Outlier Detection & Reconstruction of Lost Big Earth Data Using Machine Learning

by

Muhammad Yasir Adnan

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

July 7, 2025

College of Science and Engineering
University of Derby

Outlier Detection & Reconstruction of Lost Big Earth Data using Machine Learning

by

Muhammad Yasir Adnan

Director of Studies: Stephan Reiff-Marganiec
1st Supervisor: Richard Self

School of Computing and Engineering

July 7, 2025

Faculty of Science and Engineering
University of Derby

Acknowledgments

First and foremost, I wish to extend my deepest gratitude to my wife, whose unwavering support and understanding have been the bedrock of my perseverance throughout the journey of this PhD. Her constant encouragement, patience, and belief in my capabilities have been invaluable, providing me with the strength and motivation needed to pursue and fulfill my academic ambitions. Her role in this process has been immeasurable, and for that, I am eternally grateful.

I would also like to express my heartfelt thanks to my parents and my siblings, whose prayers and endless love have surrounded me with positivity, faith and occasional distractions have been a welcome relief and a source of joy throughout this demanding period.

To my friends, who have become more like a family during this journey, especially in the context of living abroad away from my family, your camaraderie, support, and faith in my work have been a constant source of comfort and motivation. The distance from home has been challenging, but the moments we shared, the discussions we had, and the memories we created together have enriched this journey in ways words cannot fully capture, making a foreign land feel like home.

I am profoundly grateful to my supervisory team, Professor Stephan Reif-Marganiec, Professor Yong Xue and Senior Lecturer Richard Self, for their guidance, patience, and invaluable insights throughout this research. The expertise and thoughtful advice have been crucial in shaping the direction and execution of this work. Prof. Stephan's support has not only been academic but also motivational, pushing me to achieve excellence while navigating the challenges of research. His mentorship has been a significant pillar of my PhD journey.

This dissertation is not just a reflection of my efforts but a testament to the love, support, and faith of all the incredible individuals mentioned above. Their contributions to my life and this work are deeply appreciated and will always be remembered with gratitude.

Abstract

This dissertation thoroughly examines enhancing outlier detection and reconstruction techniques for Earth Observation (EO) datasets, specifically focusing on Land Surface Temperature (LST) values. Addressing both outlier detection and data reconstruction is crucial for EO and LST data analysis because undetected anomalies can distort temperature patterns, and incomplete data reduces the reliability of environmental assessments.

This research focuses on addressing important difficulties related to the collecting, processing, and analysis of LST data, which is in high demand for environmental monitoring and decision-making. In particular, the high variability of EO data, presence of noise and missing values, and the large volumes of satellite imagery pose significant challenges requiring robust and scalable methods.

This effort focuses on identifying and setting boundaries for the study area, particularly the Beijing-Tianjin-Hebei (BTH) region. A high-level research method is adopted, where image raster data are processed in ArcGIS and then transformed into tabular format suitable for machine learning, enabling systematic detection and correction of anomalies.

This thesis presents new techniques for improving the accuracy of temperature intensity representations and enabling effective statistical learning by processing image raster data in ArcGIS and converting them into a tabular format suitable for machine learning analysis. These techniques significantly reduce reconstruction errors, enhancing both data completeness and usability.

The use of self-supervised learning models, specifically the TabNet regressor, is a major advancement in improving the forecasting and rectification of anomalies in LST datasets. Empirical tests show a notable increase in anomaly detection precision and a reduction in data gaps, indicating a high level of success for these methods.

The study addresses problems related to the complexity of EO data and the model's adaptability to varied datasets and situations. Despite challenges, devel-

oping and verifying a unique tabular dataset for the study area has been crucial in establishing a standard for anomaly detection, thus improving the usefulness and reliability of LST data for environmental research and monitoring. By focusing on present contributions, this dissertation demonstrates how robust outlier detection and data reconstruction methods can effectively support environmental monitoring tasks.

The developed techniques have been tested in the EO data context, but will be applicable to other image-based data with similar underlying characteristics such as obscured areas. This immediate applicability underscores the real-world impact and relevance of the contributions within the scope of this thesis.

This dissertation makes a substantial contribution to the subject of big-earth data analysis by introducing creative methods for identifying outliers and reconstructing data. These methods enhance the quality and dependability of Land Surface Temperature datasets and serve as a validated solution for improving data integrity in current EO applications.

Declaration

I declare that the thesis here submitted is original, except for the sources explicitly acknowledged. This thesis as a whole or any part of it has not been previously submitted for the same degree or for a different degree other than the Doctor of Philosophy (Ph.D.) at the University of Derby. Parts of this thesis have appeared in papers, specifically parts of chapter 5 are published in [1], and parts of chapter 6 are submitted to a journal.

Muhammad Yasir Adnan

University of Derby, 2025

Table of Contents

Acknowledgments i

Abstract ii

Declaration iv

List of Tables and Figures viii

1 Introduction 1

1.1 Background . . . 3

1.2 Motivation . . . 4

1.3 Problem Statement . . . 4

1.4 Aim and Scope of the Research . . . 5

1.4.1 Aim . . . 5

1.4.2 Objectives . . . 5

1.4.3 Significance of the Research . . . 6

1.5 Structure of the Thesis . . . 7

2 Background 10

2.1 Introduction . . . 10

2.2 Remote Sensing Image Acquisition . . . 10

2.2.1 Remote Sensing Image Processing . . . 12

2.3 Anomalies in Datasets . . . 15

2.3.1 Type of Outliers . . . 16

2.3.2 Anomalies in Spatial Data . . . 18

2.4 Missing Information in Remote Sensing . . . 18

2.4.1 Big Earth Data Processing . . . 20

2.5 Impact of Data Quality on Environmental Decision Making . . . 22

2.6 Summary 23

3 Literature Review **25**

3.1 Introduction 25

3.2 Anomalies in Datasets 25

3.2.1 Type of Outliers 26

3.3 Anomalies in Spatial Data 27

3.3.1 Spatial Outlier Detection and Removal 28

3.3.2 Suitability of Outlier Detection Techniques for Spatial-Temporal Data 32

3.4 Quantitative Remote Sensing 34

3.4.1 Errors in Quantitative Remote Sensing 35

3.4.2 Remote Sensing Information Reconstruction 37

3.5 Conclusion 56

4 Research Methodology **59**

4.1 Introduction 59

4.2 Dataset Overview 60

4.3 Exploratory Data Analysis 61

4.3.1 Spatial and temporal trends 61

4.3.2 Outliers in the Data 61

4.3.3 Seasonal LST Distribution 63

4.4 Evaluation Metrics 64

4.4.1 OSR Evaluation Metrics 64

4.4.2 TabNet Evaluation Metrics 66

4.5 Conclusion 67

5 Outlier Detection and Reconstruction of Lost Land Surface Temperature data in Remote Sensing imagery **68**

5.1 Introduction 68

5.2 Proposed Model 69

5.2.1 Outlier Detection 69
5.2.2 Missing Data Reconstruction 71
5.3 Experimental Results 73
5.3.1 Outlier Detection Evaluation 74
5.3.2 Reconstruction Evaluation 75
5.3.3 Bland–Altman Plot Across Scenarios 77
5.4 Conclusion 79

