Leveraging Big Data solutions and APRIORI algorithms for environmental sustainability in port cities a case study of Mombasa.

By

Chann Isaac.

Admission No. 145110

Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in Data Science & Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

Expected month of graduation, year

2024

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's Name: Chann Otieno Isaac

Sign:_____ Date ___8th April 2024_

Approval

The thesis of Chann Otieno Isaac was reviewed and approved for examination by the following:

Dr. Odipo, Victor Onyango

Institute of Mathematical Sciences,

Strathmore University

Sign:_____ Date 06/04/2024

Abstract

Big data has been dubbed "the next frontier for innovation, competition, and productivity," and is broadly described as "society's ability to harness information in creative ways to produce meaningful insights or commodities and services of significant value". The main research question for this study is: What are the potential contributions of spatial association mining based on the APRIORI algorithm for improving environmental sustainability in port cities in the global South?

The study will employ a mixed-methods approach, which will include both qualitative and quantitative data collection and analysis techniques. Qualitative data will be collected through in-depth interviews with relevant stakeholders, such as government agencies, port authorities, and environmental organizations. Quantitative data will be collected through the analysis of satellite imagery and existing databases of environmental and shipping data.

The APRIORI algorithm will be used to identify the significant associations between different environmental sustainability variables in the port cities. The expected findings from this research will provide evidence for the potential of spatial association mining based on the APRIORI algorithm for addressing environmental sustainability issues in port cities in the global South. The findings will also inform the development of more effective and efficient environmental sustainability policies and practices for these cities.

This research will contribute to the field of environmental sustainability by demonstrating the potential of Big Data solutions for addressing environmental challenges in the global South. The results will provide a basis for further research on the use of data-driven approaches for environmental sustainability in other regions and sectors.

Table of Contents

List of Figures

List of Tables

List of abbreviations

MGI - McKinsey Global Institute

IPCC - Intergovernmental Panel on Climate Change

KDD - Knowledge Discovery in Databases

OGC - Open Geospatial Consortium

WIO – Western Indian Ocean

WIOMSA – Western Indian Ocean Marine Science Association (WIOMSA)

GIST – Generalized Search Tree

KNBS – Kenya National Bureau of Statistics

IOM – International Maritime Organisation

UN – United Nations

GIS - Geospatial Information Science

SPORT - Satellite pour l'Observation de la Terre

Acknowledgements

First, I would like to thank the almighty God for the gift of life and wisdom. Second, I would like to express my heartfelt gratitude to my supervisor, **Dr. Victor Onyango Odipo** and **Dr Joseph Maina Mbui** for their invaluable guidance, support, and encouragement throughout the entire process of writing this thesis. Their expertise and feedback have been instrumental in shaping this research.

I am also grateful to my family and friends for their unwavering support and understanding during this challenging journey. Their encouragement has been a constant source of motivation for me.

Lastly, I would like to acknowledge the participants of this study for their willingness to share their experiences and insights, without which this research would not have been possible.

Thank you to everyone who has contributed to the completion of this thesis. Your support and encouragement have been invaluable to me.

Dedication

I dedicate this thesis to my parents, for their unwavering support and encouragement throughout my academic journey. Their love and guidance have been my driving force, and I am forever grateful for everything they have done for me. This thesis is also dedicated to my professors and mentors, whose wisdom and guidance have shaped me into the researcher I am today. Thank you for believing in me and pushing me to reach my full potential. Lastly, this thesis is dedicated to all the individuals who have inspired and motivated me along the way Your words of encouragement and belief in my abilities have been instrumental in my success.

Chapter 1: Introduction

The last and best opportunity for humanity to address the present environmental and climate challenges is in this decade(Carter et al. 2015). Environmental sustainability is a critical issue facing port cities in the global South today, with a growing need to understand the relationships between these cities and their ports. Traditional approaches need to be replaced with comprehensive, integrated approaches that help policymakers and other stakeholders put the environment at the centre of their sustainable development objectives(O'Neill et al. 2020).

Transformation is only possible through utilizing the potential of spatial big data tools to support data-informed planning(Ma et al. 2020). The rise of big data provides an opportunity to gain insights into the complex relationships between ports and cities and to develop solutions to environmental sustainability issues in port cities in the global South. In this context, spatial association rule mining based APRIORI algorithm has been proposed as a solution to this challenge.

The APRIORI algorithm is a well-established data mining technique that is widely used for discovering patterns in large datasets. The spatial association rule mining based APRIORI algorithm is a modification of the traditional APRIORI algorithm, which takes into account the spatial location of data points and is particularly well-suited to analysing large, complex spatial datasets(Pampoore-Thampi, Varde, and Yu 2021).

In this research, I will investigate the use of spatial association rule mining based APRIORI algorithm to address environmental sustainability issues in port cities in the global South. i will review the existing literature on big data solutions to environmental sustainability issues in port cities, as well as on spatial association rule mining based APRIORI algorithm, and identify gaps in current research. We will then design and implement a study to evaluate the effectiveness of this approach in addressing environmental sustainability issues in port cities in the global South.

Overall, my goal is to contribute to the body of knowledge on environmental sustainability in port cities in the global South and to provide insights into the relationships between these cities and their ports. We believe that our research will be of great interest to scholars, policymakers, and practitioners who are working to address these critical issues.

1.2 Problem statement

Ports are essential for global trade and economic development, but they also pose significant environmental challenges. The activities associated with shipping, such as fuel combustion and waste generation, contribute to air and water pollution, loss of biodiversity, and climate change (IOM, 2018). These environmental impacts are particularly pronounced in port cities in the global South, which face a range of additional challenges, such as limited resources for environmental protection (World Bank, 2012), weak governance structures (UN, 2016), and rapid urbanization (UN, 2014).

It is without a doubt that ports and cities have one of the most complex and intriguing relationships, for instance, ports are major sources of air, water, noise, and waste pollution, which can have significant impacts on the local environment and public health. For example, shipping emissions contribute to air pollution, including particulate matter, nitrogen oxides, and sulphur dioxide, which can have adverse effects on human health and the environment (UNEP, 2013). Furthermore, as cities grow and expand, they often encroach on port areas, leading to conflicts between port development and urbanization. For example, port expansion can result in the loss of valuable coastal habitats, such as wetlands, and can lead to the displacement of local communities (World Bank, 2012). Lastly, Ports are vulnerable to the impacts of climate change, including rising sea levels, stronger storms, and more frequent and intense flooding. At the same time, ports and shipping routes are also significant sources of greenhouse gas emissions, which contribute to climate change (International Maritime Organization, 2018). These complex relationships highlight the need for a multi-disciplinary and integrated approach to addressing environmental sustainability issues in port cities, which considers both the environmental and economic impacts of ports.

Despite the recognition of these challenges, there is a lack of comprehensive and systematic approaches to addressing environmental sustainability issues in port cities, especially in the global South. The existing approaches to addressing these issues have been fragmented and have not effectively integrated multiple perspectives and data sources (UN Environment, 2017). There is a need for a more holistic and data-driven approach to addressing environmental sustainability in port cities, especially in the global South (UN-Habitat, 2014).

This research aims to address this gap by exploring the use of Big Data solutions, specifically spatial association mining based on the APRIORI algorithm, for improving environmental sustainability in port cities in the global South. The goal is to identify the significant

associations between environmental sustainability variables and economic activities and to inform the development of more effective and efficient environmental sustainability policies and practices for these cities. The research will contribute to the field of environmental sustainability by demonstrating the potential of Big Data solutions for addressing environmental challenges in the global South and will provide a basis for further research on the use of data-driven approaches for environmental sustainability in other regions and sectors.

1.2 Research Objective

The overall objective of this study is to explore the use of Big Data solutions, specifically spatial association mining based on the APRIORI algorithm, for improving environmental sustainability in port cities in the global South.

1.3 Specific Objectives

1 To analyse the significant associations between environmental sustainability variables and economic activities in port city.

2 To assess the feasibility and effectiveness of using Big Data solutions, specifically spatial association mining based on the APRIORI algorithm, for addressing environmental sustainability issues in port cities.

3 To contribute to the field of environmental sustainability by demonstrating the potential of Big Data solutions for addressing environmental challenges in port cities.

4 To establish a comprehensive, adaptable framework for leveraging data-driven methodology, utilizing spatial association mining with the Apriori algorithm to analyse and enhance environmental sustainability in port cities.

Chapter 2: Literature Review

Environmental sustainability has become a major concern for many cities around the world, including port cities. Port cities are facing a range of environmental sustainability challenges, such as air and water pollution, waste management, and sustainable transportation. Addressing these challenges requires innovative approaches, and big data and GIS are playing an increasingly important role in providing solutions.

In this literature review, I will examine the recent research on the use of big data and GIS for environmental sustainability in port cities, I will also identify key findings and contributions, as well as research gaps and opportunities for future research.

2.1 The evolution of big data

Big Data is not a single 'object,' but rather a collection of data sources, technologies, and approaches that have developed from and seek to harness the rapid explosion of data. Big data is a catchphrase that refers to a huge number of both organized and unstructured data that is difficult to manage using conventional databases and software approaches. More data is being created and is being recognized in the popular literature as "big data," its use is becoming more widespread, and it is potential for policy-making and international development is only beginning to be explored (Nyairo et al. 2022)

Big data is a term used to describe vast amounts of fast-moving, complicated, and variable data that need sophisticated methods and tools to be collected, stored, distributed, managed, and analysed. The immense volume of data, the vast variety of data formats, and the speed at which the data can be handled are the three Vs that can be used to define big data (Favaretto et al. 2020),(De Mauro, Greco, and Grimaldi 2016). Although the term "big data" doesn't relate to a particular amount of data, it is frequently used to refer to petabytes and exabytes of data, much of which is difficult to integrate. It is notable that a fourth characteristic, veracity, or "data assurance," has lately been added by several data scientists and researchers. In other words, big data analytics and results are accurate and reliable. However, integrity is still an ideal and is not (yet) a reality (Reimer and Madigan 2019).

