caused by the related input. S_{Ti} account for the entire contributions to output variance related to the related input, and include both first-order and higher-order effects related to input interactions. As a result, the difference between $S_{Ti} - S_i$ measures the interaction of *i* with other factors. However, as compared to other GSA approaches, the Sobol method is far more computationally costly [36].

Chapter 4 - Research Methodology

This chapter describes how RFR and Sobol-GSA can help in determining what is causing the problem (data center energy inefficiency), as well as the features required to build a model and the mathematical explanation for GSA.

4.1Data Analysis

In data centers and other telecom infrastructures, several sensors, environmental (temperature and humidity) and energy-related (current and voltage) sensors, have been installed to detect alerts (warning, minor, major, and critical) and measurements in real-time. These sensors continuously collect measurements and transmit them to a central office monitoring system for remote monitoring [23]. The data collected is used for analysis like fault prediction, equipment performance analysis, and a variety of other applications. This thesis collected and analyzed energy and environmental data in order to compute data center energy efficiency (PUE), and RCA. Table 4-1 shows a list of features collected from the NetEco monitoring system. It was collected on a weekly basis. A total of 37 weeks of data were used, with 17 features and a 5-minute sample time.

Table 4-1 List of Features Source: [NetEco monitoring tool]

No.	Parameter	Unit	Description
Environment			
1	Outdoor Temperature	℃	the outside temperature
2	Outdoor Relative humidity	%	The amount of water vapor in a given amount of air
3	Temperature **Set** point	20℃	Air conditioner current set points
4	Relative Humidity Set point	45 %	
5	IT Room humidity	%	
6	IT Room temperature	℃	
7	Battery room Temperature	℃	
Energy Related			
1	Fan Power	KW	Air conditioner supply and return air fan energy consumption
2	Compressor Power	KW	Air conditioner compressor energy consumption
3	Battery charging rate	%	
4	Battery Voltage	V	
5	Battery Backup Time	Min	
6	PDU Output Power	KW	IT equipment (Server, Storage and Network) energy consumption (Supply power for IT equipment)
7	Input Power of UPS	KW	Three Phase Input Power for UPS
8	UPS Output Load Percentage Per Phase	%	Current UPS Load status per phase
9	UPS Efficiency	%	
10	Accessory Load	KW	Lighting, Door Access and Security Camera

4.2 Evaluation of the selected techniques

4.2.1 ML Algorithm Selection

Algorithm selection is a necessary step before feature selection. Identifying a suitable algorithm for our dataset among numerous algorithms is a challenging issue, various authors apply a trial-and-error technique and then select the one with the highest performance in the test. However, this thesis takes advantage of three error metrics: the coefficient of determination (R-squared), mean absolute error, and Root Mean Squared Error.

4.2.1.1 Mean Absolute Error

Mean Absolute Error (MAE) measures the prediction error it is the difference between the target value and predicted value then takes the mean of all absolute values of all errors [38].

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|y_i - x_i| \quad (1)$$

Where n is the number of samples,

y_i are the target values, and

x_i are the predicted values.

4.2.1.2 Root Mean Squared Error

The root mean square error (RMSE) is a statistic that assesses how well prediction errors are dispersed. In other words, it shows how concentrated the data is on the best-fit line. Because RMSE is scale-dependent, it is only useful for comparing prediction errors of various models or model configurations for a single variable, not between variables [39].

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=0}^{n}(y_i - x_i)^2} \quad (2)$$

Where: y_i are the observations.

x_i Predicted values of a variable, and

n the number of observations available for analysis.

∴ MAE and RMSE closer to 0 means that the model predicts with lower error and that the algorithm is better [38].

4.2.1.3 Coefficient of determination

The coefficient of determination (R-squared) is the ratio of the predicted variance to the overall variance is known as the R-squared. It's a common metric for determining how well a model matches the data's "goodness of fit"[40]. The R-language and several libraries (Packages) are used to perform the comparison experiment for KNN, SVM, DT, and RFR, and the results are displayed in the table below.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \quad (3)$$

Where:- $\hat{y}$the predicted value of the i^{th} sample,

y_i is the corresponding actual value over n-samples and

$\bar{y}$ is the mean

∴ The coefficient value varies from 0 to 1, with 1 being the best and 0 being the worst [40].

Table 4-2 ML algorithm comparison experiment result

Algorithm	MAE	RMSE	Rsquared
k-Nearest neighbours	0.01307494	0.01719802	0.9382102
Random Forest Regression	**0.01096575**	**0.01531134**	**0.9460464**
Decision Tree	0.01685037	0.02681028	0.8676008
Support Vector Machine	0.04630664	0.05396108	0.8122725

As summarized in Table 4-2 the errors for each algorithm. K-NN, SVM, RFR, and DT algorithms are compared in terms of their error rates. It is evident that the RFR algorithm performs better than the rest algorithm in all three measures.

In addition to the metrics discussed above, a brief discussion of the benefits and drawbacks of the algorithms based on the conditions given in Table 4-3 for DT, SVM, KNN, and RFR is provided. The mixed data column indicates the algorithm's capacity to learn from several data sources. The robustness column represents the algorithm's ability to deal with noise and erroneous training data. The large data column indicates the algorithm's ability to scale in size and hence deal with massive datasets. The Accuracy column refers to the accuracy of the algorithms. The parameter tuning column in the final one is adjusted to improve the algorithm's ability to change and modify its outputs depending on feedback[30].

Table 4-3 ML algorithm

Algorithms Property	RFR	DT	KNN	SVM
Handles mixed data	☑	☑	☒	☒
Robustness	☑	☑	☑	☒
Small dataset	☒	☒	☒	☑
Scalability (large dataset)	☑	☑	☒	☒
Predictive Accuracy	☑	☒	☒	☑
Parameter Tuning	☑	☑	☒	☒

4.2.2 GSA Methods Comparison

There are various ways of performing GSA, such as Morris, FAST, and Sobol. A brief explanation of the advantages and drawbacks of the methods based on the circumstances described in Table 4-3 has been used to provide a decision-making tool [35].

Table 4-4 GSA methods comparison Source [35]

Criteria for comparison	Commonly used GSA methods		
	Morris	FAST	Sobol
Discrete inputs	Yes	No (not suitable for discrete inputs)	Yes
Model independence	No	Yes	Yes
Non-linear, input-output relationship	No	Yes	Yes
Non-monotonic input-output relationship	Yes	Yes	Yes
Robustness	Yes	Yes	Yes
Ability to apportion the output variance	No	Yes	Yes
Higher-order interaction of parameters	No	Only 1st order	Yes
Quantitative measure for ranking	Yes	Yes	Yes
Computational efficiency	less	high	high

According to Table 4-3, the Sobol technique is the algorithm that meets the specified requirements. As a result, the approach was chosen based on more than just a comparison of the techniques listed above. Moreover, it has the ability to express graphically.

4.3 Random Forest Method for RCA

In Machine learning, there are procedures to use the data set. Split the data set into train and test data: first to train the algorithm, and then to test it. But using these methods has its drawbacks. If the test data is too small, the result is less convincing, and increasing the amount of the test data improves reliability but decreases the number of rows in the training data, making the model forecast poorly. Using the k-fold cross-validation technique is a better solution to overcome the drawback.

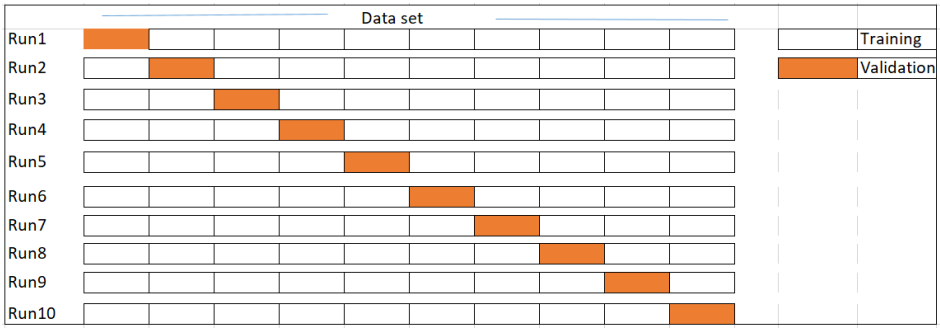

Figure 4-1 10-fold cross-validation

As illustrated in Figure 4-1, samples are subdivided into k pieces and then k-1 portions are utilized to create the model, and the remaining part is used for evaluation. The procedure is repeated k-times, with each part serving once as a test set and k-1 times contributing to the training set [41]. Unfortunately, even with a small sample size, the findings may still be dependent on a certain data split. In fact, a specific split might skew both training and testing. To solve this problem, Repeat cross-validation multiple times with independent data splits at each iteration. The 10-fold cross-validation method is repeated three times in our case (see the R code below). Each iteration begins with the selection of a collection of relevant features for the training set. The RFR model is then created using specified variables, and the model's quality is assessed using the test set.

$$\boldsymbol{R > Control < -trainControl\,(method =' repeatedcv', number = 10, repeats = 3)}$$

The resampling technique, which is based on three repeats of 10-fold cross-validation, is used to select the features. Each iteration divides the training into ten pieces, which are then combined to generate ten different samples. After that, each sample's RFR model is created, and the variable importance (VarImp) algorithm for each variable is gathered. Variables are then ranked using the total of VarImp from the 30 samples. The 30 variables with the highest VarImp are chosen. These variables are utilized to construct an RFR model for the training set and to evaluate the prediction on the test set [41].

4.3.1 Feature Selection Techniques

As explained previously, feature selection approaches used to determine which features contribute to the root cause of a problem RFR is the chosen method for ranking features according to their significance.

4.3.1.1 VarImp

In the VarImp feature selection algorithm, relevant features are identified simultaneously while the machine learning process is carried out. This study prefers this technique, i.e. build a model using all data set on the basis of the data center attributes, and develop a model for root cause analysis. Due to the assumption that the model attributes correspond to the data center energy efficiency problem. It's important to note that the aim of creating a model in this situation is not to accurately forecast like a failure prediction, but rather to analyze the relevance of each variable

included in the model based on the known energy and environmental features. The problem can be traced down to a root cause by using this approach [42].

As previously mentioned, this study uses the RStudio randomForest library that helps for the tree building process, for each tree, there is a subset of items that were not utilized in the tree's creation, known as OOB objects, i.e., for each tree, the MSE is computed and the OOB data is recorded. This enables unbiased estimation of the classification error and VarImp. For a given variable X, a subset S_X of trees that used X is identified. The prediction error on the OOB objects is then calculated for each tree from S_X. The values of X are then randomly permuted among OOB objects, eliminating all information about the true values of X, and the prediction error for these objects is computed. The differences are averaged and normalized by the standard error on OOB objects caused by the loss of information on X, which is a measure of its significance[41].