6 Outlier Detection and Reconstruction in Big Earth Data Using Tabular Self-Supervised Learning 81

6.1 Introduction 81
6.2 Data Collection and Processing 82
6.2.1 Data Collection 82
6.2.2 Data Processing 82
6.3 Experimental Results 84
6.3.1 Exploratory Data Analysis of Generated Dataset 84
6.3.2 Temporal Trend Analysis 85
6.3.3 Seasonal Trend Analysis 88
6.4 Proposed Model 89
6.4.1 TabNet Architecture 91
6.5 Experiment & Results 94
6.5.1 Bland–Altman Plot Evaluation for TabNet 99
6.5.2 Comparison Between TabNet and OSR Reconstruction Results 101
6.6 Conclusion 101

7 Discussion, Conclusion and Future Works 103

7.1 Discussion 103
7.2 Challenges and Limitations 104
7.3 Conclusion 105
7.4 Future Works 106

List of Tables and Figures

Table 3.1 Comparison of Outlier Detection Techniques 57

Table 3.2 Comparison of Data Reconstruction Techniques 58

Table 5.1 Scenario 1 — Varying Window Sizes and Outlier Counts 74

Table 5.2 Scenario 2 — Fixed 30 Outliers with Varying Window Sizes . . 74

Table 5.3 Scenario 3 — Increasing Outliers with Fixed 10×10 Window . 75

Table 5.4 Scenario 1 — Different Window Sizes with Varying Outlier Counts 75

Table 5.5 Scenario 2 — Fixed 30 Outliers with Different Window Sizes . 76

Table 5.6 Scenario 3 — Fixed 10×10 Window with Increasing Number of Outliers 76

Table 6.1 Tabular LST data of BTH Region 84

Table 6.2 Seasonal Average Temperatures 88

Figure 2.1 Basic Structure of Remote Sensing System 12

Figure 2.2 Remote Sensing Image Analysis 13

Figure 2.3 Illustration of Anomaly/Outlier 16

Figure 2.4 Classification of Outliers 17

Figure 2.5 Missing Information in Remote Sensing Images 19

Figure 3.1 Classification of Outliers 27

Figure 3.2 Exemplar-based Region Filling and Object Removal 42

Figure 3.3 Flowchart of multi-temporal similar replacement-based TSAM 54

Figure 4.1 Study Area of Beijing-Tianjin-Hebei region 60

Figure 4.2 Scatter Plots of LST from January 2017 to June 201 61

Figure 4.3 Box Plot of Land Surface Temperature (LST) with Outliers and Median Values 62
Figure 4.4 Density Plot of Land Surface Temperature (LST) by Season . 63

Figure 5.1 Mosaic image constructed by stacking $M \times N$ windowed views of Input Images 70
Figure 5.2 Sequential Daily MODIS Images Across a Four-Day Period . . 73
Figure 5.3 Outlier and Reconstructed Image Pair 77
Figure 5.4 Bland–Altman Plot for Scenario 1: Varying window sizes and outlier counts. 10×10 window exhibits minimal bias and narrow agreement range. 78
Figure 5.5 Bland–Altman Plot for Scenario 2: Fixed 30 outliers. 10×10 window shows superior consistency and reduced bias. 78
Figure 5.6 Bland–Altman Plot for Scenario 3: Fixed 10×10 window with increasing outlier count. Spread increases with more outliers, reducing accuracy. 79

Figure 6.1 Data Collection & Processing 83
Figure 6.2 Daily LST Average Trend 85
Figure 6.3 Daily average Land Surface Temperature (LST) with El Niño (red) and La Niña (blue) 86
Figure 6.4 Weekly Average LST Trend 87
Figure 6.5 Trend of Average Land Surface Temperature 87
Figure 6.6 Seasonal Average LST Trend Analysis 88
Figure 6.7 : TabNet encoder architecture 92
Figure 6.8 Feature and Attentive Transformer 94
Figure 6.9 Self Supervised Learning on Tabular LST Data 95
Figure 6.10 Prediction Error Plot: Actual vs. Predicted LST Values . . 97
Figure 6.11 Actual vs. Predicted LST Values 98
Figure 6.12 Residuals of Predictions (Sampled) 99

Figure 6.13 Bland–Altman plot of TabNet predictions vs. true LST values on the test dataset. 100

List of Publications

Conference Papers

1. Adnan, M. Y., Xue, Y., & Self, R. (2022). Outlier Detection and Reconstruction of Lost Land Surface Temperature Data in Remote Sensing. In Proceedings of the 12th International Conference on Computer Science and Information Technology (CCSIT 2022) (Vol. 12, Issue 13, pp. 197).

Chapter 1

Introduction

The challenge of handling and examining vast quantities of data, commonly referred to as 'Big Data', has become increasingly crucial. The field of remote sensing has undergone rapid growth, establishing itself as a fundamental tool for comprehending and analysing the Earth's ever-changing processes.

This research specifically addresses two crucial problems: (a) accurate outlier detection to identify and separate irregular data points from legitimate observations, and (b) effective reconstruction of missing values to ensure continuity and reliability in Earth Observation (EO) and Land Surface Temperature (LST) datasets [2].

EO datasets, particularly those related to Land Surface Temperature, are highly dimensional and typically cover extensive geographical regions with significant variability. Complexity arises from the vast volume of satellite imagery data, typically involving millions of pixel-based observations daily. Additionally, the presence of missing data and outliers significantly complicates data analysis, affecting environmental assessments and decision-making processes reliant on accurate LST measurements.

The utilisation and evaluation of land surface temperature (LST) data using remote sensing methods represent a distinct convergence of computer science. and Geographical Information Systems (GIS). Nevertheless, the process of integrating these components presents certain difficulties. An important concern in the analysis of LST data is the existence of outliers, which can greatly skew research findings and result in inaccurate conclusions. To tackle these issues, it is necessary to employ advanced data processing techniques and algorithms to guarantee

the precision and dependability of the data. This research focuses on the idea of outliers, which refers to data points that exhibit considerable deviation from the usual trend within a dataset [3]. Ensuring the integrity and accuracy of data is a crucial concern in the context of remote sensing. The identification and handling of outliers extend beyond simple data cleaning and encompass complex challenges that require advanced algorithms and analytical methodologies. These problems are crucial for ensuring the dependability of environmental evaluations and predictions obtained from remote sensing data. Moreover, this study explores the concept of outlier reconstruction, which encompasses techniques aimed at reconstructing abnormalities in the datasets induced by anomalous data. This aspect highlights the interdisciplinary nature of the work, connecting computer science approaches with actual applications in Earth observation. The accurate identification and reconstruction of outliers in remote sensing data is of utmost importance. These processes are essential for preserving the quality and dependability of the data, which in turn has a substantial impact on the decisions and policies made using this knowledge. The objective of this thesis is to investigate and propose innovative methods for identifying anomalies in remote sensing data and then reconstructing them to ensure the reliability and precision of the dataset. By doing so, this research also enhances the effectiveness of remote sensing to understand and address environmental concerns on Earth by prioritising these features.