Data sets are growing larger partly due to low-cost and widespread information-sensing, mobile devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks being used to collect data (Peters et al. 2014). Since the 1980s, the world's technical per-person capacity for data storage has typically doubled every 40 months [15]; as of 2012, 2.5 exabytes (2.51018) of data were produced every

day. Superpower High Tech Corporation produced 2.3 zettabytes (2.3 1021) of data every day as of 2014(Berendsen 2020).

Letouzé, a Big Data for Development pioneer, has defined "the 3 Cs" of Big Data, giving a different perspective. The 3Cs stand for Big Data 'Crumbs,' Big Data 'Capabilities,' and Big Data 'Community,' and they fundamentally describe Big Data as an ecosystem, a complex system in fact, rather than data sources, collections, or streams. It also refers to and contradicts the three Vs of Big Data (2012).

According to his view, Big Data is more than just data, no matter how large or diverse it is; this is why and where Big Data as a field—an ecosystem—exists. Gary King's Harvard lecture on "Big Data is Not About the Data" likewise and maybe emphasizes that big data is first and foremost "about" analytics, the tools, and processes that are used to create insights, turn data into information, and then, perhaps, knowledge (Silva, Lokanathan, and Kreindler 2016).

The second "C" of big data, capacity, is all about: the tools and methodologies, the needs and advancements in hardware and software, and the human skills. Crumbs are useless without the consideration and development of capacities. However, it's not only about knowledge and aptitude; it's also about how the question is phrased. Of course, this relates to the idea of "Data Literacy" and the requirement to become knowledgeable users and commenters.

The third C of community refers to the group of participants—both the creators and consumers of these resources and capacities; in reality, it's the human element, and it might include the entire planet. A complex ecosystem is created by the concentric circles that result, with the community serving as the bigger set and having feedback loops between them. The scope of this research can then be extended to include stakeholders in sustainable development who are thinking about joining the big data ecosystem and utilizing big data. Additionally, development actors would serve in this capacity as convening forces, knowledge brokers, and facilitators.

2.2 Application of Big Data and GIS for Environmental Sustainability

The use of Big Data and Geographic Information Systems (GIS) in the Global South has been increasingly relevant in promoting environmental sustainability (Kwakkel, 2018). According to Wu(Wu, 2019), These technologies have been used to monitor and manage natural resources, track the impact of human activities on the environment, and support decision-making processes towards sustainable development. In recent years, the Global South has

experienced significant economic and demographic growth, leading to an increase in environmental pressures and challenges. In this context, Big Data and GIS have been valuable tools to help decision-makers address these issues. For example, they have been used to map and monitor deforestation, track wildlife populations, and manage water resources (Bisht, 2019).

Big Data technologies, such as remote sensing, provide high-resolution and frequent data on environmental variables. This information can be combined with GIS tools to create maps and visualizations that highlight the distribution of environmental features and changes over time (Kwakkel, 2018). This allows for a more comprehensive and spatially explicit understanding of environmental dynamics and their impact on ecosystems and human communities. GIS also supports decision-making by providing data on environmental risks and vulnerabilities, allowing stakeholders to prioritize and allocate resources towards the most critical issues. For example, GIS has been used to assess the vulnerability of coastal communities to sea-level rise and hurricanes, and to design adaptation strategies that are both effective and sustainable. Moreover, the use of Big Data and GIS has been crucial in promoting environmental sustainability in the Global South. These technologies have been valuable in monitoring and managing natural resources, tracking the impact of human activities, and supporting decision-making towards sustainable development. However, it is important to consider the limitations and challenges of these technologies, such as data availability, accuracy, and privacy, to ensure their effective use for environmental sustainability.

2.3 The Use of Big Data for Environmental Sustainability in Port Cities

According to the literature, big data applications are mostly not yet linked to environmental sustainability. Industries including retail, manufacturing, and healthcare are being transformed by big data. However, the environmental sustainability effort, which, according to the Environmental Protection Agency, “sustains the life on earth is yet to benefit from the explosion of big data solutions.

Global issues continue to be at the forefront, including climate change. The most recent report from the Intergovernmental Panel on Climate Change (IPCC) signalled the apparent increase in global temperatures and stressed the crucial role that human activity has played in causing this change (IPCC 2013). The Environment Impact Assessment reports that people have

significantly affected the world's ecosystems, leading to "a major and largely permanent loss in the diversity of species on Earth," indicating that ecosystem health is declining (Bawa et al. 2021).

Considering the emerging varieties of analytics and insight that big data could produce, statements have been made that call for ramped-up efforts to solve issues of environmental sustainability. According to John Elkington, the creator of the consulting firm's sustainability and Volans, sustainable outcomes will be greatly influenced by big data and how we use statistics and data. In turn, this will lead to a transition in what Elkington (Conti 2021)refers to as a "Breakthrough Decade" from "values as a driver of change to valuations."

2.4 APRIORI Algorithms

The APRIORI algorithm, introduced by Agrawal and Srikant in 1994, is a fundamental technique in data mining that focuses on discovering frequent itemsets in a dataset. The algorithm uses the concept of the Apriori property, where any subset of a frequent itemset must also be frequent. This property allows the algorithm to prune the search space efficiently, reducing the number of candidate itemsets that need to be examined. By iteratively generating larger and larger candidate itemsets, APRIORI gradually builds up the frequent itemsets, ultimately revealing patterns and associations within the data. These discovered itemsets can be further utilized in market basket analysis, recommendation systems, and association rule mining. The APRIORI algorithm plays a crucial role in handling large-scale datasets efficiently and has been widely adopted in various fields for extracting valuable insights from complex datasets. (Hana Bernika Sabila et al., 2023).

A. Basic Concepts of APRIORI Algorithm

The APRIORI algorithm, a fundamental concept in association rule mining, operates on the principle of generating frequent itemsets based on minimum support thresholds. Initially proposed by Agrawal and Srikant in 1994, the algorithm's efficiency lies in its capability to reduce the search space by exploiting the Apriori property, which states that any subset of a frequent itemset must also be frequent. By iteratively pruning infrequent itemsets, the APRIORI algorithm gradually builds up higher order itemsets until no more are found to meet the support threshold. This methodology significantly decreases computational overhead compared to exhaustive enumeration, particularly beneficial for large datasets where identifying significant associations becomes a daunting task. As a result, the APRIORI algorithm serves as a cornerstone for association rule mining applications in various domains,

including geospatial data analysis. However, the algorithm's performance can be highly dependent on factors such as the threshold selection and dataset characteristics, requiring careful tuning for optimal results. The foundational principles of the APRIORI algorithm fundamentally shape its adaptive nature and pave the way for leveraging association rules in complex data scenarios. (Haoyu Xie, 2021).

B. Working Principle of APRIORI Algorithm

The APRIORI algorithm, a fundamental association rule mining technique widely used in various domains, operates on the principle of frequent itemset generation to discover significant patterns within datasets. The APRIORI algorithm employs support and confidence measures to identify itemsets that frequently co-occur above predefined thresholds. The algorithm consists of two main steps: candidate generation and pruning. In the candidate generation step, APRIORI generates potential itemsets that may meet the support threshold, while in the pruning step, it eliminates itemsets that do not satisfy the minimum support criteria. This process continues iteratively, building larger itemsets from smaller ones until no further frequent itemsets can be found. The efficiency of the APRIORI algorithm lies in its ability to reduce the search space by exploiting the downward closure property, which guarantees that all subsets of a frequent itemset are also frequent. Thus, the APRIORI algorithm efficiently uncovers significant associations within large datasets through its systematic approach to itemset generation and pruning (Longzhi Yang et al., 2021).

2.5 Geospatial Data

A. Definition and Types of Geospatial Data

Geospatial data refers to information that has a geographic component attached to it, allowing for the representation of features on the Earth's surface. This data is typically stored and analysed using geographic information systems (GIS) (Jiahao Xia et al., 2024). Various types of geospatial data exist, including vector data, and raster data. Vector data represents information using points, lines, and polygons, making it suitable for precise mapping of discrete features such as roads or buildings. On the other hand, raster data organizes information into a grid format, enabling the analysis of continuous data like elevation or temperature. Satellite imagery involves data captured by satellites orbiting the Earth, providing detailed visual representations of the planet's surface. Understanding the different types of geospatial data is crucial for the effective application of the APRIORI algorithm in spatial data mining research.

Geospatial data plays a vital role across various sectors due to its ability to provide precise location-based information for decision-making processes. In the agriculture sector, geospatial data aids in monitoring crop health, predicting yields, and optimizing irrigation systems based on soil moisture levels and topography. Within urban planning, such data assists in analysing population density, land-use patterns, and infrastructure planning for sustainable development. Moreover, in disaster management, geospatial data facilitates rapid assessment of affected areas, enabling efficient allocation of resources and timely response efforts. These examples underscore the significance of geospatial data in enhancing operational efficiency, resource management, and informed decision-making in diverse sectors, ultimately leading to improved outcomes and cost-effectiveness. It is evident that geospatial data is a valuable asset with numerous practical applications across industries (Tosin Harold Akingbemisilu, 2024).