The two factors used to assess the RFR algorithm: the number of decision trees in the RFR (N-tree) and the number of randomly selected variables (m-try).

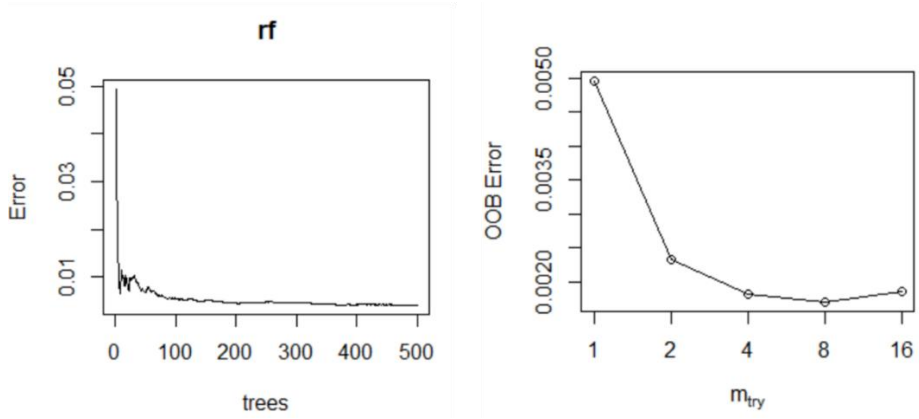

Figure 4-2 RFR model n-Tree & M-try

Figure 4-2 shows the results, where the values are compared by OOB error, which is minimized when the number of trees exceeds 200, and m-try was set to the lowest value possible when the randomly chosen variables were 8.

4.4 Computing PUE

The power consumption flow from the main distribution board to IT loads was addressed in the previous chapter, but the real measurement and quantitative findings didn't appear until this section. This section describes the data center power flow in detail and describes PUE.

The "support infrastructure," which includes power supply devices (UPS and PDU), cooling systems, and accessory loads, consumes a significant amount of electrical energy in a data center.

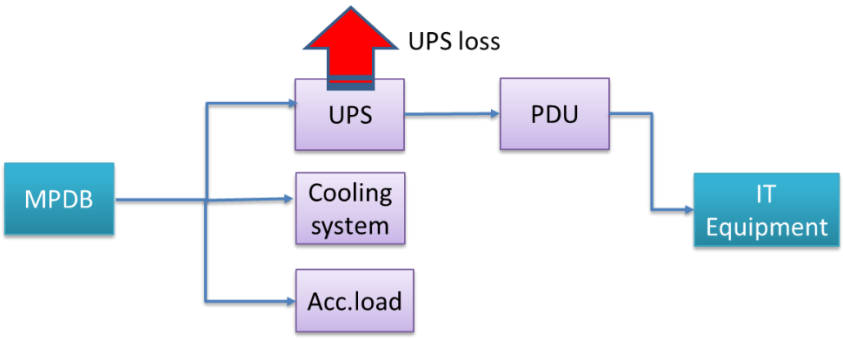

Figure 4-3 Data center energy efficiency (PUE) computation Source [19]

Figure 4-3 depicts the power flow of a typical data center power system[19]. According to the green grid association, PUE can be defined as the ratio of all energy consumed by the data center to the actual energy consumed by IT equipment. It tells how much energy IT equipment uses and how much energy is overhead.

Features required for performing the PUE calculation are: First, energy-related, the total electrical energy usage by the IT equipment's (IT_load) includes energy consumption of servers, network, and storage, The energy consumed by the cooling system is the summation of the energy required by the fans (supply and return) and the compressor (Fan_P and Comp_P), which require the most energy. A UPS is utilized as a power conditioning in the data center to provide uninterruptable power for IT equipment. The performance of UPS is determined by two parameters; the efficiency of UPS (UPS_Eff) and the load percentage of UPS (output_load_per). These two factors are related to UPS loss because a significant quantity of power is wasted during the conversion process inside UPS. The other major component for PUE computation is the energy used by accessory load (Acc_load) it includes lighting, UPS room, and battery room air conditioners, power for security cameras, and access doors. All the above-stated features are collected in kilo-watt hours.

The second most relevant data collected was the supply and return air of the air conditioner in the IT room and exterior environment in degrees Celsius and the indoor and outdoor relative humidity (RH%), which are classified as environmental factors.

Since the IT equipment area generates a lot of heat air conditioners are in charge of maintaining IT equipment at a safe working temperature. A typical metric for determining a cooling system's efficiency is the coefficient of performance (CoP) [2].

$$PUE = \frac{P_{Total}}{IT_{load}} \quad (1)$$

$$PUE = \frac{P_{Cooling} + P_{UPS\_loss} + P_{Acces\_load} + P_{IT\_load}}{P_{IT\_load}} \quad (2)$$

$P_{Cooling} = P_{fan} + P_{compressor}$

$P_{compressor} = \frac{IT_{load}}{CoP * T}$

$CoP = 0.0068 * T^2 + 0.0008 * T + 0.458$ *and*

$T = T_{supply} + T_{Setpoint}$

Where: $P_{Cooling} = power\ consumption\ of\ cooling\ system$

$P_{UPS\_loss} = power\ loss\ in\ UPS$

P_{Acces_load} = $power\ consumption\ of\ Accessory\ load$

$P_{IT\_load} = Power\ consumption\ of\ IT\ equipment$

CoP = Coefficient of Performance

T_{supply} - Supply Temperature and Temperature set point

$T_{set\ point}$ - Temperature set point air conditioners configuration to control the IT room.

$$PUE = \frac{P_{fan} + P_{compressor} + P_{UPS\_loss} + P_{Acces\_load} + P_{IT\_load}}{P_{IT\_load}} \quad (3)$$

$$PUE = 1 + \frac{P_{fan} + P_{compressor} + P_{UPS\_loss} + P_{Acces\_load}}{P_{IT\_load}} \quad (4)$$

4.5 Mathematical Explanation of Sobol method

GSA includes Sobol indices. It separates the variance of the models' output into fractions that may be attributed to single or groups of inputs. for example, In a model with two inputs 'a' and 'b' and one output, variation in the first input 'a' may account for 60% of the output variance, 25% of the variance in the second input 'b', and 15% of the variance in the output related to interactions

between 'a' and 'b'. Because they measure sensitivity throughout the entire input space, these percentages can be simply understood as GSA measures[43].Any model can be viewed as a function $y = f(x)$,

Where x is a vector of n model inputs $[x_1, x_2, x_3, x_4 \cdots x_n]$ and

Y is output vector $y = [y_1, y, y_3, \cdots y_n]$ related by:

$$y = f(x) = f(x_1, x_2, x_3, x_4 \cdots x_n) \quad (1)$$

The sobol-GSA sensitivity analysis methods consist of a decomposition of the variance of the output into a sum of contributions due to the input factors and their interactions.

$$f(x) = f_0 + \sum_{i=0}^{n} f_i(x_i) + \sum_{i<j}^{n} f_{ij}(x_i, x_j) + \cdots + f_{1\ldots n}(x_{1,\ldots}, x_n) \quad (2)$$

Where f_0 is a constant

f_i is a function ofx_i,

f_{ij} A function of x_i and x_j

$$f_0 = E(Y)$$

$$f_i(x_i) = E(Y|x_i) - f_0$$

$$f_{ij}(x_i, x_j) = E(Y|x_i, x_j) - f_0 - f_i - f_j$$

It can be observed that f_i represents the impact of modifying x_i alone (first-order indices) and f_{ij} represents the effect of varied x_i and x_j (second-order indices) which is the effect of individual variations and interaction.[44]. Sobol provides the Sobol indices to quantify the division of the output variance by applying the variance decomposition based on the decomposition Equation (2), as follow: $V(Y) = \sum_{i=1}^{n} V_j(Y) + \sum_{i<j}^{n} V_{ij}(Y) + \cdots + V_{123\ldots n}$ (3)

$$Var(Y) = \sum_{i=1}^{n} V_i + \sum_{i<j}^{n} V_{ij} + \cdots + V_{12\ldots n} \quad (4)$$

Where $V_i = Var_{xi}(E_{x\sim i}(Y|x_i))$, and $V_{ij} = Var_{xij}\left(E_{x\sim ij}(Y|x_i, x_j)\right) - V_i - V_j$ And so on.

The $x_{\sim i}$ notation represents all variables except x_i. The variance decomposition provided in Equations (3&4) shows how the variance of the model output can be divided into terms attributed to each input as well as the interactions between them. All of these factors add up to the overall variance of the model output [44].

4.5.1 First-order index

The first order (S_i) Sobol index measures the proportion of Y's variation that is caused byx_i, also known as the main effect.

The relevance of the interaction between two input variables x_i and x_j measured using Sobol indices of second order. The Sobol indices of orders 3, 4, and so on can be calculated using the same technique. A direct variance-based measure of sensitivity called the "first-order ("main effect index") sensitivity index" S_i for the input factor x_i [44]**:**

$$S_i = \frac{V_i}{V(Y)} \quad = \quad \frac{V[E(Y|x_i)]}{V(Y)} \qquad (5)$$

Second-order Sobol index S_{ij} for the interaction between X_i andX_j [44]:

$$\mathrm{S_{ij}} = \frac{\mathrm{V_{ij}}}{\mathrm{V(Y)}} \quad = \frac{\mathrm{V\left[E\left(Y|x_i, x_j\right)\right] - V_i - V_j}}{\mathrm{V(Y)}} \qquad (6)$$

4.5.2 Total-order index

We can build a picture of the relevance of each variable in influencing the output variance using the indices provided above (S_i andS_{ij}). However, when the number of variables is large this necessitates the use of indices$2^n - 1$, which can be computationally intensive. As a result, a metric called the "Total-order index"(S_{Ti}) is used. This metric assessesX_i, contribution to output variance, which includes all variation caused by interactions with other input variables. It is given as follows:

$$(S_{T_i}) = S_i + S_{ij} + \cdots + S_n \qquad (7)$$

($S_{T_i}$are the sum of all the Sobol indices relative to X_i) [36, 45, 46].