Outliers, missing values, and anomalies are often encountered terms in remote sensing data analysis, typically used interchangeably, however each possesses distinct implications within this research environment. An outlier denotes a data point that markedly diverges from the overall trend [4], typically signifying possible inaccuracies or excessive fluctuations. Missing values arise when anticipated data points are absent owing to sensor faults, meteorological obstructions like clouds, or transmission problems [5]. This study defines anomalies as encompassing both outliers and missing values, signifying any irregularity or contradiction

within the dataset [6].

1.1 Background

Remote sensing has become an essential tool in various disciplines, such as environmental studies, oceanography, and surveying. The significance of remote sensing data in making precise forecasts and analyses is emphasised. Nevertheless, this dependency is full of challenges. The presence of errors in geographical data, which can originate from factors such as attribute definition, data sources, data modelling, and analysis procedures, is just as significant as the accuracy of the actual data measurement tools. These flaws might appear as systematic, random, or significant errors, resulting in data that is incomplete, erroneous, repetitive, and contradictory. Furthermore, the presence of anomalies arising from many data sources is a distinct challenge, frequently resulting in the concealment and overwhelming of the data [7]. Atmospheric conditions play a vital role in the information-capturing abilities of remote sensing instruments. These instruments help to gather information about the atmosphere, ocean, and land surface and are the most frequently used way of gathering information. But due to certain problems with remote sensing instruments and diverse environmental conditions, the acquired information is incomplete, or we can say it has missing information in it, which greatly reduces the usability of the data. There are many types of missing information, broadly classified into these:

- Sensor Failure
- Cloud Obscuration

In the broader context of big data analytics and decision-making, data cleaning emerges as a major aspect, particularly in the field of remote sensing. Sensor failures, for example, can lead to missing information and outliers [3], as seen with some detectors in the MODIS Aqua band 6, which provided malicious readings [8]. The significance of satellite remote sensing (SRS) is highlighted in studies

related to weather predictions and the effects of global warming. However, the acquisition of remote sensing data is prone to limitations caused by weather conditions, with a substantial portion of the Earth's surface often covered by clouds, reducing visibility and data accuracy [9]. In light of these challenges, the reconstruction of missing data and the removal of outliers have become crucial aspects of the analysis of Big Earth data, ensuring the reliability and usability of the data for various applications.

1.2 Motivation

This research is motivated by the critical need to guarantee the precision and dependability of land surface temperature (LST) data acquired using remote sensing. The LST data is crucial in various environmental, climatological, and geographical research. Nevertheless, the integrity and practicality of this data are frequently undermined by discrepancies, omissions, and anomalies. These problems can originate from diverse sources, such as sensor faults, meteorological factors like cloud cover, and the inherent unpredictability of the Earth's surface.

1.3 Problem Statement

The primary focus of this research is to tackle the presence of outliers and missing data in Land Surface Temperature (LST) data acquired using remote sensing, termed anomalies. Anomalies in LST (Land Surface Temperature) data pose a substantial obstacle since they have the potential to skew the understanding and examination of this vital environmental data. Outliers can arise from multiple sources, including sensor inadequacies, climatic variables like cloud cover, and the fundamental complexity of the Earth's surface.

Outliers and extreme values, although they are commonly used interchangeably, have fundamentally different causes and implications. An outlier is commonly characterised as a data point that exhibits a substantial deviation from the other

observations in the collection, potentially suggesting a measurement error or an uncommon occurrence. On the other hand, extreme numbers are those that, while uncommon, are still within the realm of possibilities and may indicate inherent fluctuations in the data. Missing data can arise from several factors, including weather conditions and equipment malfunction limitations.

1.4 Aim and Scope of the Research

1.4.1 Aim

To propose a novel approach to address the prevailing challenges in enhancing the detection of outliers and the reconstruction of lost Big Earth data, thereby contributing to improved accuracy and reliability of Earth observation datasets.

1.4.2 Objectives

1. - Define and delimit the Beijing-Tianjin-Hebei (BTH) study region, acknowledging the challenge of acquiring consistent and high-quality satellite data over vast and varied regions. - Identify and select an appropriate satellite, recognizing potential discrepancies and gaps in satellite data coverage. - Systematically collect and preprocess satellite data, specifying the satellite product and ensuring data integrity and quality.

2. - Develop robust methodologies to accurately detect outliers within Earth Observation (EO) datasets by processing the collected image raster data using ArcGIS, thus handling complexities arising from heterogeneity and dimensionality. - Establish reliable thresholds through extensive experimentation to facilitate clear pixel-to-pixel comparisons and enhance outlier detection accuracy.

3. - Propose and validate reconstruction methods that address missing data by emphasizing spatial-temporal correlations and utilizing neighboring pixel

values. - Provide clear methodological steps and statistical metrics for validating reconstruction accuracy, enhancing the integrity and reliability of reconstructed data.

4. - Transform high-dimensional satellite raster data into a structured tabular format suitable for machine learning analysis, overcoming challenges associated with data complexity and facilitating efficient statistical learning and visualization.

5. - Employ and evaluate advanced self-supervised learning models, specifically TabNet, to rigorously validate reconstructed data. - Assess the model's performance by comparing predicted values against reconstructed image values using established metrics like Mean Squared Error (MSE), F1 score, and Bland-Altman plots, thereby ensuring reliability and accuracy.

1.4.3 Significance of the Research

This thesis represents an important step forward in the processing and analysis of large-scale Earth data, with the goal of enhancing the quality and usefulness of Earth observation (EO) datasets. The research addresses significant obstacles in this domain, providing multiple essential contributions.

The main objective of this research is to develop sophisticated techniques for identifying outliers in extensive Earth data. Detecting data points that differ significantly from the average is essential to preserving the accuracy and dependability of Earth observation datasets. Accurate identification of exceptional data points is crucial for maintaining data integrity, which, in turn, facilitates diverse scientific studies and applications.

Another notable accomplishment of this research is the restoration of data that was previously lost or corrupted. The work improves the comprehensiveness and usefulness of Earth observation data by retrieving missing information. The restoration is crucial for obtaining a comprehensive understanding of environ-

mental analysis and guaranteeing that datasets are suitable for various tasks.

The development of a unique tabular dataset specifically designed for the analysis of large-scale Earth data. This dataset addresses a deficiency in current data resources and enables the utilisation of deep learning models that are specifically designed for tabular data in the analysis of Earth data. The presence of such a dataset is a significant advancement in this field, providing fresh prospects for research and innovation.