B. Challenges in Geospatial Data Analysis

Challenges in Geospatial Data Analysis are multifaceted and can impede the effective application of the APRIORI algorithm. One significant challenge is the reliability and accuracy of geospatial data sources. This includes issues such as data inconsistencies, incompleteness, and errors which can significantly impact the outcomes of the algorithm. Moreover, the scalability of geospatial data poses another challenge, particularly when dealing with large datasets that require substantial computational power and storage capacity for analysis. Additionally, the interpretability of the results generated through geospatial data analysis using the APRIORI algorithm can be complex, making it challenging for non-experts to understand and utilize effectively. Addressing these challenges is crucial for maximizing the potential of the algorithm in geospatial applications and ensuring its accuracy and reliability in decision-making processes (Terry Devara, 2023)

2.6 Integration of APRIORI Algorithm in Geospatial Databases

A. Applications of APRIORI Algorithm in Geospatial Data Analysis

Spatial databases play a crucial role in various fields such as urban planning, transportation systems, environmental monitoring, and geospatial analysis. The Apriori algorithm, originally developed for market basket analysis, has found application in spatial databases to discover frequent patterns and associations among spatial objects. The use of the Apriori algorithm in spatial databases has been explored in several studies, with a focus on its effectiveness in discovering meaningful relationships and patterns in spatial data. One study by Deren Li et al., (2016) examined the application of the Apriori algorithm in spatial databases for discovering

frequent spatial association rules. They applied the algorithm to a real-world dataset of traffic accidents, aiming to uncover associations between different spatial attributes such as road type, weather conditions, and accident severity. The results showed that the Apriori algorithm was able to identify significant associations between these attributes, providing valuable insights for improving road safety measures. (Viet-Ha Nhu et al., 2024) evaluated the use of the Apriori algorithm in spatial databases for land use classification. They utilized remote sensing data and spatial attributes such as soil type, vegetation cover, and topography to classify land use patterns. The results showed that the Apriori algorithm effectively identified frequent patterns and associations among these attributes, enabling accurate land use classification. Furthermore, the Apriori algorithm has also been applied in spatial clustering analysis. For example, a study by (Matthias et al., 2022) utilized the Apriori algorithm to identify spatial clusters of disease outbreaks based on various attributes such as population density, environmental factors, and socioeconomic variables. Their results demonstrated that the Apriori algorithm could successfully identify significant patterns and associations among these attributes, aiding in the understanding of disease transmission and informing public health interventions. Overall, the use of the Apriori algorithm in spatial databases has shown promise in discovering frequent patterns and associations among spatial objects. These studies highlight the effectiveness of the Apriori algorithm in uncovering meaningful relationships and patterns within spatial data.

B. Case Studies Demonstrating APRIORI Algorithm in Geospatial Data

Case studies have demonstrated the efficacy of employing the Apriori algorithm within the context of geospatial data analysis. For instance, a study by (Viet-Ha Nhu et al., 2024) applied the APRIORI algorithm to mine association rules from remote sensing data, revealing valuable patterns and relationships between different geographic features. The algorithm's ability to handle large datasets efficiently and extract meaningful associations has been pivotal in uncovering hidden insights within spatial data. By utilizing the APRIORI algorithm, researchers can identify significant patterns in geospatial datasets that would have otherwise been challenging to detect using traditional methods. These case studies showcase the algorithm's utility in geospatial data analysis and its potential to enhance decision-making processes in various fields such as environmental monitoring, urban planning, and natural resource management. The application of the Apriori algorithm in geospatial data holds promise for advancing spatial data analysis capabilities and extracting valuable knowledge from complex geographical datasets.

Chapter 3: Methodology

3.1 Introduction

The goal of this study was to investigate the impact of various factors on environmental sustainability in port cities using spatial association rule mining techniques. A case study approach was employed to conduct an in-depth analysis in Mombasa, a port city located in southeast part of Kenya.

3.2 Spatial Database

The study area boundary was defined based on the geographical limits of the port city. The area was then divided into a grid of 10 by 10-meter cells using a Geographic Information System (GIS), creating the basic units of analysis for the spatial database. A total of over one Million plus grid cells were generated covering the entire study area.

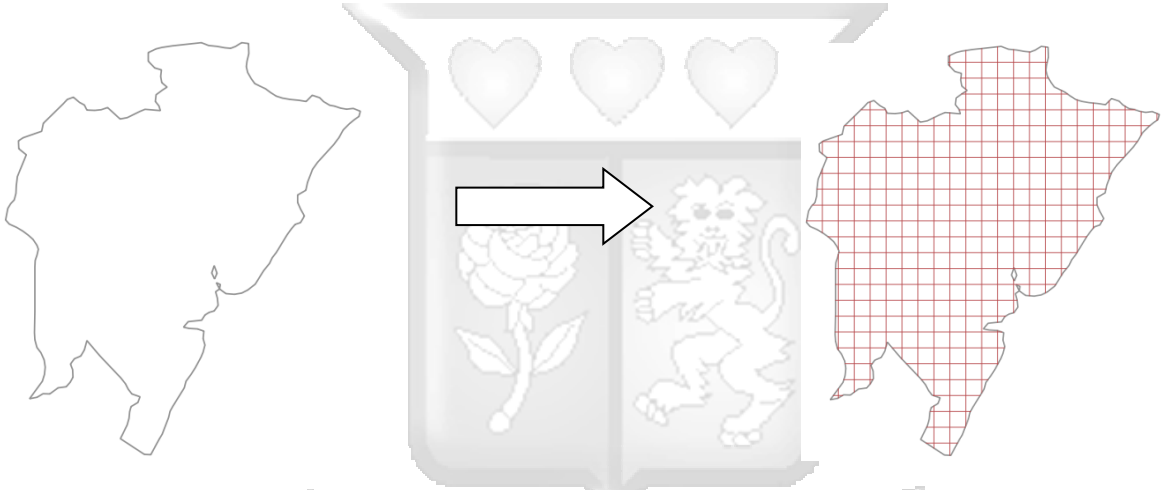

Figure 3. 1: 10 by 10 Meter grid, Source Author

Data collection and Data extraction

Primary and secondary data were collected from a variety of sources between 2010 to 2022. Spatial data on environmental, economic, and social variables were extracted for each grid cell based on their geographic coordinates. A total of 10 attributes were recorded for analysis representing categories such as land use, population density, industrial facilities, transportation infrastructure, coastal and marine ecosystems, air and water quality indicators.

Majority of the spatial data for the study was extracted through analysis of high-resolution satellite images acquired between 2000 and 2021. Table 3.1 shows detailed description of the satellite images used in this analysis.

Table 3. 1:Description of Satellite Images acquired.

Year	Description
2021	SPOT 6/7 satellite image at 1.50m spatial resolution collected on 29th March 2021 with less than 5%. Cloud Cover
2018	SPOT 6/7 satellite image at 1.50m spatial resolution collected on 7th March 2018 with 0% Cloud cover
2014	SPOT 6/7 satellite image at 1.50m spatial resolution collected on 9th March 2014 and 24th Nov. 2014 with less than 12% Cloud cover
2010	GeoEye-1and WorldView-2 satellite image at 0.50m spatial resolution collected between 4th March 2010 and 3rd July 2010 with less than 10% Cloud cover
2006	Quick Bird and IKONOS at 0.61m spatial resolution collected between 17th January 2006 and 24th. December 2006 with less than 10% Cloud cover
2000	Quick Bird at 0.61m spatial resolution collected between 17th March 2002 and 10th May 2003 with less than 10%

ERDAS Imagine software was used to generate Land use and land cover maps through supervised classification of the satellite images for each year. Sample areas of different land types such as residential, commercial, industrial, agricultural, forested, water bodies among others, were first identified based on visual interpretation.

Population density data was estimated by first identifying built-up areas through an urban mask generated from the normalized difference vegetation index. Dwelling structures were then extracted using object-based image analysis incorporating texture and context. Population counts were allocated to each structure based on ancillary census data which was extracted

from 2019 census report from Kenya National Bureau of Statistics (KNBS)[1] , generating gridded population density surfaces.

Budling dataset was obtained from Google Open Building datasets using Google Earth engine[2]. While Transportation networks data were extracted from Open Street Map using R package osmdata[3].

Coastal and marine ecosystem boundaries like mangroves and seagrass were obtained from Global Ocean data viewer. While Urban green space data was generated from the NDVI image whereby the image was reclassified based on threshold values >0.2 – 0.3 for vegetation.

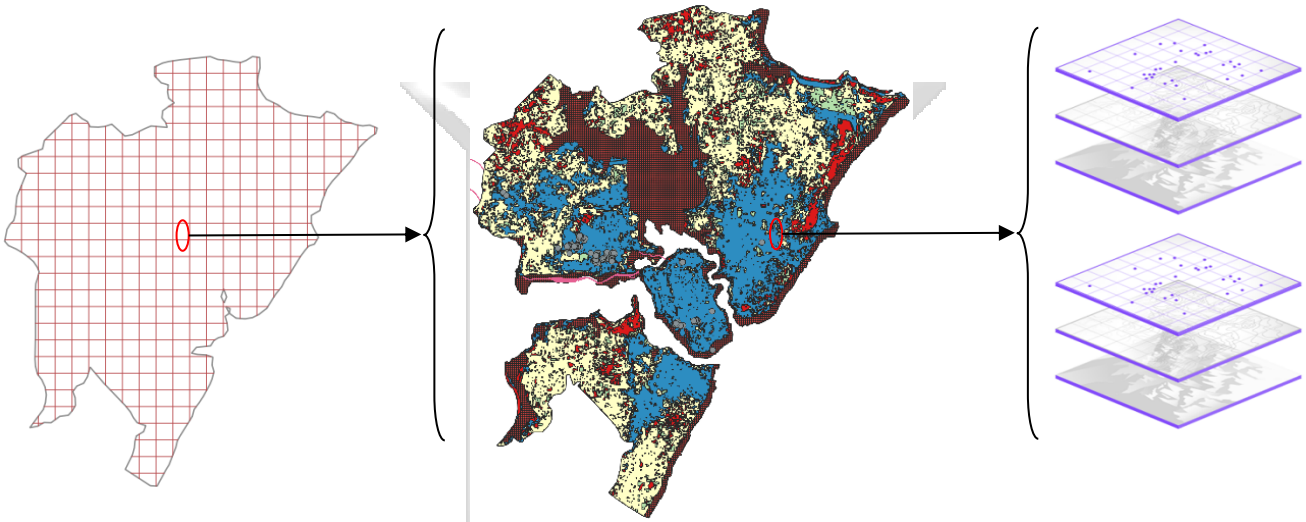

Figure 3. 2 : Data extraction process based on grids, Source Author

[1] https://statistics.knbs.or.ke/nada/index.php/catalog/68

[2] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y.S.E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, J.A. Quinn. Continental-scale building detection from high resolution satellite imagery. arXiv:2107.12283, 2021.