4.6 Sobol-GSA Implementation

There are two key prerequisites for GSA implementation[46]:

1. **Sampling design**: a technique for organizing sample points in the multidimensional space of input variables. This consists of specifying the sensitivity analysis settings, i.e., setting the sample size "N" of the basic sample matrix and the number of variables "k" and constructing a vector with the parameter names, number of bootstrap replicas R, and finally, setting the confidence intervals to 0.95. The sample matrix will then be created using function ***Sobol matrices ()***. As seen in the R software platform **Appendix-I**, develop

sample designs that use random integers (type ="R"), which produce more accurate sensitivity indices (first and total-order indices) [46].

2. **Sensitivity estimators**: a formula for calculating the sensitivity measures described in the mathematical expression section above [46].

We need to know how the model output maps onto the model input space before calculate Sobol's indices. The SenSobol R package offers two methods to accomplish this goal: ***plot_ scatter () and plot multiscatter ()***. The first plots the model output PUE against each input variable (Fan_P vs. PUE, Acc_load vs. PUE, and so on) and displays the first-order effect, whereas the second plots the interactions between the PUE value and the input variables (second-order Sobol index). As seen in **Appendix-I**, allows the user to discover patterns that indicate sensitivity[46].

The scatter plots in Figure A-1 and A-2 shows the sensitivity of inputs and outputs, the more patterns there are, the higher the sensitivity. Scatter plots, on the other hand, do not always allow for the determination of which parameters have a combined influence on the model output. To acquire a better understanding of these interactions, the function ***plot multiscatter ()*** plots, displays interactions by colored patterns. Over-plotting may occur if the number of input variables is large, making it difficult to interpret the pattern. To solve this problem, we must concentrate on a limited number of factors here six input factors are selected[46].

In general, for GSA implementation, the initial stage is to identify the range of the inputs as well as their probability distributions. The second stage is to build and run a model, in this case, a PUE model. Finally, a Sobol sensitivity analysis has been performed, and simulation results are collected. Display scatter plots of results to investigate the relationship between inputs and outputs and quantify the impact level to analyze the contribution of data center features on energy performance [36].

Chapter 5 - Results and Discussions

This chapter describes the results and analysis of the root causes of the data center energy efficiency problem. The PUE analysis, RFR, and Sobol-GSA results are presented.

5.1 PUE Analysis

As defined in the previous chapter, the metric used to measure data center efficiency is PUE. It determines the ratio of the data center's overall energy consumption to the IT equipment's energy consumption[47], [48]. The ideal data center would have a PUE of 1.0, indicating that the IT equipment consumes all of the power entering the data center. Any value greater than 1.0 indicates that some of the total power is redirected to support systems such as cooling, lighting, and the power conditioning system. The greater the PUE value, the more power is consumed by support infrastructure, resulting in a less energy-efficient data center.

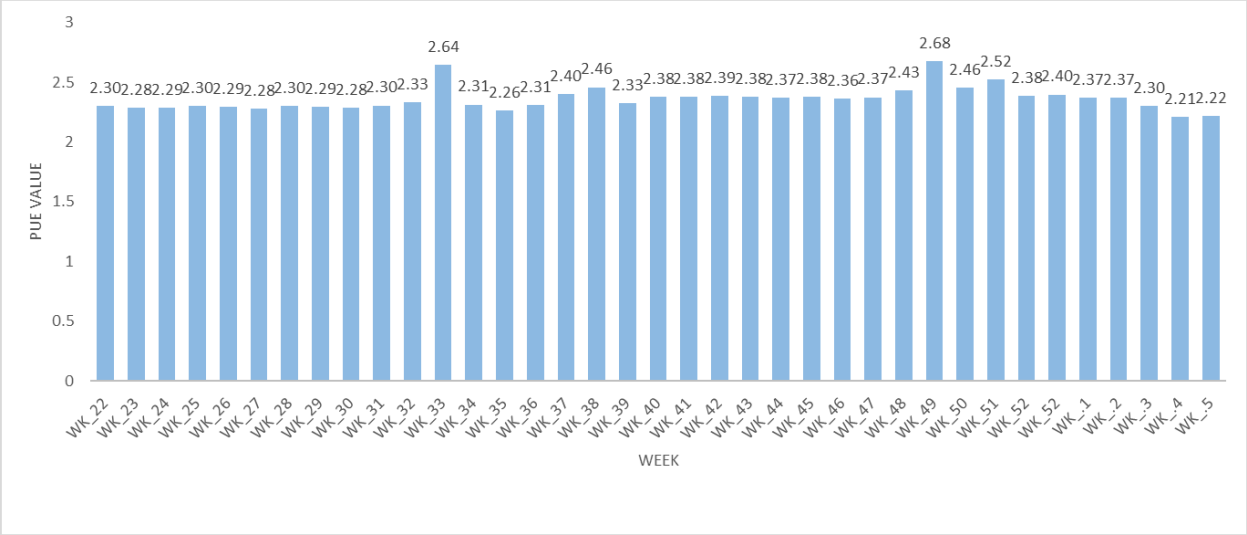

Figure 5-1 Legehar weekly PUE values

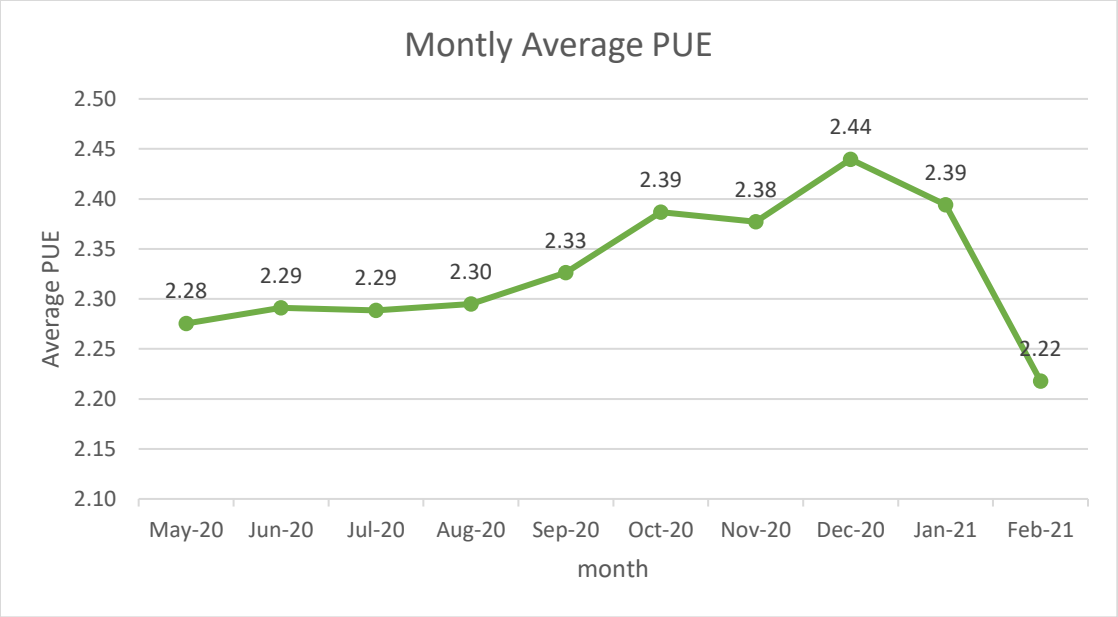

Figure 5-2 Monthly Average PUE

The figures 5-1 and 5-2 depict the weekly and monthly PUE values for the Legehar data center respectively. The total average PUE is 2.34, which is higher than the global standard and indicates that the data center's performance is inefficient. The PUE was computed using data gathered between May 25, 2020, and February 7, 2021.

5.2 Results of RFR

Using conventional ways to find the root cause of a problem, it can be challenging to figure out which feature influences a data center's energy efficiency. This thesis provides a hybrid method to the RCA technique that uses RFR and GSA to identify features that can be the source of the problem. To select features, the RFR-VarImp approach was used, as explained in the previous section, features that are redundant or irrelevant to the dataset are appropriately removed. We'll examine a smaller number of features that might be the source of the problem.

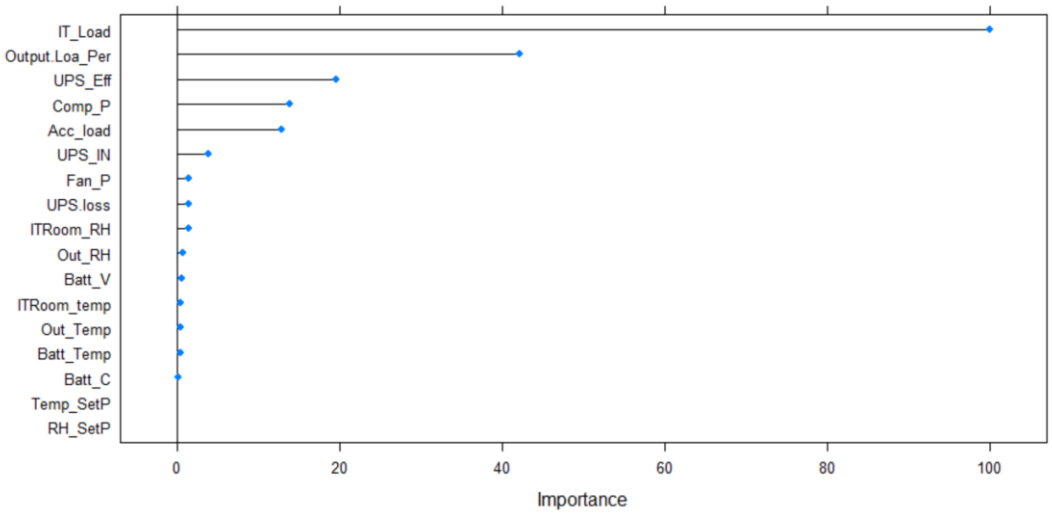

Figure 5-3 RFR VaImp result

After applying the model, we select features to find the most important driving factors of energy efficiency (PUE) for the next level. Figure 5-3 illustrates the RFR-VarImp feature selection method findings, which are energy efficiency features in data centers.