Furthermore, the research indicates significant advancements in identifying and reconstructing anomalous data points using deep learning techniques. This signifies a substantial enhancement compared to conventional approaches, providing more resilient and effective solutions. The incorporation of deep learning in this particular setting presents fresh opportunities for sophisticated data analysis and processing.

This research has a significant and wide-ranging influence. Effectively tackling these obstacles not only raises the calibre of Big Earth data but also amplifies its applicability in several scientific fields. Dependable and precise Earth observation databases are essential for making well-informed judgements in crucial fields such as climate studies, urban planning, and environmental conservation. The research, which concentrates on identifying and restoring anomalies, has the capacity to significantly contribute to these wider scientific and societal goals. The results of this thesis, which improve the accuracy and usefulness of data, have the potential to significantly influence the development of future environmental policies and initiatives.

1.5 Structure of the Thesis

This thesis is organized into seven main chapters, each contributing a unique perspective to the intersection of computer science and Earth Observation (EO), with a focus on remote sensing data analysis, particularly concerning outliers and missing data. The structure is as follows:

- Chapter 1: Introduction Provides an overview of the thesis, introducing the main research topics and setting the context for the study. This chapter includes the background, motivation, problem statement, and the aims, objectives, and significance of the research.

- Chapter 2: Background Offers a detailed exploration of remote sensing, including an overview of the datasets used, the principles of remote sensing image processing, types of anomalies in datasets, and methods for handling missing information. This chapter also discusses exploratory data analysis techniques applied to the datasets.

- Chapter 3: Literature Review Presents a comprehensive review of the existing literature on anomalies in datasets, spatial data analysis, remote sensing information systems, and various aspects of data processing and reconstruction in the field of remote sensing.

- Chapter 4: Research Methodology Describes the overall research design, including dataset selection, data collection procedures, preprocessing steps, exploratory data analysis methods, and the evaluation metrics used to assess both outlier detection and reconstruction models.

- Chapter 5: Outlier Detection and Reconstruction of Lost Land Surface Temperature Data in Remote Sensing Imagery Introduces and discusses the proposed models for outlier detection and missing data reconstruction, along with experimental results validating these models.

- Chapter 6: Outlier Detection and Reconstruction in Big Earth Data Using Self-Supervised Learning Explores advanced models for data processing, specifically the use of self-supervised learning techniques (e.g., TabNet) for outlier detection and data reconstruction in large-scale EO datasets.

- Chapter 7: Discussion, Conclusion and Future Works Provides a thorough discussion of the results, implications, and significance of the research find-

ings, along with the challenges encountered and the solutions developed. Concludes the thesis by summarizing the key findings and contributions of the research, and outlines potential future directions and areas for further investigation.

Chapter 2

Background

2.1 Introduction

In the rapidly evolving field of remote sensing and Earth observation (EO), there is a growing demand for accurately determining climate variables with detailed spatial and temporal resolution. This is especially important in regions where there are not enough meteorological observations available for effective environmental monitoring and modelling. The importance of this requirement is emphasised by the constraints of conventional climate reanalysis datasets, which frequently fall short of capturing the subtle fluctuations of climate at smaller scales. As we explore the domain of massive Earth Observation (EO) data, which is characterised by its substantial volume derived from Earth observations and climate models, we encounter obstacles that are as vast as the data itself. To address these difficulties, it is essential to implement creative solutions, particularly by combining cloud computing and machine intelligence. These technologies provide scalable and efficient computer resources and advanced analytical capabilities, respectively, which are essential for analysing and helping in outlier detection and reconstruction of lost big earth data.

2.2 Remote Sensing Image Acquisition

Remote sensing is the process of acquiring data without getting physically in contact with the object. Remote sensing data acquisition involves sensors mounted on satellites, aircraft, or drones capturing electromagnetic radiation reflected or

emitted from Earth's surface [10] as shown in 2.1 . These data are then transmitted through satellite communication channels to ground stations for storage and archival purposes. Common storage formats include GeoTIFF, HDF, and netCDF, each selected for specific strengths: GeoTIFF supports geo-referenced imagery and compatibility with Geographic Information Systems (GIS), HDF is ideal for complex multi-dimensional data, and netCDF efficiently handles large arrays of scientific data, especially suited for time-series analysis [11]. Effective handling of large remote sensing datasets requires extensive preprocessing to address challenges such as sensor noise, calibration inaccuracies, atmospheric interference, and data compatibility. These preprocessing steps are crucial for ensuring the reliability and accuracy of analyses, as neglecting them can lead to misleading results.

The remote sensing platforms consist of the equipment or vehicles used to capture data. The sensors mounted have a number of characteristics, including time of image accusation, distance from the object,interval between accusations of image location, and range of coverage. The remote sensing platforms have been divided into three broad categories[12].

- Ground based platform
- Airborne platform
- Satellite Platform
- Static Platform
- Mobile Platform

Satellite-based platforms are mainly used for meteorological data acquisition. The first successful weather satellite was launched on April 1st, 1961 [13].

Since then, the advancement of satellites and sensors to capture data has opened new frontiers of research in remote sensing. These advancements are also subject to some limitations that introduce outliers in remote sensing imagery. Therefore,

remote sensing imagery goes through a number of different processing steps before it can be utilised for an actual task. The basic structure of a remote sensing system is shown in the figure.

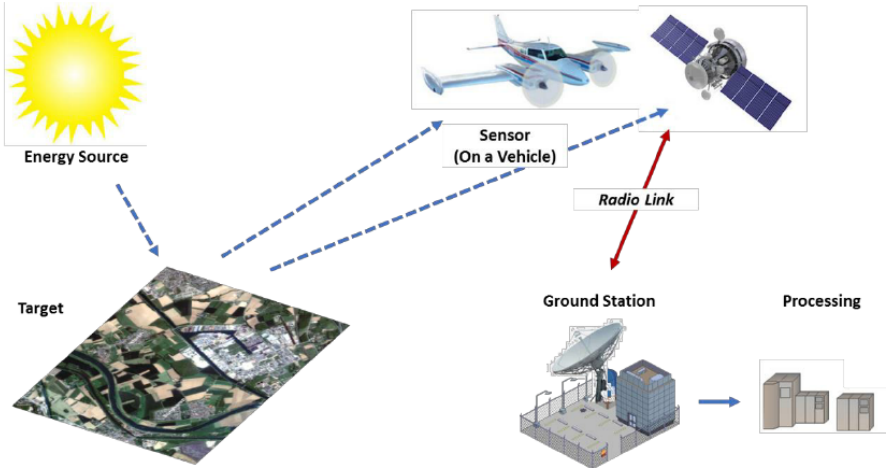

Figure 2.1: Basic Structure of Remote Sensing System

2.2.1 Remote Sensing Image Processing

The utilisation of digital image processing in remote sensing is essential for rectifying mistakes and improving image quality. Diverse methodologies are utilised to mitigate problems including noise, distortion, and clarity, hence enhancing the interpretability of remote sensing data. The subsequent sections describe essential image processing techniques employed in this domain, mainly divided into four categories [14].