[3] Padgham, Mark. 2016. Osmplotr: Customisable Images of OpenStreetMap Data. https://cran.r-project.org/package=osmplotr.

Table 3. 2: Data layers generated for this study.

Major Category	Subcategory	Description	Indicators
Ecological	Geology/Water	Identifies geophysical vulnerabilities	1. Percent of Landscape that is Arable Land (land-use & land cover) 2. Percent of Bodies of Water with High Water Quality
	Ecosystems	This category assesses the extent and health of ecosystem services	1. Level of Mangrove Coverage 2. Level of Coral Reefs Coverage 3. Health of Existing Coral Reefs 4. Level of Mangrove Coverage
	Environmental	climate-related impacts on cities	1. Presence of air pollution 2. Presence of water pollution 3. Change in Sea Surface Temperature 4. % if CO2 in the atmosphere
Socioeconomic/ Economic	Economics	economic factors of the coastal city	1. Urban Unemployment Rate 2. GDP Per Capita 3. Number of occupants per household
	Major Industries	Certain economic sectors are highly vulnerable to climate change	1. Percent of National Economy Based in Offshore Fisheries 2. Percent of National Economy Based in Port and Shipping Industries 3. Percent of National Economy Based in Nearshore Fishing Industry
	Infrastructure		1. Roads, railways, airport 2. Unplanned Settlement 3. Level of Resilience for Ports and Shipping

Spatial database construction

The spatial database was constructed using PostgreSQL with the PostGIS spatial extension to take advantage of its robust capabilities for storing, querying, and analysing location-based data. The grid cell feature geometries were loaded into a table with a geometry column defined using the Spatial Reference System for the study area. Grid cell IDs were indexed as the primary key to enable fast retrieval.

Dimensional tables were created for each attribute category with columns for the grid ID foreign key and the attribute value. Attributes like land use were stored as lookup codes to reduce storage size. Numeric attributes were normalized for efficient aggregation and comparison. PostgreSQL offered transactional support throughout the ETL process to load the vast amounts of spatial and non-spatial data into the database tables from source files in a consistent and recoverable manner.

Spatial indexes GIST was used to create geometry columns to accelerate queries involving spatial predicates and functions. Multicolumn indexes were also applied to high-cardinality columns to optimize performance of complex JOIN operations.

The robust ACID compliance of PostgreSQL ensured data integrity even with concurrent access during the knowledge discovery process. Row-level locking prevented anomalies during simultaneous read and write transactions.

PostGIS extended the database with over 300 spatial functions and operators enabling advanced location analytics like overlay, proximity, density, and network analysis. User-defined functions were also created for custom modelling and validation tasks. Figure 3.3 shows the spatial database schema.

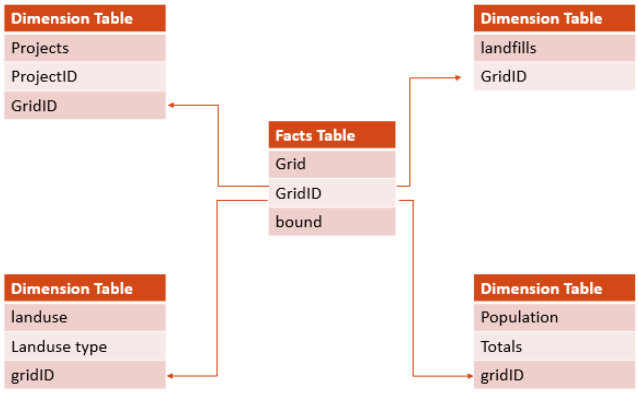

Figure 3. 3 : Database Schema, Source Author

Knowledge discovery

For the knowledge discovery process, association rule mining was performed on the spatial database using Python. The Apriori algorithm as implemented in the mlxtend library was used to generate rules from the transactional grid cell data. Each grid cell represented a transaction containing items in the form of attribute values. The minimum support threshold was set to 0.1%, meaning rules had to occur in at least 0.1% of all grid cells. Minimum confidence was set to 70%.

The algorithm first determined the frequency of occurrence of individual attributes and attribute pairs through one and two item sets. Those meeting minimum support proceeded to larger item sets. Candidates were generated through the Apriori property that any subset of a frequent itemset must also be frequent. To calculate support, the algorithm counted the number of transactions where the itemset occurred and divided by the total number of transactions. For confidence, it counted transactions where both the antecedent (if) and consequent (then) items occurred and divided by those with just the antecedent.

The support, Lift and confidence will be calculated as follows:

Support: Support(X) = P(X) Where P(X) is the probability of occurrence of itemset X in the transaction database

Support

$$\boldsymbol{Supp}(x_1 \cap x_2 \cap x_m) = \frac{Number\ of\ spatial\ objects\ to\ satisfy\ x_1 \cap x_2 \cap x_m}{total\ number\ of\ spatial\ object\ in\ SDB}$$

Confidence: It measures the strength of implication or certainty of rule that is, confidence is the probability that B will occur given A.

Confidence

$$\boldsymbol{Conf}(x_1 \cap x_2 \cap x_m \rightarrow y_1 \cap y_2 \cap y_m) = \frac{\boldsymbol{Supp}(x_1 \cap x_2 \cap x_m \cap y_1 \cap y_2 \cap y_m)}{\boldsymbol{Supp}(x_1 \cap x_2 \cap x_m)}$$

Lift: Lift measures the strength of rule or degree of dependency between antecedent and consequent. Lift (A → B) = P(B|A) / P(B)

Lift

$$\boldsymbol{Lift}(x_1 \cap x_2 \cap x_m \rightarrow y_1 \cap y_2 \cap y_m) = \frac{\boldsymbol{Supp}(x_1 \cap x_2 \cap x_m \cap y_1 \cap y_2 \cap y_m)}{\boldsymbol{Supp}(x_1 \cap x_2 \cap x_m) * \boldsymbol{Supp}(y_1 \cap y_2 \cap y_m)}$$

Chapter 4: Discussion of Results

An in-depth analysis of the association rules generated by the APRIORI algorithm from the dataset reveals intricate relationships between environmental sustainability variables and economic activities in port cities. These rules provide insights into the dynamics at play in these urban settings, offering a nuanced understanding of how different elements interact within the context of environmental sustainability.

4.1 Significant association between environmental sustainability variables and economic activities in port city.

The entwining of the realm of environmental preservation with of the port cities region's economic prosperity an intricate phenomenon, resulting in the corresponding evolution of both. In this section, the detailed investigation is given about the highly important relationships which were discovered and reinforced by the rigorous analysis of data cadging from our APRIORI algorithm application. The results of the analysis also displayed interesting patterns which uncovered that the environmental influencers that form the economic backbone of the port cities, are not isolated factors. The environmental forces are rather woven insolubly with the economic factors that define the life of the busily inhabited port cities.

By looking into the influence of these factors in a data-driven manner, the complex layers of their relationship and the consequences to the sustainability of port environments have been revealed. Urban/industrial infrastructure and ecosystems sustaining each other, and industrial activities affecting the surroundings have been investigated and showed the network of interdependency of this system in all its aspects.

Following the introductory summary that emphasizes the intricate linkages between environmental sustainability and economic activities in port cities, we transition into a more granular analysis of these connections. This deeper dive is exemplified by examining the relationship between economic activities and environmental sustainability indicators.

Table 4. 1 : Spatial Association Rules

Rule number	Antecedents	Consequents	Confidence Level	Lift	Implications
Rule 0	Presence of mangroves (mangroves_0.1)	Presence of roads (roads_0.1)	0.98	1.0	• High, indicating a strong likelihood that where mangroves are present, roads are also present, and vice versa. • Close to 1, suggesting that the presence of one is independent of the presence of the other, yet they occur together more frequently than would be expected by chance. This rule suggests a spatial correlation between mangrove ecosystems and road infrastructure. It could imply that in port cities, road development is undertaken with consideration to existing natural ecosystems, or possibly that roads are impacting areas where mangroves are present.
Rule 1	Presence of mangroves (mangroves_0.1)	Certain land use patterns (landuse_0.0).	0.99	1.0	• Similarly high, pointing to a significant association between mangrove presence and specific land use types. • Close to 1, suggesting that the presence of one is independent of the presence of the other, yet they occur together more frequently than would be expected by chance. This rule suggests a spatial correlation between mangrove ecosystems and road infrastructure. It could imply that in port cities, road development is undertaken with consideration to existing natural ecosystems, or possibly that roads are impacting areas where mangroves are present.
Rule 3	Environmental challenges	Negative impacts of ports	1.0	1.006	• Very high confidence level, suggests a strong co-occurrence of environmental challenges and impacts of port on environmental sustainability. • Slightly above 1, which indicates that these items are more likely to be found together than independently. The correlation between environmental challenge and Negative impacts may reflect a comprehensive approach to port development in port cities. This association can have implications for urban planning and development,

					potentially affecting environmental sustainability through changes in land use and habitat disruptions
Rule 4	roads_0.0	landuse_0.0	0.98	1.0	• High confidence level showing strong relationship where road tend to be present with certain type of land use. • High lift also reinforces. This could indicate that roads are often built in response to or in anticipation of certain land uses, reflecting urban or regional planning strategies

4.2 Exploring Big Data Solutions for Environmental Sustainability in Port Cities: Spatial Association Mining with APRIORI algorithm.

To have more holistic understanding on how this complex interaction occurs at bird eye perspective, a scenario base analysis was adopted with aid of a network graph to illustrates the intricate relationships between various environmental factors and activities within the port city.