Table 5-1 List of significant and non-significant variables

No.	Variable	Description
Significant Variables		
1	IT_load	IT load
2	Output.Loa_per	UPS output load percentage
3	UPS_Eff	Efficiency of UPS
4	Comp_P	Air conditioner compressor power
5	Acc_load	Accessory load (lighting, Security camera, and access door)
6	UPS_In	Total input power to the data center
7	Fan_P	Air conditioner supply and return fan power
8	UPS_loss	UPS power loss
9	ITRoom_RH	IT room relative humidity
Non-Significant Variables		
10	Out_RH	Outdoor Relative humidity
11	Batt_V	Battery terminal voltage
12	ITRoom_temp	IT room temperature
13	Out_Temp	Outdoor temperature
14	Batt_Temp	Battery temperature
15	Batt_C	Battery charging current
16	Temp_SetP	Temperature set pint
17	RH_SetP	Relative humidity set point

As shown in Table 5-1 also RFR-VarImp feature selection to distinguish between significant and non-significant variables, assuming that importance levels below three are considered insignificant. The top-ranked features in the graph are the most significant contributors to data center energy efficiency. The Sobol-GSA method is used to explore further the important variables that have been selected. Because the RFR-VarImp feature selection technique lacks information on the degree of the impact level, as well as its individual and interactions with PUE, necessitating the use of Sobol-GSA.

5.3 SOBOL-GSA Result

In this subsection, the above-selected features that influence the PUE are considered in the following sensitivity analysis. The analysis consists of two scenarios, to see which parameter causes the largest changes in PUE. These two scenarios are first-order indices and total order indices. First-order indices are individual variable contributions with PUE and total order indices are the interaction between variables i.e., the sum of the first order and higher indices. Within each scenario, the parameter was changed simultaneously because our method is used SOBOL-GSA.

As described in the methodology chapter, the Sobol indices were calculated using Sobol sensitivity analysis, which is available in the sensitivity package of **R** software.

The first stage is to simulate the first and total order indices for each input feature. To determine the number of simulations using $N(2K + 2)$.

Where N is the sample size (a recommended N of 500–1000), and K stands for the number of model input parameters.

The N value in this study was 1000, and the K value was 9. As a result, equal to 20000.

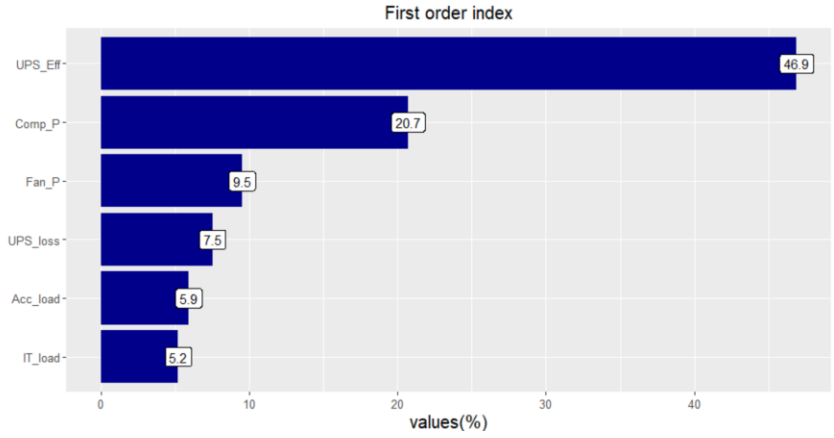

Figure 5-4 First order sensitivity result

The number of important parameters identified is less than the initial 17 features, with the RFR model identifying nine significant features as input for Sobol-GSA and the Sobol method identifying six parameters among the nine significant variables, namely UPS_Eff, Comp_P, Fan_P, UPS_ loss, Acc_load, and IT_load. As shown in Figure 5-4, in the case of the first-order effect, the parameter UPS_ Eff was discovered to be the most sensitive, accounting for 46.9 percent of the model output variability. Low sensitivity was seen in Acc_load and IT_load, with lower values for both the first order and total order.

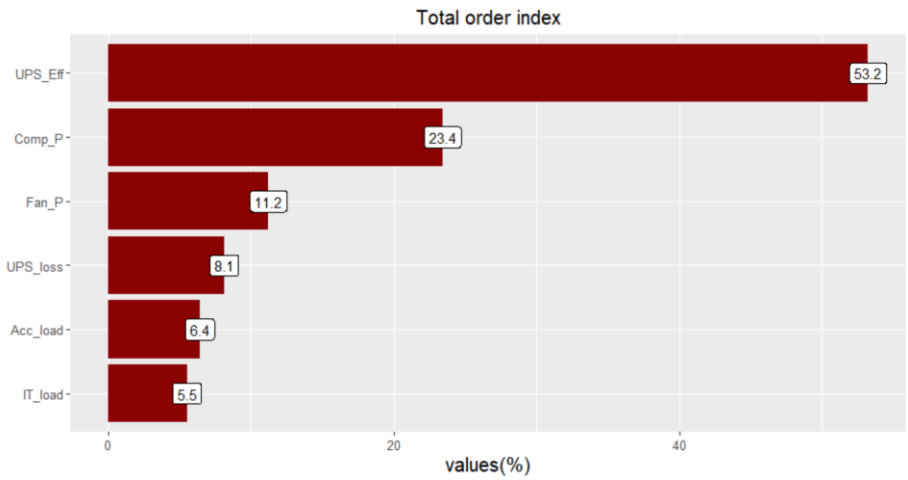

Figure 5-5 Total order sensitivity result

The difference between S_{Ti} and $S_i$reveals how much factor X_i has a role in the interactions. S_{Ti} is commonly larger than S_i. If $S_{Ti} = S_i$, the factor X_i does not interact with other variables, meaning that the component X_i is insignificant and can be adjusted to any value without impacting the model output variable's variance.

In the second scenario (total order index), figure 5-5, all six PUE model sensitivity parameters have S_{Ti} values greater thanS_i, the smallest difference between the first-order and total-order effects for individual parameters, as well as the difference between the sum of the first-order indices ($\sum S_i = 95.7$) and unity, which shows parameter interaction. It can also be observed that UPS and cooling system (the sum of Fan power and compressor power) parameters had the highest sensitivity, being the most important variables in data center energy inefficiency. Thus, varying UPS and cooling system variables had the biggest influence on the PUE value.

5.4 Discussion

For data center operators (owners), there is a challenge in identifying the root cause of the energy efficiency problem. This thesis reflected the data center energy inefficiency of Ethio-Telecom. Since there are various reasons for the inefficiency, it cannot be easily diagnosed manually. The reasons are individually and the interactions of data center components. In order to figure out the energy efficiency problem, RCA is required. There are various problem identification methods. Among the methods, combing RFR and Sobol-GSA is selected. The major contributor components for the data center low overall energy performance are the UPS and the cooling system.

5.4.1 UPS system

As described in the previous sections, all Ethio-Telecom data centers use utility power as their primary power source, a UPS battery, and diesel generators for backup power. The battery will provide backup power for IT devices in the case of a power outage. The primary influencing element of the data center energy efficiency problem was the UPS system. The reasons for the loss of electrical energy in UPS were twofold: the type of the UPS system, which is of the double-conversion type repeated conversions (AC-DC-AC), and the UPS load percentage (a poor load factor). According to UPS manufacturers, UPS efficiency curves show UPS load percentage affects UPS efficiency, which in turn influences the PUE value. The UPS load at the research location was between 40 and 45 percent, while the recommended load percentage for UPS is 92-96 percent [22]. The UPS is operating much less than the optimum output load percentage, which results in a poor UPS efficiency of 83-85 percent [4, 49].

From the definition of PUE, as the load on the UPS (IT load) increases, so does its efficiency, which is expected to improve data center efficiency because IT load and PUE are inversely proportional. However, as the IT load increases, so does the temperature generated in the IT room, causing the air conditioners to work for an extra time, consuming more energy. Therefore, energy-saving (efficiency) methods must take into account the interaction impact of various parameters.

Therefore, this study's analysis shows that UPS losses accounted for 9 percent to 11 percent of the total energy utilized by the data center by taking the optimum UPS efficiency of 92-96%.

For example, a data center with an average IT load of 4700 KWh and an average PUE of 2.34

$$PUE = \frac{Total\ Power}{IT\ load} \Rightarrow Total\ Power = PUE * ITload$$

Total power consumption will be 10998 KWh and 1210 kWh in electrical losses and will spend approximately birr 1197.00 on wasted electrical energy[7].

5.4.2 Cooling System

As previously mentioned, data centers with raised floors, dropped ceilings, and hot aisle and cold aisle arrangements are standard data center designs. This arrangement helps with efficient air distribution, which in turn helps reduce the energy consumption of the cooling system by improving air circulation, i.e., improving air conditioning cooling efficiency by protecting against air recirculation (mixing of hot and cold air in the IT room). However, the data center at the research location does not have a raised floor but was designed using the hot and cold Aisle arrangement, which means the cold inlets and hot discharge sides are arranged to face each other. In addition to this, the cooling system's efficiency is also a key factor and it mainly depends on whether the cooling system is air-cooled or water-cooled and up-flow or downflow. In our case the type of air conditioner is a UP flow type i.e., cold air is supplied at the top and return air (hot) at the bottom of the air conditioners, According to the principle of thermodynamic temperature moves from hotter to a cooler area. So the probability of air recirculation is high, which leads to increase air conditioners consuming more energy. Moreover, to the factors for the data center energy inefficiency discussed above, air-cooled cooling systems since they are outdoor temperature-dependent consume more energy than water-cooled cooling systems. In general, down flow and water-cooled cooling systems are more energy-efficient than up flow and air-cooled cooling systems.

Besides the air conditioner efficiency and the type of air conditioners the other factor contributing to the inefficient energy consumption of the data center was the total number of air conditioner units installed. The IT room consists of seven stand-alone air conditioners with various capacities, three additional air conditioner units were also installed for the battery and UPS rooms. An oversized cooling system was deployed in the data center. This was verified by examining the cooling capacity factor (CCF), i.e. the ratio of total installed cooling capacity to IT load. According to ASHRAE, one of the metrics to know the cooling system capacity is CCF. This ratio should be around 1.2times, or 120%[50]. The Legehar data center cooling system, on the other hand, is 1.37times or 137 %. This is mainly due to redundancy and the overestimation of the heat emission of IT equipment.

According to the findings of this study, the cooling system is a key contributor to data center energy efficiency. Previous data center energy efficiency studies have concentrated on the cooling system and the opportunity of using free cooling (using outdoor low temperature) to decrease PUE and improve the efficiency of the data center. However, the case study site data center environment (Addis Ababa) is not suitable to free cooling throughout the year. On average, the maximum temperature was 19.2°C and the low temperature was 10.8°C [51].