- Pre-processing
- Enhancement
- Transformation
- Classification

The smallest unit is pixel in which images are represented. Remote sensing image analysis helps to obtain meaningful information from the images. The figure shows a general procedure for remote sensing image analysis.

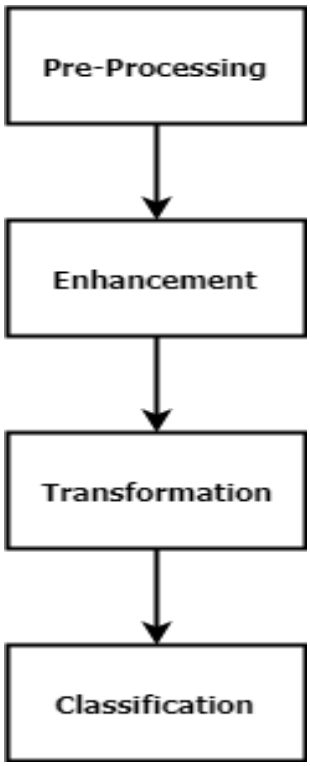

Figure 2.2: Remote Sensing Image Analysis

Image Pre-processing

This is the first and foremost step prior to data analysis [15]. This initial processing is carried out to remove any distortions caused by the equipment used or imaging conditions. It involves radiometric correction,geometric correction, and atmospheric correction. Radiometric correction smooths the global radiation reflection in the image. [16] is an example of a radimetric correction strategy for satellite images. Geometric correction removes the distortion caused by the angle of the sensor from which the image was taken or the distortion caused by the earth's rotation. [17]. A typical example of geometric correction by [18] proposes a correction function based on orthogonal polynomials. The weather conditions tend to hinder the clear view of the scene from the satellite, for which atmospheric correction is necessary. [19] is an example of atmospheric correction in satellite imagery.

Image Enhancement

Image enhancing refers to the idea of improving the information content of an image before getting into processing [20]. Image enhancement can be broadly classified into two categories [21].

- Spatial enhancement
- Frequency domain enhancement

Spatial domain enhancement works at the pixel level for the enhancement of the image, while frequency domain enhancement first takes a Fourier transform of the image and performs enhancement on this transform. After that, the inverse Fourier transform is used to get the final image. Some of the techniques used for image enhancement are [22].

- Contrast Enhancement
- Composite Generation
- Digital Filtering

Contrast enhancement stretches the original grey levels of the image for better visualization. It is also known as contrast stretching. Composite generation combines different bands from the image and also uses contrast stretching to produce an image for visual analysis. Digital filtering or spatial filtering removes the blurring effect by making edges clearer for a better and crisper image. [23]. Image restoration is also a form of image enhancement that deals with the reconstruction or recovery of lost or degraded pixels.

Image Transformation

Image transformation typically uses information from other bands of the same image or multiple images of a similar area acquired at different intervals. It produces a new image with better information as compared to the original image [14, 24].

Typically, arithmetic operations, including subtraction, addition, multiplication, and division, are used in image transformation.[25].

Image Classification

Image classification methods are mainly used in image segmentation and object detection [26]. Image classification helps understand and differentiate different types of land cover in a remotely sensed image. The image classification procedure is divided into two types [27].

- Supervised
- Unsupervised

Classification of pixels based on training data or images so that every pixel is associated with a certain group based on its spectral characteristics refers to supervised classification. While an unsupervised classifier automatically classifies a pixel and forms clusters of relevant spectral characteristics, Unsupervised classification, though not directly applied in this research, refers to grouping pixels based on spectral similarity without prior training data. It contrasts with the anomaly detection techniques used herein, which do not aim to classify pixels but rather identify significant deviations from expected patterns

2.3 Anomalies in Datasets

Over the last few decades, big data has gotten the attention of industry, academia, government, and other organisations. During the analysis of real-world data, a common practice is to identify data or instances that are outliers or dissimilar to the actual data. These are known as anomalies [7]. Normally, anomalies give the idea of erroneous values in the dataset, but anomalies can introduce two factors in any dataset, given as [3]:

Outlier: Anomalies frequently introduce outliers, which represent data points inconsistent with the general behavior or pattern of the dataset. Such

outliers may arise due to measurement errors, data corruption, or genuine extreme variations [7].

Novelties: Novelties refer to previously unrecognized patterns or new data characteristics. Novelty detection focuses on recognizing previously unseen patterns, assigning novelty scores and thresholds to distinguish new patterns from familiar ones [28].

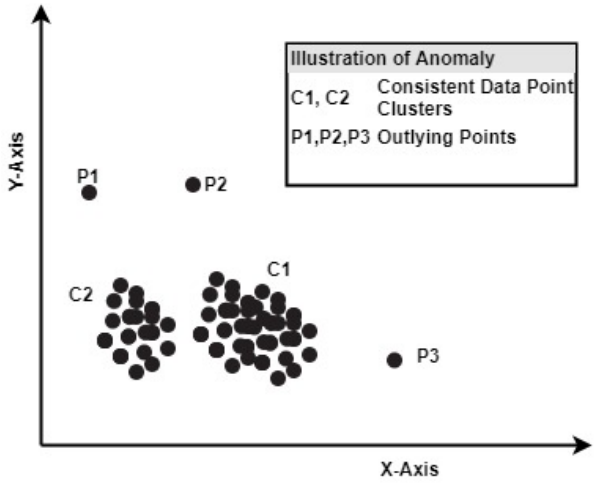

Figure 2.3: Illustration of Anomaly/Outlier

Figure 2.3 explicitly illustrates the concept of anomalies within datasets, focusing primarily on outliers. The clusters labeled C_1 and C_2 represent regions with consistent and typical data observations, showing densely grouped data points. Points labeled P_1, P_2, and P_3 are depicted as anomalies or outliers because they significantly deviate from the established normal data patterns of clusters C_1 and C_2. Such outliers can negatively impact data quality and analytical accuracy, making their detection and subsequent handling crucial for robust environmental monitoring and remote sensing data processing.

2.3.1 Type of Outliers

In general, outliers are divided into three categories, and their detection strategies varies the same.[29]. A classification of outliers is shown in figure 2.3.

Point Outliers

Points outliers are considered data points that are far away from the rest of the data. Consider points P1, P2, or P3 in Figure 1. These are examples of point outliers.

Contextual Outliers

Contextual outliers are also known as conditional outliers [30]. For example, a temperature peak of 70 °F would be considered normal in the summer, but it would be treated as an outlier in the winter. The context in which the outliers are being identified must be specified beforehand to proceed with outlier detection. In particular, data instances are divided by two sets of attributes.

1. **Contextual attributes.** The contextual attribute defines the context for a data instance. For example, in the case of time series data, time is a contextual attribute.