A. Scenario one

From the graph we can strongly deduce that environmental challenges **(enviromental_challenges_1.0)** are closely linked to both negative impacts of port **(negative_port_effects_1.0)** and positive impacts of port**(positive_port_effects_1.0)** which also have a strong influence on land use changes **(landuse_ch_1.0)** withing the study area. Figure 4.1 Shows Network graph of association rules and how they interact with one another.

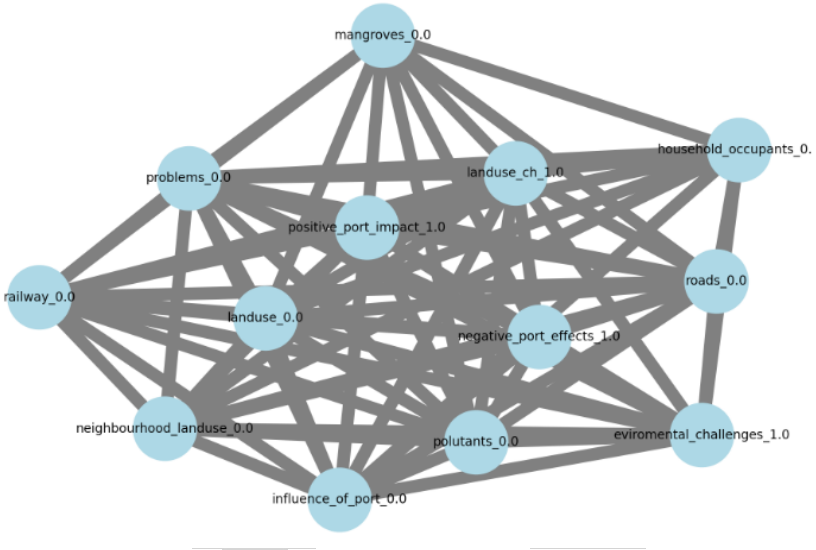

Figure 4. 1:Network Graph of Association Rules, Source Author

This interaction can be shown further on a map where we have land use change (built up areas closely linked to environmental pollution (mostly air quality and water pollution) being represented with points on the map. Figure 4.2 shows a map of land use land cover change overlayed by in-situ data of environmental pollutions.

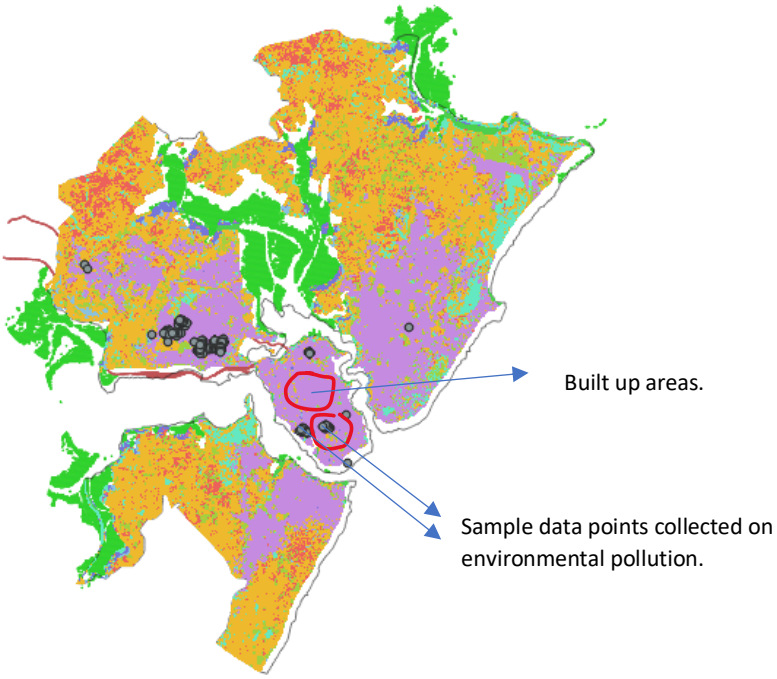

Figure 4. 2 : Land use land cover change with environmental variables, Source Author

B. Scenario two

Scenario two shows interaction between environmental challenges and pollutants. Each node represents an item (e.g., environmental conditions, land use patterns), and the edges represent the association rules, where the edge labels indicate the confidence of the association. As shown in Figure 4.3

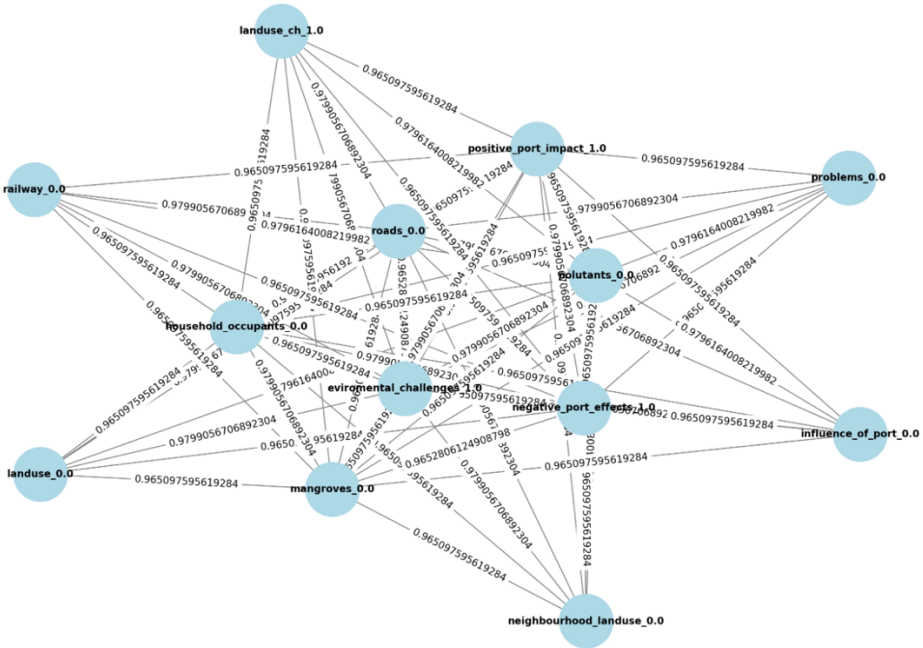

Figure 4. 3 : A Network Graph of association rules, Source Author

The graph suggests a direct association between "environmental_challenges_1.0" and "landuse_0.0". This implies that the presence of environmental challenges is likely to be associated with specific land use patterns, particularly those categorized under "landuse_0.0". from urban planning and development perspective, Environmental challenges could significantly influence urban and regional planning decisions, leading to the adoption of land use patterns that either mitigate these challenges or adapt to them. For instance, areas with high environmental challenges may see a shift towards more sustainable or resilient land use practices.

The presence or absence of pollutants, as indicated by "pollutants_0.0", also shows a connection to "landuse_0.0", suggesting that how an area is used or zoned can be directly influenced by the level of pollution. This could reflect policies or planning strategies aimed at reducing exposure to pollutants or rehabilitating polluted areas. The relationship between pollutants and environmental challenges is pivotal. High pollutant levels can exacerbate environmental challenges, leading to more stringent land use requirements and potentially fostering the development of green spaces, pollution buffers, and zones with restricted industrial activity.

The combined effect of environmental challenges and pollutants on land use is notably complex. Areas facing both high pollution levels and significant environmental challenges might be prioritized for interventions that address both issues simultaneously, such as the creation of green infrastructure, implementation of stricter environmental regulations, and the promotion of clean industries. The association rules indicating strong relationships between these factors and specific land use patterns could provide predictive insights for urban planners and environmental scientists. By understanding these associations, policymakers can better predict the impacts of environmental challenges and pollution on land use changes, allowing for more informed decision-making.

In summary, the confidence levels associated with the rules connecting these elements to specific land use patterns highlight the predictability of these relationships. High confidence in these rules suggests that when environmental challenges and pollutants are present (or absent), specific land use patterns are likely to follow. These insights could guide policy adjustments and urban planning strategies. Recognizing the strong linkage between environmental factors and land use patterns can lead to more sustainable urban development policies that account for environmental resilience, pollution control, and sustainable land management.

This analysis underscores the significant impact of environmental challenges and pollutants on land use patterns, revealing a complex interplay where both factors influence and are influenced by how land is utilized. The relationships deduced from the network graph illustrate that areas experiencing environmental challenges and high pollution levels may prioritize interventions like green infrastructure, stricter environmental regulations, and clean industry

promotion. Such insights are crucial for urban planners and policymakers, enabling the design of sustainable urban development policies that consider environmental resilience and pollution control.

Furthermore, this analysis has demonstrated the effectiveness how we can leverage Big Data solutions, particularly spatial association mining using APRIORI algorithm, in uncovering patterns that might not be apparent through traditional research methods. The ability to process large datasets and identify meaningful relationships between variables is essential for addressing environmental sustainability issues, which are often characterized by complex interactions between natural and man-made systems. By harnessing the power of Big Data analytics, stakeholders can better predict the impacts of various environmental factors on land use changes, allowing for more informed decision-making. This predictive capacity is instrumental in fostering long-term viability and resilience in both urban and rural landscapes, contributing significantly to the advancement of environmental sustainability efforts.

4.3 Contribution to Environmental Sustainability: Demonstrating Big Data Solutions' Potential for Port City Environmental Challenges.

Port cities, serving as hubs of commerce and transportation, inherently face a unique set of environmental challenges, from managing coastal ecosystems to integrating sustainable practices into urban and industrial activities. These challenges are exacerbated by the complex interdependencies of ecological, social, and economic factors that characterize urban coastal environments. The potential of Big Data solutions to address these environmental challenges has not been fully realized in the discourse of sustainable urban development.

One of the objectives of this study was to bridge this gap by harnessing spatial association mining techniques, grounded in the robust Apriori algorithm, to dissect and analyse large-scale urban data sets. Through a detailed examination of spatial co-occurrences and correlations between various sustainability variables and economic activities, we have shed light on the intricate patterns that underpin the environmental landscape of a bustling port city.