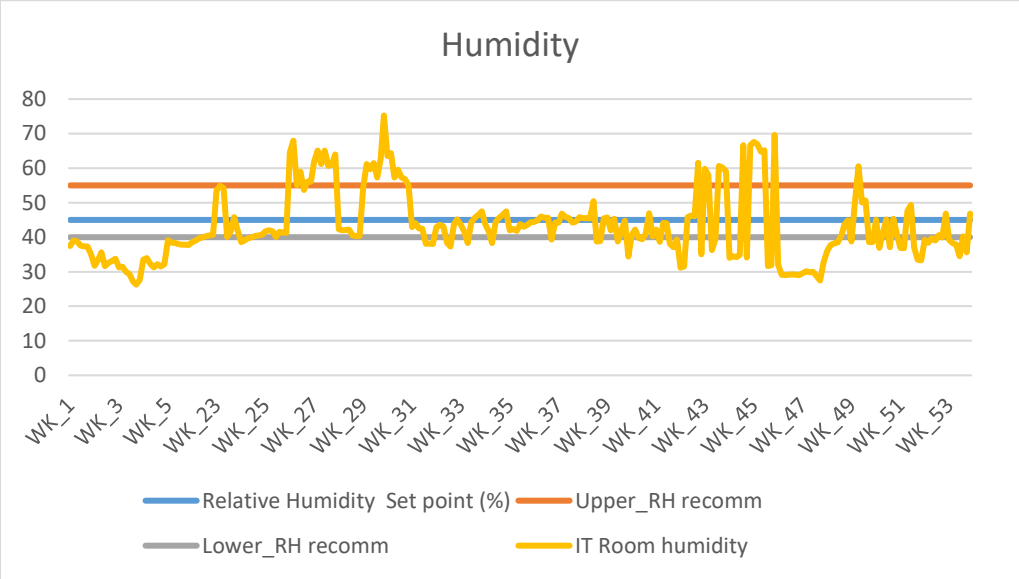

Figure 5-6 Legehar data center Relative humidity

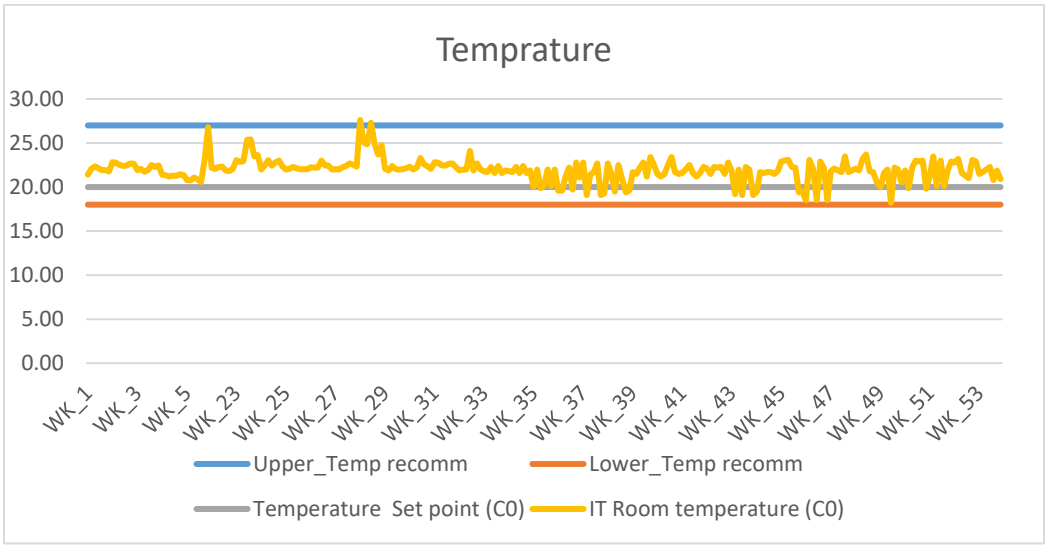

Figure 5-7 Legehar data center indoor temperature and set point

Figures 5-7 and 5-8 show humidity and temperature measurements from the Legehar data center. Even though the temperature is generally within the allowable range, humidity levels are higher than allowed by ASHRAE guidelines[27].

Finally, the RCA results reveal that the cooling system contributes significantly to energy inefficiency, which is due to the fact that the data center cooling system is not built on a raised floor and dropped ceiling, and the type of air-conditioning is also UP flow type, which leads to an air recirculation problem. This will increase the room temperature. As a result, air conditioners are run for longer periods of time. Consequently, energy consumption increased. In addition to this, the Legehar data center is configured to operate at a lower temperature (20°C) even if the operating temperature range as recommended by ASHRAE, i.e., from 18°C to 27°C (64°F to 81°F) [25]. Ethio-Telecom power & environment experts' justification for this was the need to operate devices within their "safe" operating range and to reduce data center failures due to temperature increase during mains power interruption.

Chapter 6 - Conclusion and Future works

This chapter presents the study's conclusions and future research directions are suggested.

6.1 Conclusion

In this study, two techniques, RFR and Sobol-GSA, were combined to identify the fundamental causes of data center energy inefficiency. The primary conclusion was that combining the RFR feature selection technique (VarImp) with the Sobol-GSA approach performed here indicates that out of the 17 initial model inputs, 6 are important for the considered model outputs moreover, helps in identifying the root causes of a problem. The biggest contributors to the energy efficiency problem are UPS and cooling systems. Following the identification of the problem's contributor, the final stage in the RCA stages is to provide solution recommendations that successfully prevent the problem from reoccurring.

IT room temperatures and RH should be carefully maintained from an energy-saving standpoint for optimal data center performance. Energy savings were realized in the free cooling study by taking advantage of the low outdoor temperature. However, environmental conditions in our country do not allow for the use of a free cooling system. The other option is to raise the set point of IT room temperature. Changing the set point in our data center from 20℃ to 22℃ will enhance overall data center performance while following ASHRAE guidelines and sustaining QoS.

Finally, the combination of RFR feature selection and GSA aids in finding the root causes of energy inefficiency in data centers. The main idea is to go through two stages. First, the RFR VarImp algorithm is utilized to prioritize feature importance and remove the feature with the lowest ranking. Second, SOBOL-GSA uses the selected features to find the most important contributors to the data center energy efficiency problem. The SOBOL-GSA computes first and total order effects were computed and parameters were ranked according to their contribution to the overall variance of the PUE function.

6.2 Future Works

Based on the findings of the study and the conclusions made, the study recommends that:

- The study is limited to the Legehar data center in Addis Ababa. However, more research on a broader scale that encompasses different geographical locations is needed to either validate or refine the conclusions of this study.
- This study investigates the aspects of energy efficiency based on energy and environmental-related datasets. The variables included in the study were not exhaustive and future research should be carried out to determine the effect of other variables which are not identified in the present study but affect data center efficiency, for example, computing equipment (Server) utilization.
- Furthermore, in order to ensure long-term efficient energy use in data centers, a regular energy performance tracking or energy management system is highly recommended. It can be useful in building an energy-use database for future energy efficiency studies and analyses if measurements are taken on a regular basis by providing real-time monitoring of mission-critical IT equipment and its supporting systems.

References

[1] A. Grishina, "DATA CENTER ENERGY EFFICIENCY ASSESSMENT BASED ON REAL DATA ANALYSIS," LUT University, 2012.

[2] C. Dumitrescu, A. Plesca, L. Dumitrescu, M. Adam, C. Nituca, and A. Dragomir, "Assessment of Data Center Energy Efficiency. Methods and Metrics," in *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, Iasi, Oct. 2018, pp. 0487–0492. doi: 10.1109/ICEPE.2018.8559745.

[3] M. Pärssinen, "Analysis and Forming of Energy Efficiency and GreenIT Metrics Framework for Sonera Helsinki Data Center HDC," Aalto University, 2016.

[4] M. Milad and M. Darwish, "UPS system: How can future technology and topology improve the energy efficiency in data centers?," in *2014 49th International Universities Power Engineering Conference (UPEC)*, Cluj-Napoca, Romania, Sep. 2014, pp. 1–4. doi: 10.1109/UPEC.2014.6934608.

[5] B. Whitehead, D. Andrews, A. Shah, and G. Maidment, "Assessing the environmental impact of data centres part 1: Background, energy use and metrics," *Building and Environment*, vol. 82, pp. 151–159, Dec. 2014, doi: 10.1016/j.buildenv.2014.08.021.

[6] G. Koutitas and P. Demestichas, "Challenges for Energy Efficiency in Local and Regional Data Centers," p. 33.

[7] "EEU New_Tariff_English_Version." EEU, 2019.

[8] Z. Song, X. Zhang, and C. Eriksson, "Data Center Energy and Cost Saving Evaluation," *Energy Procedia*, vol. 75, pp. 1255–1260, Aug. 2015, doi: 10.1016/j.egypro.2015.07.178.

[9] Shally, Sanjay Kumar Sharma, Sunil Kumar, "Measuring Energy Efficiency of Cloud Datacenters," *IJRTE*, vol. 8, no. 3, pp. 5428–5433, Sep. 2019, doi: 10.35940/ijrte.B3548.098319.

[10] N. Lei and E. Masanet, "Statistical analysis for predicting location-specific data center PUE and its improvement potential," Energy, vol. 201, p. 117556, Jun. 2020, doi: 10.1016/j.energy.2020.117556.

[11] Jim Gao, "Machine Learning Applications for Data Center Optimization." Google, 2013.

[12] Eduard Oró Jaume Salom, "Data Centre Overview." Catalonia Institute for Energy Research – IREC, SpainSpain.

[13] D. Kliazovich, P. Bouvry, F. Granelli, and N. L. S. da Fonseca, "Energy Consumption Optimization in Cloud Data Centers," in *Cloud Services, Networking, and Management*, N. L. S. da Fonseca and R. Boutaba, Eds. Hoboken, NJ: John Wiley & Sons, Inc, 2015, pp. 191–215. doi: 10.1002/9781119042655.ch8.

[14] A. Inc, "Data Center Infrastructure Resource Guide," p. 64.

[15] C. Belady *et al.*, "PUE™: A COMPREHENSIVE EXAMINATION OF THE METRIC," p. 83, 2012.

[16] R. Ascierto, "Uptime Institute global data center survey 2020," p. 32, 2020.

[17] "High-Performance Computing Data Center Power Usage Effectiveness," *National Renewable Energy Laboratory (NREL)*. https://www.nrel.gov/computational-science/measuring-efficiency-pue.html (accessed Nov. 17, 2021).

[18] X. Wang, "Optimal DC Power Distribution System Design for Data Center with Efficiency Improvement," University of Wisconsin Milwaukee, 2014.

[19] S. Jayathilake, "ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPROACH TO MEASURING ENERGY CONSUMPTION IN DATA CENTRE FACILITIES," University of East London, 2019.