2. **Behavioural attributes.** Behavioural attributes are also known as non-contextual attributes, which determine the non-contextual characteristics of a data point.

Collective Outliers

One or more data point gather and form a cluster which is anomalous as compared to entire dataset, known as collective outliers. The individual data point in that cluster may not be outlier but due to their occurrence they are treated as collective outliers.

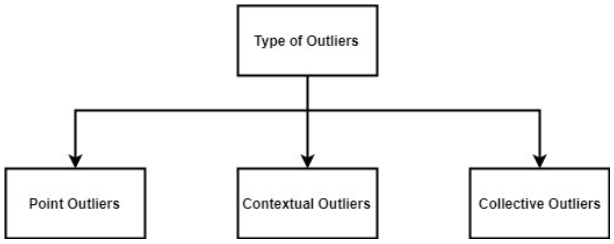

Figure 2.4: Classification of Outliers

2.3.2 Anomalies in Spatial Data

Remote sensing has been widely used for the past few decades for environmental, oceanic, and surveying applications. All the applications are entirely dependent on remote sensing data, how it is acquired, and how accurately we can make predictions with this data. Errors in spatial data are as important as actual data, measurement, and data structure. These errors mainly occur due to attribute definition, data sources, data modelling, and the analysis process. Spatial data mainly consists of systematic error, stochastic error, and gross errors, such as incomplete, inaccurate, redundant, and inconsistent data and anomalies due to multiple sources of data. Masking and swamping effects are common while detecting outliers. Masking states the occurrence of a second outlier after detection and removal of the first outlier, and swamping occurs when a correct data point is considered an outlier due to the actual outlier [31].

2.4 Missing Information in Remote Sensing

Missing information is another huge source of inaccurate or incomplete data in remote sensing, apart from outliers. Atmospheric conditions play a vital role in the information-capturing abilities of remote sensing instruments. These instruments help to gather information about the atmosphere, ocean, and land surface and are the most frequently used way of gathering information. But due to certain problems with remote sensing instruments and diverse environmental conditions, the acquired information is incomplete, or we can say it has missing information in it, which greatly reduces the usability of the data. Various categories of missing information can be broadly described as follows:

- Sensor Failure
- Cloud Obscuration

Sensor Failure

Sensors are crucial in remote sensing and data collection. Missing information results from the failure of these equipment. 15 out of 20 detectors in MODIS Aqua Band 6 provide inaccurate data. [32]. [33] is a common instance of data retrieval failure due to sensor malfunction. Remote sensing utilises two types of sensors:

- Active Sensors
- Passive Sensors

Active sensors utilise their own source of energy in the form of radiation to illuminate the object under observation. They can measure the energy reflected back from the object. On the other hand, passive sensors use a natural source of energy in the form of radiation in the form of sunlight to capture objects [34]. Sensor malfunctions lead to the phenomenon of stripping in remote sensing images. An example of striping in a remote sensing image is shown in figure 3 (a,b) [35]:

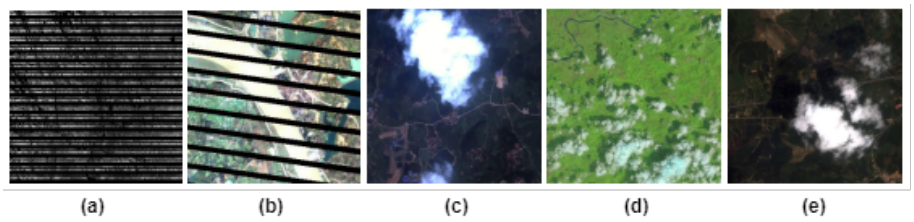

Figure 2.5: Missing Information in Remote Sensing Images

Cloud Obscuration

Data acquisition in passive remote sensing is subject to limitations caused by weather conditions [9]. Cloud cover obstructs the data acquired in remote sensing imagery, thereby limiting the data's utility in future applications. At one point, 35 % of the Earth's surface is covered by clouds [9] , while for specific countries like Canada, 50 % to 80 % of its land is covered with clouds in the

morning [36]. 35% of Landsat ETM+ scenes are affected by cloud cover, resulting in a substantial data loss. [37]. As passive sensors are effected by weather conditions, considering the fact that cloud contamination is inevitable, passive sensor radiation cannot reach the target object, thus hindering the actual desired information. An example image showing cloud obscuration is shown in figure 3 (c,d,e) [38]. [39] considered 10 bands from different data sources to detect thin and thick clouds and proposed a multiscale feature-convolutional neural network. The MF-CNN is used to learn the global features of input images. The input images are combination images from the Landsat 8 satellite. Low-level spatial information and high-level semantic information, which are the essence of MF-CNN, are better suited to acquiring detailed information about clouds, and thus cloud detection at the pixel level is performed accurately.

2.4.1 Big Earth Data Processing

The emergence of Big Earth Data has required the creation of sophisticated processing systems capable of managing intricate geospatial datasets [40]. ArcGIS is a prominent platform that facilitates the use of maps and geographic information through its complete geographic information system (GIS). This software application serves the purpose of generating and utilising maps, consolidating geographical data, examining mapped data, exchanging and uncovering geographical information, and administering geographical information inside a database. These systems are designed to handle the challenges posed by big geospatial data, including the management and analysis of multi-source vector and raster data [41].

ArcGIS operates by structuring GIS data into layers and visual depictions, using spatial and statistical techniques to examine and manipulate data, and ultimately producing intricate visuals citedevelopment-method-of-arcgis-data-processing-tool-1s1ih78ejd. The process entails the amalgamation of diverse data sources, such as satellite imaging, aerial pictures, and extensive databases of geographically ref-

erenced information. The data undergoes processing using advanced algorithms capable of detecting patterns, recognising trends, and generating prediction models.

ArcGIS possesses a range of functionalities that surpass mere data visualisation. It facilitates sophisticated geographic data analysis, empowering users to generate their own geoprocessing scripts. ArcPy, a Python site package, is a very effective tool for performing geographic data analysis, data translation, data management, and map automation [42].

ArcPy offers an efficient method for conducting geographic data analysis, automating geoprocessing activities, and executing intricate map processes in ArcGIS. ArcPy enables users to manage ArcGIS functions and alter map documents with Python scripts, facilitating the efficient conversion of large-scale geospatial data into informative graphics and reports.

The Google Earth Engine (GEE) has become a leading platform for Earth scientific data and analysis, alongside ArcGIS and ArcPy. GEE integrates a vast collection of satellite imagery and geographic datasets, amounting to many petabytes, with the ability to perform analysis on a global scale. Google's cloud architecture enables scientists, academics, and developers to identify alterations, chart patterns, and measure disparities on the Earth's surface.

Google Earth Engine efficiently analyses geospatial data on a large scale and offers a collection of algorithms and computational abilities that simplify intricate raster computations, efficient vector manipulation, and machine learning applications. Its excellent data visualisation capabilities make it an indispensable tool for analysing environmental changes and charting global events [43].