The analysis conducted has illuminated how Big Data analytics can not only identify existing correlations but also uncover latent associations that traditional methods might overlook. This subsection details the pivotal contributions of our research to the realm of environmental sustainability, particularly within the context of port cities, by revealing the profound capabilities of Big Data analytics. It underscores the role such technologies can play in facilitating more informed decision-making, enabling stakeholders to devise strategies that align with sustainability goals.

Through the demonstration of clear, actionable insights drawn from the confluence of environmental variables and economic indicators, the findings advocate for a data-driven

approach in urban environmental policy. This approach recognizes the nuanced balance between economic growth and environmental stewardship, a balance crucial to the long-term health and viability of port cities. In the following section, we expound on the specific contributions our study makes to this critical field, laying the groundwork for a future where Big Data is an integral part of sustainable urban planning and environmental resilience in port cities.

A. Discovery of Critical Associations

The application of the Apriori algorithm has unearthed critical associations between environmental sustainability factors and urban development patterns within port cities. The analysis underscores the recurring proximity of ecological areas, such as mangrove forests, urban green spaces, to bustling economic zones. Such juxtapositions suggest potential synergies or conflicts between environmental stewardship and economic development priorities.

These associations shed light on the spatial dynamics that underpin sustainable development strategies. Mangrove forests, for instance, are crucial for coastal protection, carbon sequestration, and supporting marine biodiversity. Their prevalence near economic zones may reflect an underlying recognition of these ecosystem services in urban planning. However, it also raises questions about the pressures exerted by urban expansion and economic activities on these fragile ecosystems.

Furthermore, the findings provide a nuanced understanding of the environmental impacts of economic activities. For example, the analysis reveals that areas with dense transportation networks, such as roads and railways, often coincide with high-value ecological zones. This raises critical considerations about the environmental costs of connectivity and the need for strategic environmental assessments in transportation planning.

The Big Data approach has enabled a holistic examination of environmental sustainability that factors in the multidimensional nature of urban development. This study moves beyond simplistic cause-and-effect assumptions, allowing for a more sophisticated understanding of the environmental challenges in port cities. The patterns identified point towards the need for integrated environmental policies that are tailored to the unique contexts of these urban centres.

The implications of these findings are manifold. They provide a foundation for policymakers to initiate dialogues on the implementation of sustainable urban development frameworks that reconcile economic growth with environmental conservation. Additionally, these insights can guide investments in infrastructure and development projects, ensuring that they contribute positively to the sustainability trajectory of port cities.

Lastly, the associations discovered through this study highlight the critical role of data-driven decision-making in environmental management. By leveraging Big Data analytics, city planners and environmental managers can gain unprecedented insight into the complex interplay of factors that define the sustainability of port cities. This has the potential to transform environmental policymaking, leading to more effective and targeted interventions that support the dual objectives of economic prosperity and environmental resilience.

B. Spatial Association insights

Spatial association mining has unravelled complex patterns of land use and transportation infrastructure that are intertwined with environmental sustainability factors in port cities. These patterns reflect the multifaceted relationships between human activities and ecological systems. Through this granular lens, we observe not just the presence of green spaces within urban sprawls but also how these spaces interact with and are impacted by surrounding developments.

The intricate tapestry of urban growth, characterized by the expansion of transportation networks and commercial areas, is often woven together with threads of natural habitats. The coexistence of such diverse land uses has illuminated the silent dialogue between urban life and nature. Highways that skirt the edges of wetlands, residential zones abutting forested areas, and ports adjacent to marine conservation zones are just a few examples of this dialogue. The spatial association insights gleaned from our analysis provide a map of these interactions, offering a visual narrative of the delicate balance that port cities must navigate.

Moreover, the findings of this study emphasize the importance of green corridors and buffer zones in urban planning. The identification of specific regions where ecological areas are sandwiched by urban developments has spurred conversations on the necessity for and potential placement of these green spaces. Such insights are crucial in fostering urban biodiversity, mitigating the urban heat island effect, and enhancing the quality of life for city residents.

This deeper understanding is pivotal for shaping future urban expansion in a way that not only respects but also harnesses the benefits of natural ecosystems. For instance, the knowledge of spatial associations between mangroves and commercial districts can inform the creation of green belts that serve as natural barriers against storm surges while potentially boosting eco-tourism.

The spatial patterns discovered also raise questions regarding the sustainability of current urban planning approaches. By providing evidence of the current state of land use, our analysis sets a benchmark for measuring future changes and assessing the effectiveness of interventions aimed at enhancing environmental sustainability. The Big Data-driven insights compel stakeholders to consider the long-term ramifications of development and the transformative power of planning that prioritizes ecological integrity.

Furthermore, the nuances captured in our spatial data analysis highlight the potential for using Big Data solutions in the creation of dynamic, responsive urban environments. These environments can react to changes in both human activity and natural phenomena, adjusting and evolving in ways that protect and enhance urban ecosystems.

In summary, the spatial association insights offered by our Big Data analysis do more than just map the current landscape of port cities. They serve as a guide for strategic urban development, providing a foundation for policies that can lead to more sustainable, resilient, and harmonious cityscapes. As port cities continue to grow and face environmental challenges, these insights will be invaluable for policymakers and urban planners in crafting a sustainable future.

C. Contribution to sustainability practises

The research conducted has a potential to contribute to the field of environmental sustainability by exploiting the extensive datasets inherent to a port city's landscape. By delving into this rich data, we have not only identified but also quantified the often subtle and intricate patterns that characterize the interplay between urban development and the environment. These patterns, which might remain hidden using traditional analytical methods, are brought to the forefront with spatial Big Data analytics, providing a new dimension to our understanding of sustainable urban ecosystems.

The power of Big Data to cut through the complexity of urban systems has revealed actionable insights into how environmental factors are influenced by, and in turn influence, the economic and social fabric of port cities. For example, the correlation between transportation infrastructure and green spaces has been delineated with unprecedented clarity, allowing us to contemplate not only the current state but also the trajectory of urban development. This clarity is crucial for the creation of sustainable strategies that can accommodate the fast-paced growth of port cities while preserving the natural environment.

Moreover, this analysis extended beyond merely mapping current conditions; it offers predictive insights that are essential for proactive sustainability practices. By understanding the existing spatial relationships, we can forecast the environmental impact of planned developments, anticipate the challenges that may arise from urban growth, and propose mitigation strategies that align with sustainability goals. Such foresight is instrumental in transforming urban planning into a process that is not merely reactive but anticipatory, paving the way for resilience and adaptability in the face of environmental changes.

The contribution of this research also lies in its ability to facilitate a more integrated approach to environmental policymaking. The granular details captured through Big Data analytics enable a more nuanced consideration of how policies can be tailored to address specific local conditions. This tailored approach is particularly pertinent in port cities, where the proximity to coastal and marine ecosystems requires a delicate balance between utilization and

conservation. Our findings advocate for the adoption of data-driven methodologies in policy formulation, ensuring that environmental strategies are not only evidence-based but also contextually relevant.

Furthermore, the integration of Big Data into environmental sustainability frameworks represents a paradigm shift in how data is perceived and utilized. The traditional view of data as a mere tool for retrospective analysis is replaced by an understanding of its predictive and prescriptive potential. This shift highlights the transformative role that data can play in guiding sustainable practices and in the formulation of adaptive management strategies that can evolve with both the city and its environment.

4.4 Framework for future strategies

Building on the solid foundation laid by our methodical approach to gathering and analysing spatial data within the port city of Mombasa, we propose a comprehensive framework to guide policymakers and stakeholders. This framework is not only adaptable but is designed to evolve with the dynamic and complex interplay between urban development and environmental sustainability in different regions. This proposed framework is divided into two steps namely: A. Spatial database construction, B. Knowledge discovery and spatial association mining.

A. Spatial database construction.

The first step in building a spatial transactional database is to define the boundaries of the study area that will be analysed. This could be a specific city, county, region, or other geographic extent depending on the scope of the research. Once the study area is established, it needs to be divided into discrete grid cells or tiles that will form the basic spatial transactions. The size of these grid cells should be determined based on the scale and level of detail required for the analysis - smaller grid cells allow for more localized patterns to be detected while larger cells provide a broader view.

After the grid framework is created, the next phase involves gathering all relevant spatial datasets that relate to the research questions. This typically includes economic data such as port through put, income levels, demographic information such as population density, environmental metrics like land use or pollution concentrations, and any other attribute data that could provide insight (Refer to Table 3.2). These disparate spatial datasets then need to be processed and integrated using GIS software which performs spatial joins between each layer and the grid framework. This allows the attributes from each dataset to be linked to the specific grid cell they intersect with based on spatial relationships like containment or intersection.

The output from the spatial joining process is a transactional database where each row equates to an individual grid cell and the columns represent the different feature attributes that were associated with that location through the spatial analysis. At this point, the data is structured in a format that can be mined and analysed using various algorithms designed for transactional

datasets. Figure 4.2 Shows the spatial database construction workflow while Table 4.3 shows a sample transaction database.