[20] M. Dayarathna, Y. Wen, and R. Fan, "Data Center Energy Consumption Modeling: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 732–794, 2016, doi: 10.1109/COMST.2015.2481183.

[21] G Bennett, "UPS Systems – What Does On-Line Double Conversion Mean?," *Standby Systems*, Apr. 2021. https://standbysystems.co.za/ups-systems-what-does-on-line-double-conversion-mean/ (accessed Aug. 26, 2021).

[22] Tewodros Kibatu, "Recurrent Neural Network-based Base Transceiver Station Power System Failure Prediction," Addis Ababa University, 2019.

[23] C. Pattinson, "Critical Issues for Data Center Energy Efficiency," *In Green Information Technology*, pp. 223–248, Mar. 2015.

[24] G. Yogesh, "Fundamentals_of_Data_Centre_Part_1."

[25] J. Kaivosoja, "Green Data Centers: Evaluating Efficiency," Aalto, 2016.

[26] ASHRAE Technical Committee (TC) 9.9, "Data Center Power Equipment Thermal Guidelines and Best Practices." June_2016_REVISED.

[27] "Root Cause Analysis (RCA)," *Toolshero*. https://www.toolshero.com/problem-solving/root-cause-analysis-rca/ (accessed Aug. 25, 2021).

[28] Tesfaye Fetene, "Root Cause Analysis of Base Station Outage using Bayesian Network for Addis Ababa," Addis Ababa University, 2020.

[29] Kulcsar T., Balaton M., Nagy L., and Abonyi J., "Feature selection based root cause analysis for energy monitoring and targeting," *Chemical Engineering Transactions*, vol. 39, pp. 709–714, Aug. 2014, doi: 10.3303/CET1439119.

[30] Mesfin Geremew, "Root Cause Analysis of Mobile Site Outage Using Bayesian Network: the Case of ethio telecom," AAU, Addis Ababa, Ethiopia, 2018.

[31] Tim Josefsson, "Root-cause analysis through Machine learning in the cloud," Institutionen för informationsteknologi Department of Information Technology, 2017.

[32] C. Aldrich, "Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework," *Minerals*, vol. 10, no. 5, p. 420, May 2020, doi: 10.3390/min10050420.

[33] "Sensitivity Analysis: Types, Methods, and Use | Wikiaccounting," *OUR BLOG*. https://www.wikiaccounting.com/sensitivity-analysis (accessed Jun. 09, 2021).

[34] X. Zhang, M. Trame, L. Lesko, and S. Schmidt, "Sobol Sensitivity Analysis: A Tool to Guide the Development and Evaluation of Systems Pharmacology Models," *CPT Pharmacometrics Syst. Pharmacol.*, vol. 4, no. 2, pp. 69–79, Feb. 2015, doi: 10.1002/psp4.6.

[35] W. Tian, "A review of sensitivity analysis methods in building energy analysis," *Renewable and Sustainable Energy Reviews*, vol. 20, pp. 411–419, Apr. 2013, doi: 10.1016/j.rser.2012.12.014.

[36] "Understanding and comparisons of different sampling approaches for the Fourier AmplitudesSensitivityTest(FAST),"*ScienceDirect*.https://www.sciencedirect.com/science/article/abs/pii/S0167947310002756 (accessed Jul. 09, 2021).

[37] "WikipediaMAE,"*Meanabsoluteerror*.https://en.wikipedia.org/wiki/Mean_absolute_error (accessed Aug. 23, 2021).

[38] "Wikipedia RMSE." https://en.wikipedia.org/wiki/Root-mean-square_deviation (accessed Aug. 23, 2021).

[39] "Wikipedia rsquare." https://en.wikipedia.org/wiki/Coefficient_of_determination (accessed Aug. 23, 2021).

[40] A. Polewko-Klim, "Sensitivity analysis based on the random forest machine learning algorithm identifies candidate genes for regulation of innate and adaptive immune response of chicken," p. 14.

[41] A. Detzner and M. Eigner, "Feature selection methods for root-cause analysis among top-level product attributes," *Qual Reliab Engng Int*, vol. 37, no. 1, pp. 335–351, Feb. 2021, doi: 10.1002/qre.2738.

[42] "Wikipdia Variance-based sensitivity analysis," *Variance-based sensitivity analysis*. https://en.wikipedia.org/wiki/Variance-based_sensitivity_analysis, (accessed Aug. 24, 2021).

[43] A. Saltelli, Ed., *Global sensitivity analysis: the primer*. Chichester, England ; Hoboken, NJ: John Wiley, 2008.

[44] L. Brevault, M. Balesdent, N. Berend, and R. L. Riche, "Comparison of different global sensitivity analysis methods for aerospace vehicle optimal design," p. 12.

[45] A. Puy, S. L. Piano, A. Saltelli, and S. A. Levin, "sensobol: an R package to compute variance-based sensitivity indices," *arXiv:2101.10103 [stat]*, Apr. 2021, Accessed: Oct. 08, 2021. [Online]. Available: http://arxiv.org/abs/2101.10103

[46] ThegreenGridAsso.,"green_Data_Center_Power_Efficiency_Metrics_PUE_and_ DCiE."

[47] Dr. Aparna S. Varde , Dr. Stefan Robila and Dr. Michael P. "green_data_centers_for_sustainability.", New Jersey, USA, 2011.

[48] N. Rasmussen, "Electrical Efficiency Modeling for Data Centers White Paper 113." 2011.

[49] UPSite Technology, "CCF Calculator." https://www.upsite.com/resources/ccf-calculator/ (accessed Sep. 15, 2021).

[50] "Addis Abeba Historical Weather, ET," *World weather*. https://www.worldweatheronline.com/addis-abeba-weather-history/et.aspx (accessed Aug. 25, 2021).

A-APPENDIX – I

This appendix-I presents the GSA implantation proposed in this thesis. When we implement GSA, we need to have two basic requirements[46]:

1. Sample matrix design

```
N = 1000 # sample size
params = c("Comp_P","Fan_P", "Acc_load", "IT_Load",
           "UPS_loss","UPS_In","UPS_Eff",
          "Output.Loa_per","Out_RH") #create a vector with the parameters' name
matrices = c("A", "B", "AB", "BA")
#first = "sobol"
#total = "saltelli"
first = total = "azzini"
order = "second"
R = 10^ 3 #bootstrap replicas
type = "norm"
conf = 0.95 # confidence interval
```

2. Sensitivity estimator

Once the sample matrix has been constructed, we can now run our model using the following codes.

$$R > mat < -sensobol :: sobol\_matrics(matrics = matrics, N = N, params = params, order = order, typr = "R")$$

$$R > PUE = PUE\_run(mat)$$

$$R > plot\_scatter(data = mat, N = N, Y = PUE, params = params)$$

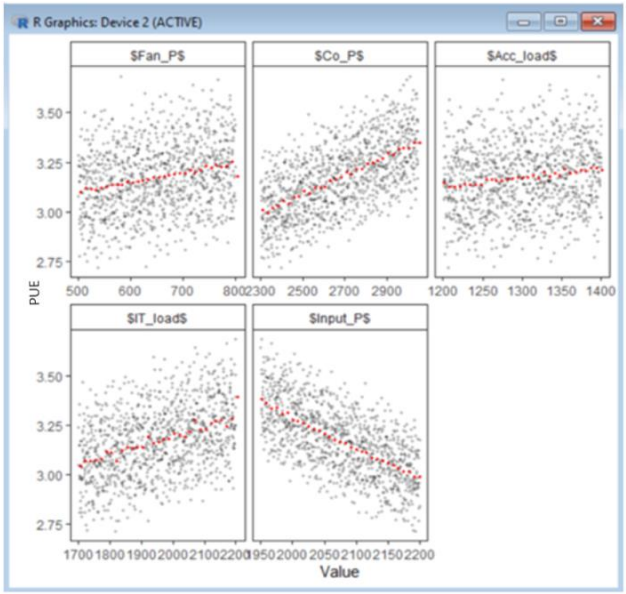

Figure A-1 scatter plot

$$R > Plot\_multiscatter(data = mat, N = N, Y = PUE, parms = params)$$

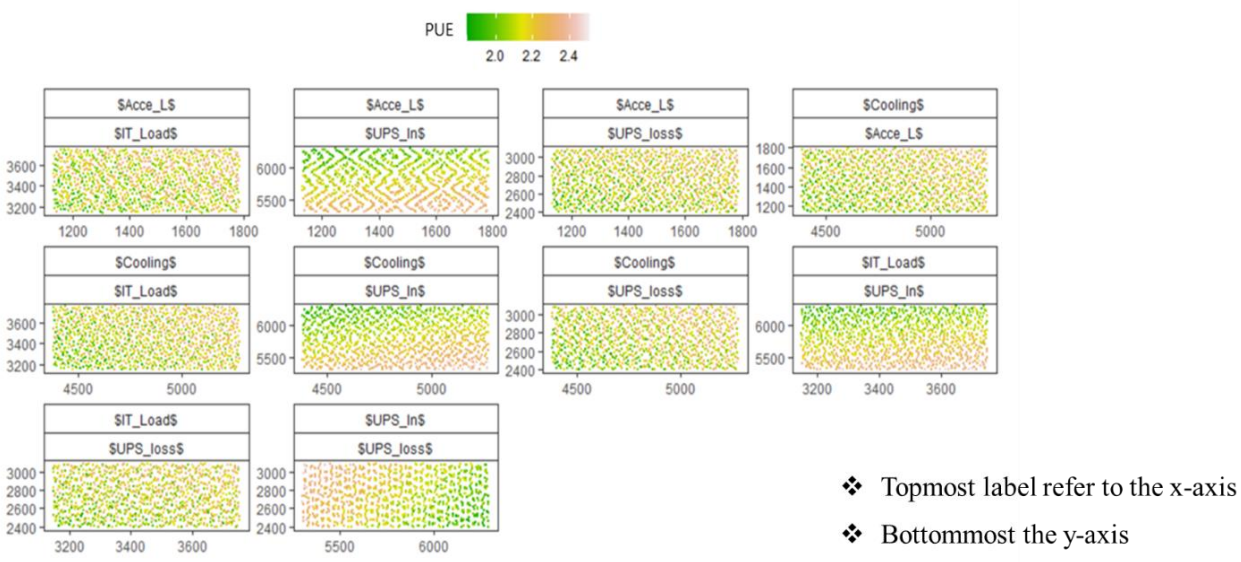

Figure A-2 Multi-scatterplot matrix of pairs of model inputs for the PUE function

For example, the results of interaction between IT_Load and UPS_In given the colored pattern of the (IT_Load, UPS_In) plane: the lowest values of the model output are concentrated on the top of the (IT_Load, UPS_In) input space. In order to know assessing the exact weight of this second-order interaction, the computation of second-order indices would be required.