The incorporation of these tools—ArcGIS, ArcPy, and GEE—has transformed the earth observation domain, facilitating the manipulation and examination of extensive earth data with exceptional precision and effectiveness. Consequently, they play a crucial role in generating comprehensive and practical visual representations of the Earth, enhancing our comprehension of the planet's ever-changing

systems.

2.5 Impact of Data Quality on Environmental Decision Making

The quality of data significantly impacts the efficacy of environmental decision-making processes [44]. High-quality remote sensing data, distinguished by accuracy, completeness, and consistency, immediately enables accurate monitoring, informed policy development, and efficient resource management. Conversely, abnormalities include sensor malfunctions, atmospheric disturbances, absent data, and cloud cover can significantly undermine data integrity, resulting in erroneous analysis and consequently misguided choices [45].

In the field of environmental monitoring, reliable Land Surface Temperature (LST) data is very vital. High-quality LST data permit correct identification and management of urban heat islands, enhance effective drought evaluations, and promote informed agriculture practices through precise vegetation monitoring [46]. This thesis examines detection and reconstruction approaches that specifically tackle significant data quality concerns, ultimately improving the dependability of environmental assessments.

The practical benefits of enhanced data quality are apparent in urban planning and management, where accurate identification of temperature anomalies facilitates improved responses to high heat occurrences [47]. In agricultural contexts [48], precise anomaly-free data allow farmers and agricultural planners to forecast crop health, optimise irrigation, and efficiently manage potential drought impacts. Enhanced data quality substantially aids climate modelling and forecasting by diminishing uncertainties caused by absent or aberrant data points. The methods presented in this research, which emphasize robust outlier detection and effective reconstruction of missing data through advanced computer science techniques, significantly enhance the quality of remote sensing datasets. Consequently, they

provide environmental stakeholders and policymakers with dependable information, enabling informed, efficient, and timely environmental decisions.

2.6 Summary

This chapter has presented a thorough summary of the present condition of remote sensing and Earth observation (EO), specifically highlighting the difficulties and approaches involved in determining climatic variables in areas with insufficient meteorological observations. The analysis of large-scale Earth observation data, which is characterised by the vast amount of information obtained from Earth observations and climate models, uncovers substantial challenges that require creative solutions, particularly the incorporation of cloud computing and machine intelligence.

The following parts explored remote sensing technologies, classifying platforms as ground-based, airborne, and satellite platforms, and providing a comprehensive analysis of the progress and constraints of remote sensing imagery. The significance of remote sensing image processing, encompassing pre-processing, enhancement, transformation, and classification, was examined to emphasise the crucial role of these procedures in enhancing the precision and applicability of remote sensing data.

The study focused on examining anomalies within datasets, specifically outliers and novelties, in order to emphasise the difficulties they present in the process of data analysis. An analysis was conducted on various categories of outliers, including point, contextual, and collective outliers. Additionally, the study highlighted the special challenges associated with anomalies in spatial data, such as masking and swamping effects.

The chapter also addressed the crucial problem of incomplete data in remote sensing, mostly attributing it to sensor malfunction and cloud cover. These issues highlight the intricacy of obtaining precise and comprehensive remote sensing data, requiring advanced methods for data recovery and correction.

The discourse on Big Earth Data Processing presented advanced technologies such as ArcGIS, ArcPy, and Google Earth Engine, which play a crucial role in handling and examining the extensive and intricate geographic datasets that are typical of contemporary Earth observation endeavours. These solutions illustrate the importance of scalable approaches and automated methods in addressing the challenges posed by data volume, quality, and processing.

Finally, this chapter underscores the significance of modern technical solutions and rigorous data analysis methods in overcoming the complexities of Earth observation data. This research significantly enhances the quality and utility of remote sensing data by improving the detection of outliers as well as reconstructing missing data. Enhanced accuracy and the absence of anomalies in Land Surface Temperature data augment the efficacy of urban heat island studies, vegetation monitoring, drought evaluation, and precision agricultural applications. Improved image quality and data integrity immediately facilitate more dependable environmental decision-making, paving the way for future investigations into determining climatic variables and monitoring the environment.

Chapter 3

Literature Review

3.1 Introduction

This chapter presents a comprehensive review of research relevant to the subject of outlier detection (OD) and reconstruction of lost big earth data. It carefully delineates the fundamental information that serves as the basis for the research described in the subsequent chapters. The review will provide a current and comprehensive summary of the latest advancements in approaches for OD and reconstruction of lost big earth data. It will specifically highlight the difficulties associated with handling extensive and intricate datasets from variable resources. This investigation will emphasise the unique approaches utilised in this field, identify current deficiencies, and clarify the underlying causes of these deficiencies in the existing methodologies. This analysis will also examine the appropriateness and possible constraints of these technologies in outlier detection and reconstruction of lost big earth data.

3.2 Anomalies in Datasets

Over the last few decades, big data has gotten the attention of industry, academia, government, and other organisations. During the analysis of real-world data, a common practice is to identify data or instances that are outliers or dissimilar to the actual data. These are known as anomalies [7]. Normally, anomalies give the idea of erroneous values in the dataset, but anomalies can introduce two factors in any dataset, given as [3]:

Outlier: Anomalies introduce outliers in the form of malicious data, which is inconsistent with respect to the rest of the data [7]. Figure **??** shows that the clusters C1 and C2 have most of the observations that are consistent, and the values of P1, P2, and P3 are data points that are located away from the regions of consistency, and these data points are considered outliers.

Novelties: Previously unrecognised patterns in data are considered novelties. A novelty score using a threshold is given to the data point. [28]

3.2.1 Type of Outliers

Outliers are data points that deviate significantly from the majority of the data. Point P1, P2, and P3 in figure 3.1 are examples of outliers.

Point Outliers

Points outliers are considered as data points which are far away from the rest of the data. Consider point P1, P2 or P3 in figure 2.3, these are example of point outliers.

Contextual Outliers

Contextual outlier are also known as conditional outlier [30]. For example a temperature peak of 70F would be considered as normal in summer but it will be treated as outlier in winter. The context for which the outliers are being identified must be specified beforehand to proceed with outlier detection. In particular data instances are divided by two set of attributes.

1. **Contextual attributes.**The contextual attribute specifies the context in which a data instance exists. For instance, with time series data, time functions as a contextual attribute.

2. **Behavioral attributes.** Behavioural attributes, often referred to as non-contextual attributes, define the non-contextual properties of a data point.

Collective Outliers

One or more data points gather and form a cluster that is anomalous as compared to the entire dataset, known as collective outliers. The individual data points in that cluster may not be outliers, but due to their occurrence, they are treated as collective outliers.