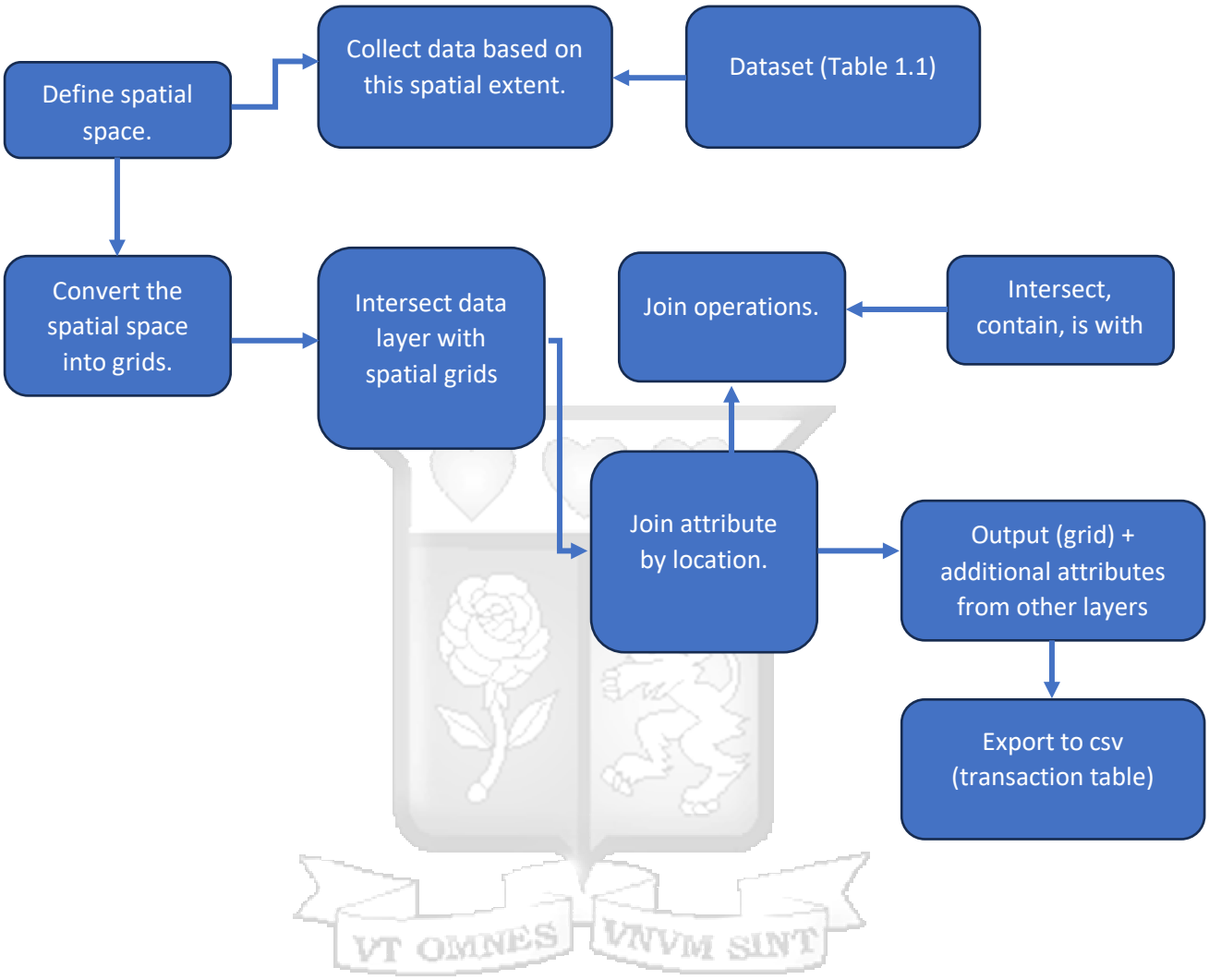

Table 4. 2 : A workflow for constructing spatial database, Source Author

Table 4. 3 : Shows a sample transactional database.

Cell ID	roads	rail way	land use	Household occupants	problems	Eviromental challenges	Influence of port	Positive port impact	negative port effects	polutants	neighbourhood_landuse
18011	unclassified		8	1	roads traffic water dust poor leadership	1	1	More clients for my business			commercial transport
18011	residential		8	1	roads traffic water dust poor leadership	1	1	More clients for my business			commercial transport
18011	unclassified		8	1	roads traffic water no_vegetation poor leadership	1	0	Most of my good customers are port employees	Unemployment due to removal of clearing by the use of sgr	0	commercial residential
18011	residential		8	1	roads traffic water no_vegetation poor leadership	1	0	Most of my good customers are port employees	Unemployment due to removal of clearing by the use of sgr	0	commercial residential
18011	unclassified		8	4	roads traffic tuktuk water air_pollution dust poor_leadership policy	0	1	My daily livelihood	Loss of employment - removal of clearing and forwarding companies and use of sgr	0	commercial transport residential
18011	residential		8	4	roads traffic tuktuk water air_pollution dust poor_leadership policy	0	1	My daily livelihood	Loss of employment - removal of clearing and forwarding companies and use of sgr	0	commercial transport residential
18013	unclassified		8	4	water infrastructure poor_leadership policy air_pollution roads	0	1		Lack of employment, removal of clearing and forwarding companies and use of sgr	0	commercial
18013	residential		8	4	water infrastructure poor_leadership policy air_pollution roads	0	1		Lack of employment, removal of clearing and forwarding companies and use of sgr	0	commercial
17488	unclassified		8	3	policy infrastructure dust roads poor_leadership	0	0		Unemployment due to temporary contracts	0	commercial
17488	residential		8	3	policy infrastructure dust roads poor_leadership	0	0		Unemployment due to temporary contracts	0	commercial
17488	unclassified		8	2	traffic roads water dust air_pollution infrastructure poor_leadership policy	0	0		Lack of employment, sgr-removal of clearing and forwarding companies	0	commercial
17488	residential		8	2	traffic roads water dust air_pollution infrastructure poor_leadership policy	0	0		Lack of employment, sgr-removal of clearing and forwarding companies	0	commercial
17489	unclassified		8	3	policy water air_pollution	1	1		Lack of customers dueto relocation of port employees	0	commercial
17489	residential		8	3	policy water air_pollution	1	1		Lack of customers dueto relocation of port employees	0	commercial
17488	unclassified		8	4	water roads poor_leadership policy air_pollution	0	0		Before was good, currently complicated	0	commercial residential
17488	residential		8	4	water roads poor_leadership policy air_pollution	0	0		Before was good, currently complicated	0	commercial residential
17751	path		8	4	water air_pollution roads	0	1		Before they used to get customers from Port, since SGR number of their customers have reduced	0	commercial
17751	residential		8	4	water air_pollution roads	0	1		Before they used to get customers from Port, since SGR number of their customers have reduced	0	commercial
18013	unclassified		8	4	air_pollution policy	0	1		Shifting of port activities to Naivasha led to loss of jobs and slow small businesses	0	commercial
18013	residential		8	4	air_pollution policy	0	1		Shifting of port activities to Naivasha led to loss of jobs and slow small businesses	0	commercial
18014	path		8	3	infrastructure air_pollution policy	0	1		Customers she depended on have moved	0	none
18014	residential		8	3	infrastructure air_pollution policy	0	1		Customers she depended on have moved	0	none
18013	unclassified		8	8	water air_pollution dust policy	0	1		Chemicals on eyes	1	commercial residential
18013	residential		8	8	water air_pollution dust policy	0	1		Chemicals on eyes	1	commercial residential

18013	unclassified		8	4	policy dust air_pollution water	0	0		When settlement shifts and moves there are no customers	0	none
18013	residential		8	4	policy dust air_pollution water	0	0		When settlement shifts and moves there are no customers	0	none

B. Knowledge discovery and spatial association mining

The second step of this proposed framework involves the implementation of the Apriori algorithm on the output data from the first step. Apriori is a classic algorithm for learning association rules between items in large transactional databases. Apriori works by finding frequent itemsets - sets of items that often occur together in the database above a predefined minimum support threshold. It makes use of the downward closure property, which states that any subset of a frequent itemset must also be frequent. Apriori basically follows the following steps.

Purpose: It is used to find frequent itemsets - sets of items, feature pairs, etc. that often occur together in a transaction database. This allows things like "**port expansion** also tend to have negative effects on ecological system such as **urban green space** " type of association rule mining.

Approach: It works in two main steps - first finding all frequent itemsets (that occur together above a minimum threshold), then generating strong association rules from those itemsets.

Frequent itemset generation: It relies on the property that any subset of a frequent itemset must also be frequent. It counts item occurrences to determine frequent items and item pairs, then longer itemsets incrementally.

Association rule generation: After finding all frequent itemsets, it generates rules by considering all subsets of each itemset. Rules must satisfy minimum confidence (conditional probability).

Parameters: It takes a minimum support threshold (minimum number of transactions an itemset must appear in) and minimum confidence as parameters

These steps can be summarized as follows:

```
1. L1 ← find_frequent_items(DB, minSup)
2. k ← 2
3. Lk ← ∅
4. while (Lk-1 ≠ ∅) do
5. Ck ← apriori_gen(Lk-1) // Candidate generation step
   // Generate candidate
   itemsets of size k from Lk-1
```

```
6.  for each transaction t in DB do

7.  Ct ← subsets(Ck, t)
8.  for each candidate c in Ct do
        c.count++ // Increment count of candidates contained.

9.  Lk ← {c in Ck | c.count ≥ minSup} // Eliminate candidates not meeting minSup

10. k++

11. return Uk Lk // Union of all frequent itemsets

12.
```

For this study, the Apriori algorithm is implemented in the Python programming environment using Google Colab notebooks[4]. Python provides a versatile platform for data analysis and machine learning tasks. The mlxtend library is leveraged to easily apply the Apriori algorithm on the output dataset from the first step of the framework. Mlxtend is a powerful Python library that implements many popular machine learning algorithms for extracting patterns from data.

Specifically, mlxtend.frequent_patterns.apriori is used to mine frequent itemsets and generate association rules. This function takes the transactional dataset, minimum support threshold, and other parameters as input.

The algorithm is then executed to find all frequent itemsets that meet or exceed the minimum support. Mlxtend handles all the steps of candidate generation, pruning, and support calculation internally using an Apriori approach. Finally, strong association rules are generated from the frequent itemsets based on a minimum confidence threshold and written to csv file for further analysis & interpretation.

[4] https://colab.research.google.com/drive/1x4Xd_N4Sjez4AqXeLXyN0oZXH8vaxzM2?usp=sharing

Chapter 5: Conclusions, Challenges, Recommendations and Future Work

A. Conclusion

The findings from the APRIORI algorithm analysis align well with the research objectives, showcasing the potential of Big Data solutions in enhancing our understanding of environmental sustainability in the global South's port cities. The significant associations identified between environmental variables and economic activities underscore the intricate connections that need to be considered in sustainable urban planning and development strategies.