The final step is to compute the Sobol indices. To obtain the Sobol indices in this specific case, run the code below and the result is described in Chapter Five.

$$R > ind <- sobol\_indices(y = PUE, N = N, params = params, boot = TRUE, R = R, type = type, conf = conf)$$

B- APPENDIX – Manuscript

Data Center Energy Inefficiency Root Cause and Sensitivity Analysis: The case of ethio Telecom Legehar Data Center

Zerihun Tesfaye † and Dereje Hailemariam‡
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
Email: zer.tesfaye65@gmail.com† and dereje.hailemariam@aait.edu.et‡

***Abstract*– Data center is the cornerstone in modern IT infrastructures. As the usage of IT grows, so does the volume of data processing and stored in data centers. As a result of this growth, the energy consumption of data center has increased significantly. The aim of this study was to identify the major contributors for the energy efficiency problem, as well as to quantify the magnitude of the influence and interaction between variables by combining Random Forest Regression (RFR) feature selection method and Sobol-Global Sensitivity analysis (Sobol-GSA). This study analyses 37 weeks (nine months) of energy and environmental data and computes Power Usage Effectiveness (PUE). PUE is the ratio of total power entering a data center to the amount of power utilized by IT equipment. The current average PUE value of Legehar data center is 2.34, meaning that more than half of electrical energy consumed by the "support equipment", such as power supply devices, UPS, cooling system and lighting.**

Key words: - Data center, PUE, Sobol-GSA, Random Forest

I. INTRODUCTION

Data center is the cornerstone of today's information age. As more people use Information Technology (IT), more data (e.g. photos, movies, financial transaction data, and so on), is generated leading to a rise in the amount of data processed, stored, and transported. As a result, data centers' energy usage has increased[1]. To function more effectively and reliably, IT equipment's require the support and interaction of various dependable systems such as electrical systems (power source) and mechanical Systems (cooling systems).

literatures have used different approaches to determine the fundamental cause of the energy efficiency problem in data centers. In[2] the major energy consume equipment and its factors are cooling and ventilation system, consumes about 40% of the total energy and the two factors for this energy consumptions are air flow management and climate condition (data center location). The authors proposed solution for data center energy efficiency which is ventilation and air flow management with vertically placed server racks have good efficiency than horizontally placed servers by providing efficient heat transfer with reduced air flow pressure drop and contribute to data center energy saving. Data center location selection also one of the significant procedures for reliable operation of data center to reduce energy cost and improve the performance of data centers. And the paper in [3] is a study on data center energy efficiency using statistical analysis to predict location-specific data center PUE and its improvement. The method utilized in this study is based on Sobol's method for this analysis taking climate variables and energy system parameters as inputs. the results suggest that focusing on reducing key input parameters is most important for reducing uncertainties in PUE values. Climate variables and UPS efficiencies are the most important parameters.

The main goal of this study is to identify the most essential factors that influence energy inefficiency and estimate the impact level of the features based on data collected from ethio telecom in Legehar data center. For that, we have used by combining Random Forest Regression (RFR) feature selection with Sobol-Global Sensitivity Analysis (Sobol-GSA). Our main contributions are:

- Several previous studies on energy efficiency have relied on free cooling while taking environmental advantages into consideration. This work, on the other hand, contributes to filling gaps by providing a statistical analysis of a PUE-

based environmental and energy-related model.

- It varies in a number of ways from prior research, combining RFR feature selection with Sobol-GSA which takes into account variables interaction effects and finds the most critical factors for energy efficiency problem. Previous research has mostly focused on local sensitivity analysis that is, analyzing a single variable at a time ignoring the interaction between variables.

The rest of the paper is organized as follows. In Section II, PUE calculations, and a brief discussion on the model formulation and Sobol mathematical explanation are described. In Section III. Explanation about the dataset used for the research and procedures followed in developing the analysis are presented in detail finally, in Section IV presents analysis results and conclusion of the research work and discussion on future research work directions.

II. *MODELLING COMPONENTS OF DATA CENTER*

The dataset for this study was obtained from ethio telecom NetEco Power and Environment Monitoring System. It offers statistics on 17 energy and environmental features for 37 weeks, every 5 minutes. The overall process for the study consists of the following steps:

1. Analyze the PUE level of the data center.
2. Feature selection based on RFR to identify major contributor for energy inefficiency.
3. Sobol-GSA to quantify the impact level of each parameter. It contains first order (main Effect) and total order indices.

The metric used to measure data center efficiency is PUE. It determines the ratio of the data center's overall energy consumption to the IT equipment's energy consumption. The ideal data center would have a PUE of 1.0, indicating that the IT equipment consumes all of the power entering the data center. Any value greater than 1.0 indicates that some of the total power is redirected to support systems such as cooling, lighting, and the power conditioning system. The greater the PUE value, the more power is consumed by support infrastructure, resulting in a less energy-efficient data center[4].

$$PUE = \frac{Total\ Power}{IT\ equipment\ Power} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$Total\ Power = Cooling\ Powe + UPS_{loss} + Accessory_{Load} + IT\_load$$

$$Cooling\ Power = Fan\_Power + Compressor\_Power$$

Total power is the total power provided to the data center, which includes energy for operating IT equipment, cooling infrastructure, energy losses in Uninterruptible Power Supply (UPS), and accessory load. The cooling usage is further subdivided into a supply and return fan of air conditioners, as well as compressor power[5].

Root Cause Analysis (RCA) is a problem-solving technique that aims to identify the source of problems[6].

There are two major types of RCA: data-driven RCA and non-data-driven RCA. Data-driven RCA is based on the analysis and interpretation of historical (recorded) data[7]. Using ML methods such as Decision Tree (DT), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and RFR. Whereas non-data-driven RCA is focused on qualitative analysis. The coefficient of determination (R-squared), mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate method selection. As a result, this study used a feature selection-based RFR and Sobol-GSA techniques for determining the underlying cause of energy inefficiency.

Table 2 ML algorithm comparison

Algorithm	MAE	RMSE	Rsquared
k-Nearest neighbours	0.01307494	0.01719802	0.9382102
Random Forest Regression	**0.01096575**	**0.01531134**	**0.9460464**
Decision Tree	0.01685037	0.02681028	0.8676008
Support Vector Machine	0.04630664	0.05396108	0.8122725

The feature is selected based on three repeats of 10-fold cross-validation, each iteration divides the training data into ten pieces, which are then combined to generate ten different samples. After that, each sample's RFR model is created, and the important variables (VarImp) for each variable is gathered. The variables with highest VarImp are chosen. These variables are applied to construct Sobol-GSA to quantify the impact level of features[8].

III. RESULT ANALYSIS AND DISCUSSION

The "support infrastructure," which includes power supply devices (UPS and PDU), cooling systems, and accessory loads, consumes a significant amount of electrical energy in a data center. PUE

tells how much energy IT equipment uses and how much energy is overhead.

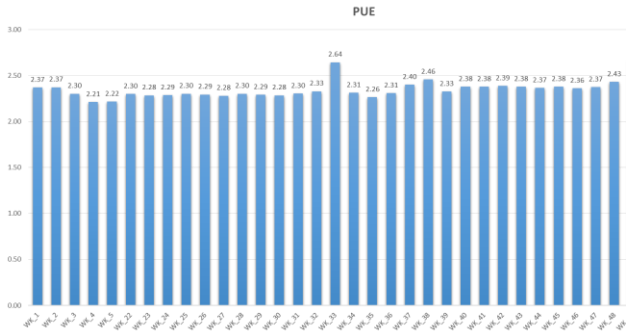

Figure 3 Weekly PUE value

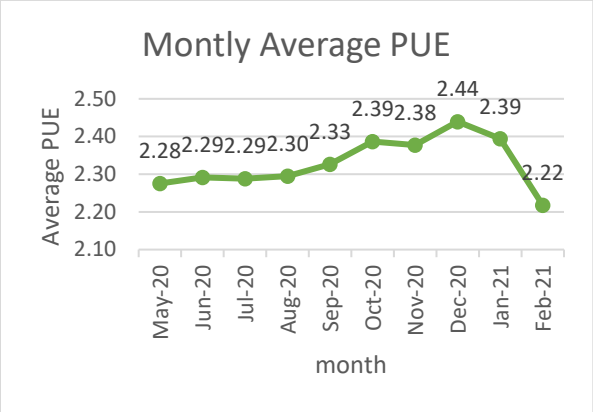

Figure 4 Average monthly PUE

The figures 1 and 2 depict the weekly and monthly PUE values for the Legehar data center respectively. The total average PUE is 2.34, which is higher than the global standard and indicates that the data center's performance is inefficient. The PUE was computed using data gathered between May 25, 2020, and February 7, 2021.

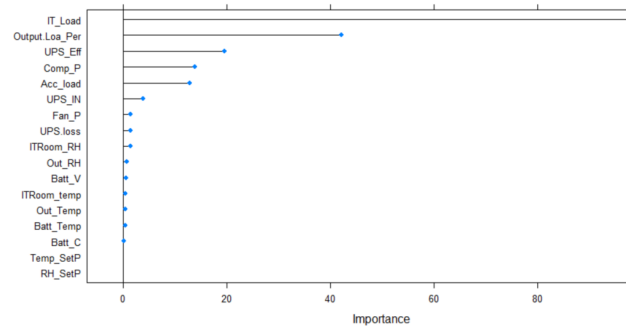

Figure 5 RFR VarImp Result

According to Figure 3 above the most important variables are the top nine features which includes: IT_load, UPS system (UPS Efficiency and output load percentage), Cooling system (Compressor and fan power consumption), and accessory load consumption (Battery room air conditioners, lighting, door access and surveillance).

The above-selected features that influence the PUE are considered in the following sensitivity analysis. The analysis consists of two scenarios, to see which parameter causes the largest changes in PUE.

Sensitivity analysis is carried out within predefined boundaries that are influenced by one or more input factors. It is sometimes referred to as a "what-if analysis". There are two major types of sensitivity analysis methods: Local Sensitivity Analysis (LSA) and Global Sensitivity Analysis (GSA). One variable is changed while all other variables remain constant, and all variables are changed at the same time respectively. The primary disadvantage of LSA method is that interactions between input variables are not taken into account, whereas GSA is more concerned with the effects of uncertain inputs throughout the whole input space[9].