Moreover, this research contributes to the field of environmental sustainability by demonstrating the applicability and effectiveness of data-driven approaches, like spatial association mining, in addressing complex environmental challenges. By uncovering the hidden patterns and relationships within the data, this study provides a foundation for further research on the use of Big Data for environmental sustainability, not only in other regions but also across different sectors.

The potential of these findings to inform policy and planning for sustainable development in port cities cannot be overstated. They offer a basis for developing targeted interventions that address the specific sustainability challenges identified through the analysis. As the global South continues to urbanize and develop, leveraging Big Data solutions like the APRIORI algorithm will be crucial for balancing economic growth with environmental preservation and sustainability.

B. Challenges in Implementing APRIORI Algorithm in Geospatial Data

One of the current challenges in implementing the APRIORI algorithm in geospatial data is the scalability issue when dealing with large datasets. As geospatial datasets continue to grow and complexity, traditional APRIORI algorithms may struggle to efficiently process the vast amounts of information involved. This can result in increased computational resource requirements and longer processing times, hindering the algorithm's practical application in real-world geospatial analysis tasks. Additionally, the interpretability of the results generated by the APRIORI algorithm in the context of geospatial data can be another challenge, as the relationships between different spatial features may not always be straightforward to comprehend. Addressing these challenges will be crucial for advancing the utilization of the APRIORI algorithm in geospatial data analysis and improving its effectiveness in extracting meaningful patterns and associations from spatial datasets (Guido Cervone et al., 2013).

C. Potential Enhancements and Modifications for Better Results

Potential enhancements and modifications to improve the results of the APRIORI algorithm in geospatial data analysis have been widely discussed in the literature. One key strategy discovered in this research is the incorporation of domain knowledge to optimize the selection of relevant itemsets, thereby enhancing the algorithm's efficiency. Additionally, introducing spatial constraints to the algorithm can improve the accuracy and relevance of the association rules generated, especially in geospatial applications where location-specific patterns are crucial. Moreover, adapting the minimum support and minimum confidence thresholds based on the characteristics of the geospatial data can lead to more meaningful results. These modifications can enhance the algorithm's performance in handling large-scale geospatial datasets and complex spatial relationships, ultimately yielding more valuable insights for decision-making processes in various domains. By integrating domain expertise and spatial constraints, the APRIORI algorithm can be tailored to deliver superior results in geospatial data analysis.

D. Recommendations for Future Research

Moving forward, there are several key areas that merit investigation to enhance the application of the APRIORI algorithm in geospatial data analysis. Firstly, future research should focus on developing more efficient strategies for handling the scalability issues encountered when applying APRIORI to large-scale geospatial datasets. Additionally, exploring the integration of spatial and temporal dimensions in association rule mining could provide valuable insights into dynamic patterns over time. Moreover, investigating the incorporation of domain knowledge or expert input to refine the rule generation process could further improve the accuracy and relevance of the results obtained. Lastly, conducting comparative studies with other data mining algorithms in geospatial applications can offer a comprehensive understanding of the strengths and limitations of Apriori in this context. By addressing these research directions, scholars can advance the efficacy and applicability of the Apriori algorithm in geospatial data mining. (Usama Awan et al., 2021).

Reference

Bawa, Kamaljit S. et al. 2021. "Securing Biodiversity, Securing Our Future: A National Mission on Biodiversity and Human Well-Being for India." *Biological Conservation* 253.

Berendsen, Clara. 2020. "Bid Data: An Engineering Marvel [Editorial]." *IEEE Potentials* 39(6).

Carter, Jeremy G. et al. 2015. "Climate Change and the City: Building Capacity for Urban Adaptation." *Progress in Planning* 95.

Conti, Diego de Melo. 2021. "Interview with John Elkington." *Sustentabilidade: Diálogos Interdisciplinares* 2.

Ding, Qin, Qiang Ding, and William Perrizo. 2008. "PARM - An Efficient Algorithm to Mine Association Rules from Spatial Data." *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(6).

Favaretto, Maddalena, Eva de Clercq, Christophe Olivier Schneble, and Bernice Simone Elger. 2020. "What Is Your Definition of Big Data? Researchers' Understanding of the Phenomenon of the Decade." *PLoS ONE* 15(2).

IPCC. 2013. "The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern Mental Panel on Climate Change." In Cambridge University Pres.

Laube, Patrick, Mark De Berg, and Marc Van Kreveld. 2008. "Spatial Support and Spatial Confidence for Spatial Association Rules." In *Lecture Notes in Geoinformation and Cartography*,.

Letouzé, Emmanuel. 2012. "Big Data for Development: Challenges & Opportunities." *Global Pulse is a United Nations initiative* (May).

Ma, Qiwei et al. 2020. "Measuring Functional Urban Shrinkage with Multi-Source Geospatial Big Data: A Case Study of the Beijing-Tianjin-Hebei Megaregion." *Remote Sensing* 12(16).

De Mauro, Andrea, Marco Greco, and Michele Grimaldi. 2016. "A Formal Definition of Big Data Based on Its Essential Features." *Library Review* 65(3).

Mou, Naixia, Hongen Wang, Hengcai Zhang, and Xin Fu. 2020. "Association Rule Mining

Method Based on the Similarity Metric of Tuple-Relation in Indoor Environment." *IEEE Access* 8.

Nejad Kousari, Alireza Mirzaie, and Ehsan Ghasemkhani. 2012. "Mining Fuzzy Multiple-Level Association Rules." *World Applied Sciences Journal* 17(12).

Nyairo, Risper et al. 2022. "Socio-Economic Trajectories, Urban Area Expansion and Ecosystem Conservation Affect Global Potential Supply of Bioenergy." *Biomass and Bioenergy* 159.

O'Neill, Brian C. et al. 2020. "Achievements and Needs for the Climate Change Scenario Framework." *Nature Climate Change* 10(12).

Pampoore-Thampi, A, A S Varde, and D Yu. 2021. "Mining GIS Data to Predict Urban Sprawl." *arXiv preprint arXiv:2103.11338*.

Peters, Debra P.C. et al. 2014. "Harnessing the Power of Big Data: Infusing the Scientific Method with Machine Learning to Transform Ecology." *Ecosphere* 5(6).

Reimer, Andrew P., and Elizabeth A. Madigan. 2019. "Veracity in Big Data: How Good Is Good Enough." *Health Informatics Journal* 25(4).

Shyu, Chi Ren, Matt Klaric, Grant Scott, and Wannapa Kay Mahamaneerat. 2006. "Knowledge Discovery by Mining Association Rules and Temporal-Spatial Information from Large-Scale Geospatial Image Databases." In *International Geoscience and Remote Sensing Symposium (IGARSS)*,.

Silva, Nisansa De, Sriganesh Lokanathan, and Gabriel Emanuel Kreindler. 2016. "The Potential of Mobile Network Big Data as a Tool in Colombo's Transportation and Urban Planning." *Information Technologies & International Development* 12(2).

Hana Bernika Sabila, Feri Candra. (2023). Implementation of Apriori Algorithm for Data Mining on Sales Transaction Data. https://www.semanticscholar.org/paper/550938948b4d2ab9e69d613a7fc0f8638ba85519

Haoyu Xie. (2021). Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets. 09, p. 458-468. https://www.semanticscholar.org/paper/5bfff2b82fcf0431ec3f46b7028b9e7f8981c5b7

Longzhi Yang, Jia Hu, Che-Lun Hung. (2021). Special issue on recent advances in data science

and systems. 38. https://www.semanticscholar.org/paper/623dcfb247fc3881a153d0e170c3b59d0f3ecb02

Jiahao Xia, Gavin Gong, Jiawei Liu, Zhigang Zhu, Hao Tang. (2024). Pedestrian-Accessible Infrastructure Inventory: Enabling and Assessing Zero-Shot Segmentation on Multi-Mode Geospatial Data for All Pedestrian Types. 10. https://www.semanticscholar.org/paper/5ae6c64a717eb355cf70991d70f2b873503572ef

Tosin Harold Akingbemisilu. (2024). A Critical Evaluation of Government Role in Spatial Data Infrastructures for Healthcare Decision-Making. https://www.semanticscholar.org/paper/cd0752367cd653748b93a7542aa8436e882dc7d2

Terry Devara. (2023). Opportunities and Challenges of Remote Sensing, Geospatial Data, and Machine Learning in Obtaining Accessibility and Location Information for Sustainable Development in Indonesia. https://www.semanticscholar.org/paper/07bee86825016f2ce77505e01a1076c00aa62d55

Deren Li, Shuliang Wang, Deyi Li. (2016-03-23). Spatial Data Mining. Springer. https://play.google.com/store/books/details?id=dq_WCwAAQBAJ&source=gbs_api

Viet-Ha Nhu, Duong Cao Phan, Pham Viet Hoa, Pham The Vinh, D. Bui. (2024). A New Approach Based on Deep Neural Networks and Multisource Geospatial Data for Spatial Prediction of Groundwater Spring Potential. 12, p. 26344-26363. https://www.semanticscholar.org/paper/822b8c0e0b0460c794c10acb4f85841fd9683ad4

Matthias Rüdiger, David Antons, Amol M. Joshi, T. Salge. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. 17. https://www.semanticscholar.org/paper/a105c338840f326888e02bd00a28d6564f42ca23

Appendix A: Ethical Approval

Strathmore UNIVERSITY

15th May 2023

Mr Otieno Isaac,
isaac.chann@strathmore.edu

Dear Mr Otieno,

RE: Big Data Solutions to Environmental Sustainability Issues in Port Cities in the Global South

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1732/23.** The approval period is from **15th May 2023 to 14th May 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Mr Ambrose Rachier,**
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
15-May-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix B : Similarity Report