There are several types of GSA methodologies and tools such as: Morris, Fourier Amplitude Sensitivity Test (FSAT) and Sobol. SALib in Python and R packages (sensobol, FAST, Morris) are the two suggested applications. However, in this study, the R senssobol package is utilized. The Sobol technique is a variance-based method for decomposing the uncertainty of outputs in relation to their corresponding inputs. The first order (Main effect) and total effects are the two major sensitivity metrics utilized in this method. The first-order effects take into account the main effects of the output changes caused by the related input. Total effects account for the total contributions to output variance due to the associated input, which includes both main effect and interactions [10].

Any model may be viewed as a function $y = f(x)$, where X is a vector of n model inputs $[x_1, x_2, x_3, x_4 \cdots x_n]$ and y is output vector $y = [y_1, y, y_3, \cdots y_m]$ related by:

$$y = f(x)$$
$$= f(x_1, x_2, x_3, x_4 \cdots x_n) \quad \cdots\cdots\cdots \quad (1)$$

The variance decomposition methods for sensitivity analysis consist in a decomposition of the variance of the output into a sum of contributions due to the input factors and their interactions.

$$f(x) = f_0 + \sum_{i=0}^{n} f_i(x_i) + \sum_{i<j}^{n} f_{ij}(x_i, x_j) + \cdots + f_{1\ldots k}(x_{1,\ldots}, x_k) \quad \cdots\cdots\cdots (2)$$

Where f_0 is a constant and f_i is a function ofx_i, f_{ij} a function of x_i and x_j

$$f_0 = E(Y)$$
$$f_i(x_i) = E(Y|x_i) - f_0$$
$$f_{ij}(x_i, x_j) = E(Y|x_i, x_j) - f_0 - f_i - f_j$$

From which it can be seen that f_i is the effect of varying x_i alone (known as the main effect of x_i) or first order, and f_{ij} is the effect of varying x_i and x_j simultaneously, additional to the effect of their individual variations. This is known as a second-order interaction.

Based on the functional decomposition Eq. (2), Sobol introduces the Sobol indices to quantify the partition of the output variance by using the decomposition of the variance [11], we have:

$$V(Y) = \sum_{i=1}^{n} V_j(Y) + \sum_{i<j}^{n} V_{ij}(Y) + \cdots + V_{123\ldots n}(Y)$$

$$Var(Y) = \sum_{i=1}^{n} V_i + \sum_{i<j}^{n} V_{ij} + \cdots + V_{12\ldots n}$$

Where $V_i = Var_{xi}\left(E_{x\sim i}(Y|x_i)\right)$,

$$V_{ij} = Var_{xij}\left(E_{x\sim ij}(Y|x_i, x_j)\right) - V_i - V_j$$

The $x_{\sim i}$ notation indicates the set of all variables except x_i. The variance of the model output may be decomposed into components attributed to each input, as well as the interaction effects between them, using the above variance decomposition. The overall variance of the model output is equal to the sum of all terms. The first order Sobol index quantifies the part of variance of Y due to x_j, refereed as main effect. The second order Sobol indices allow to measure the importance of the interaction between two input variables x_i and x_j[12].

First order Sobol index S_i for the input factor x_i

$$S_i = \frac{V_i}{V(Y)} = \frac{V[E(Y|x_i)]}{V(Y)}$$

Second order Sobol index S_{ij} for the interaction between X_i andX_j : $S_{ij} = \frac{V_{ij}}{V(Y)} = \frac{V[E(Y|x_i,x_j)]-V_i-V_j}{V(Y)}$

Total-effect index (ST_i) are the sum of all the Sobol indices relative to X_i . Using the S_i and S_{ij} indices one can build a picture of the importance of each variable in determining the output variance[12]

$$(ST_i) = S_i + S_{ij} + \cdots + S_n \ .$$

The sensitivity of a significant contributor to data center energy inefficiency is depicted in Figure 4 and 5. To determine the key factors impacting data center energy usage, a Sobol-GSA approach was used. As a result, first-order effects (S_i) can account for more than 46.9 percent UPS_Eff and 30.2 percent of the variance in and cooling system (Comp_P and Fan_p) respectively. The total order (S_{Ti}) Sobol index also shows the interaction of features, with PUE. S_{Ti} has a consistent trend as the S_i, that is, the ranking of parameters by S_i values agreed well with those from S_{Ti} values.

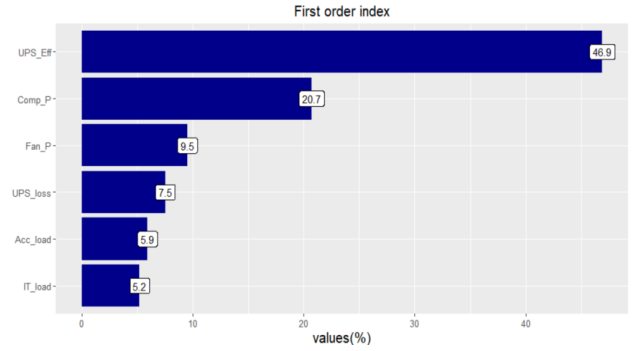

Figure 6 first order sensitivity result

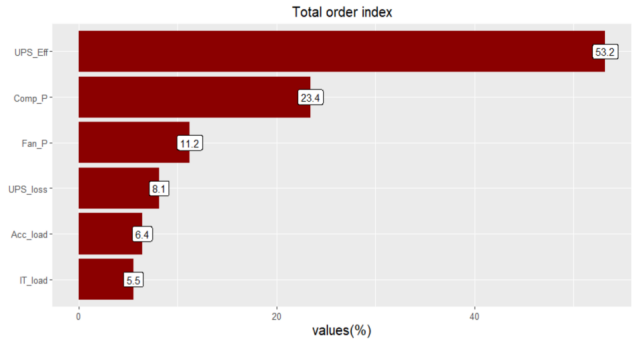

Figure 7 Total order sensitivity result

S_{Ti} slightly greater than S_i values for every parameter.

$ST_{UPS\_Eff} - S_{UPS\_Eff}$ 6.3% suggests that this value was the result of interaction between UPS_Eff and other parameters)/ When changes in the energy consumption of the IT load, i.e. related to load percentage and UPS efficiency, a larger difference is observed. The figure clearly shows that varying these variables had the biggest influence on the PUE.

IV. CONCLUSION AND FUTURE WORK

The data collected from the NetEco monitoring system help in the analysis of the overall efficiency of data center. The analysis reveals that the data center has a low overall energy performance. The major cause for the data center's energy inefficiency is the technology of the UPS system, which is of the double-conversion type. Another source of UPS inefficiency is the UPS load percentage. Because our average load level is 43 percent, UPS efficiency is around 83 percent lower than the suggested range of 92-96 percent by UPS and data center operators. Finally, cooling system energy consumption also significant. Improving UPS efficiency and cooling system can help improve PUE performance in data center[13].

The combination of RFR feature selection and global sensitivity analysis aids in finding the root causes of energy inefficiency in data centers. The main idea is to go through two stages. First, the RFR VarImp algorithm is utilized to prioritize feature importance and remove the feature with the lowest ranking. Second, SOBOL-GSA uses the selected features to find the most important contribution to the data center energy efficiency problem. The SOBOL-GSA computes the first order individual effect of features and the total order impact of feature interaction.

In the future, to improve the research's quality, the generalizability of its results and to confirm or refine the findings of this study, a larger scale that includes geographical areas throughout Ethiopia is required. This study investigates the aspects of energy efficiency based on energy and environmental-related datasets. The variables included in the study were not exhaustive and future research should be carried out to determine the effect of other variables which are not identified in the present study but affect data center efficiency, for example, computing equipment (Server and storage devices) utilization.

Furthermore, a frequent energy performance tracking or energy management system is strongly suggested to ensure long-term efficient energy consumption in data centers. It can be useful in establishing an energy-use record for future energy efficiency by enabling real-time monitoring of mission-critical IT equipment and its supporting systems.

REFERENCES

[1] A. Grishina, “DATA CENTER ENERGY EFFICIENCY ASSESSMENT BASED ON REAL DATA ANALYSIS,” LUT University, 2012.

[2] Z. Song, X. Zhang, and C. Eriksson, “Data Center Energy and Cost Saving Evaluation,” *Energy Procedia*, vol. 75, pp. 1255–1260, Aug. 2015, doi: 10.1016/j.egypro.2015.07.178.

[3] N. Lei and E. Masanet, “Statistical analysis for predicting location-specific data center PUE and its improvement potential,” *Energy*, vol. 201, p. 117556, Jun. 2020, doi: 10.1016/j.energy.2020.117556.

[4] “THE GREEN GRID DATA CENTER POWER EFFICIENCY METRICS: PUE AND DCiE.” The green grid, 2007.

[5] M. Pärssinen, “Analysis and Forming of Energy Efficiency and GreenIT Metrics Framework for Sonera Helsinki Data Center HDC,” Aalto University, 2016.

[6] Kulcsar T., Balaton M., Nagy L., and Abonyi J., “Feature selection based root cause analysis for energy monitoring and targeting,” *Chem. Eng. Trans.*, vol. 39, pp. 709–714, Aug. 2014, doi: 10.3303/CET1439119.

[7] A. Inc, “Data Center Infrastructure Resource Guide,” p. 64.

[8] A. Polewko-Klim, “Sensitivity analysis based on the random forest machine learning algorithm identifies candidate genes for regulation of innate and adaptive immune response of chicken,” p. 14.

[9] “Saltelli - 2008 - Global sensitivity analysis the primer.”

[10] A. Puy, S. L. Piano, A. Saltelli, and S. A. Levin, “sensobol: an R package to compute variance-based sensitivity indices,” *ArXiv210110103 Stat*, Apr. 2021, Accessed: Oct. 08, 2021. [Online]. Available: http://arxiv.org/abs/2101.10103

[11] X. Zhang, M. Trame, L. Lesko, and S. Schmidt, "Sobol Sensitivity Analysis: A Tool to Guide the Development and Evaluation of Systems Pharmacology Models," *CPT Pharmacomet. Syst. Pharmacol.*, vol. 4, no. 2, pp. 69–79, Feb. 2015, doi: 10.1002/psp4.6.

[12] A. Saltelli, Ed., *Global sensitivity analysis: the primer*. Chichester, England ; Hoboken, NJ: John Wiley, 2008.

[13] G Bennett, "UPS Systems – What Does On-Line Double Conversion Mean?," *Standby Systems*, Apr. 2021. https://standbysystems.co.za/ups-systems-what-does-on-line-double-conversion-mean/ (accessed Aug. 26, 2021).