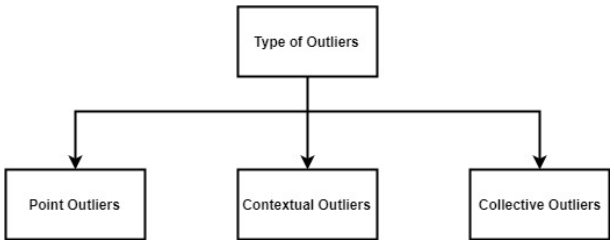

Figure 3.1: Classification of Outliers

3.3 Anomalies in Spatial Data

Remote sensing has been widely used for the past few decades for environmental, oceanic, and surveying applications. All the applications are entirely dependent on remote sensing data, how it is acquired, and how accurately we can make predictions with this data. Errors in spatial data are as important as actual data, measurement, and data structure. These errors mainly occur due to attribute definition, data sources, data modelling, and the analysis process. Spatial data mainly consists of systematic error, stochastic error, and gross errors, such as incomplete, inaccurate, redundant, and inconsistent data and anomalies due to multiple sources of data. Masking and swamping effects are common while detecting outliers. Masking states the occurrence of a second outlier after detection and removal of the first outlier, and swamping occurs when a correct data point is considered an outlier due to the actual outlier [31].

3.3.1 Spatial Outlier Detection and Removal

A spatial outlier is a spatial data point whose non-spatial attributes are not consistent with its neighbourhood. The detection of spatial outliers can reveal interesting information. Spatial outlier detection methods can be classified into six types, given as follows: [49]

- Statistical methods.
- Distance-based methods
- Density-based methods
- Depth-based methods
- Clustering
- Neural network
- Support vector machine

Statistical methods identify outliers by differentiating between actual data and ideal data. [50, 51] It follows the idea that the dataset under consideration follows a certain distribution. But as the model depends on a statistical model, it is difficult to build a model specifically suitable for the data, and if a data point does not satisfy the model criteria, it will be considered an outlier. [52] [53]..
In distance-based outlier detection methods, if a data point is separated from most of the data points, it will be an outlier. [54] [55] [56]. One simple approach is the K nearest neighbour algorithm. It assumes that an instance of a consistent data point can be found in the neighbourhood. [57]. Distance-based approaches are better in terms of understanding and implementation as compared to statistical approaches. A variant of the distance-based approach based on pruning claims that its complexity approaches linearity in practice. [58]

The density-based outlier follows the idea of degree of isolation. A simple point density is calculated, and if its value exceeds a certain threshold, the point is considered an outlier. [59] [52].

Depth-based outlier detection methods follow the concept of computation geometry. [60] These methods compute different k-d convex hulls. These algorithms lack extensibility; as the amount of data increases, their performance decreases significantly.

Cluster-based approaches form clusters of the data, and a data point that does not belong to any cluster is considered an outlier. Cluster-based approaches form clusters instead of detecting outliers, which is one of their limitations [61] [62].

The approaches discussed before do not consider the implicit relationship between data points. Neural networks utilise their learning abilities to establish this relationship between attributes of the dataset. [63].

Support vector machines have developed applications in various fields. It was initially used for binary classification, but it has been extended to apply to multi-class classification. A hyperplan is formulated, which separates the outlier from consistent data. [64] [63].

The six classifications of outlier detection methods presented before have their own pros and cons, which makes them suitable for one scenario but lowers performance in another [65].

Outlier detection is involved in every aspect of today's research, where data is considered an important point of decision-making. Data with spatiotemporal characteristics provides new aspects of studying human mobility patterns. Therefore, location-based services (LBS) are providing new ways to analyse the spatiotemporal construction of urban areas. This study [66] gives the idea of a trajectory data cleaning method for electric bicycles. This is a three-step procedure, as given below:

1. Endpoint classification from continuous and raw data from GPS

2. Remove incomplete trajectories

3. Extraction of Origin-Destination Pair

PAIRS (Physical Analytic Integrated Repository and Services)[67] is another algorithm of outlier detection for the purpose of data cleansing. It is a platform that enables high-resolution monitoring for agriculture as well as weather forecasting by utilising machine learning to enhance renewable energy forecasting. This platform enables the processing of petabytes, irrespective of traditional databases with limited processing abilities when data exceeds a limit. Another tool used for geospatial big data analytics is Spatial Hadoop and Apache Spark, which are built on top of Hadoop and the Apache Spark platform and provide big data analytics for spatial data. [68]. They provide a base for building spatial applications for the visualisation and processing of spatial data. Outlier removal is not limited to raw data alone. Remote sensing images are also pruned to outliers. This is an important aspect of spatial data cleaning, where outliers occur in the form of clouds, cloud shadows, and haze in remote sensing images. The optical characteristics of an image play an important role in satellite imagery for weather prediction. A fast algorithm is demonstrated [69] that efficiently removes the atmospheric effects on the image. Parallel implementation of the algorithms makes them perform much more efficiently. The image is processed in parts with a parallel implementation of the algorithm.

Different case studies are done in terms of outlier detection with the sole purpose of purifying data for better processing and application. A case study involving cryospheric regions of Nepal was performed by [70]. Aerosols play a vital role in climate prediction, research, and the amount of uncertainty they bring to atmospheric processes. The uncertainty due to aerosols and atmospheric processes is one of the huge gaps in current weather forecasting capabilities. An empirical model has been developed for AOD estimation over cryospheric regions of Nepal. The results are corrected using the average regression slope from the MODIS (Moderate Resolution Imaging Spectroradiometer). Similar to this, another case study of outlier detection for insect count data used a variogram model to describe

the spatial continuity of the data. These models are heavily affected by outliers. The model presented detects outliers in isolated and patchy spatial distributions of insect count data and is removed by neighbouring median filters. [71]

A geostatistical method involving outlier detection is used in the quality analysis of underground water. [72] studied two methods, including point kriging and IRF-k (intrinsic random function of order k). Both of these methods proved well in detecting outliers for the concentration of chloride and testing the hardness of water. Aerosol optical depth products are widely used in climate studies. [73] utilises the maximum likelihood method to fuse data from multiple AOD products. Outliers in this case are removed using a threshold in a 50 x 50 window to improve the quality of the fused data. Identification or detection of spatio-temporal outliers will help us understand the unexpected and interesting patterns in the data. [52] has presented an algorithm to identify outliers based on the degree of outlying. The algorithm improves performance by detecting true outliers by considering the spatial properties of the objects.

Spatial outlier detection in wireless sensor networks (WSNs) can be used to ensure the quality and accuracy of data before it is fed to the decision-making system. Traditional spatial local outlier measures (SLOM) and spatial local outlier factors (SLOF) were used to detect outliers. Citation 8009844 introduced a novel spatial local outlying value (SLOV) factor. The algorithm presented is able to detect outliers without any parameters, as compared to SLOM and SLOF, which require high user dependency.

Isolated and assembled outliers are present in multidimensional spatial series data. A self-organising map (SOM) neural network algorithm is presented for the spatial series dataset [63]. First, clustering is performed based on the SOM neural network, and then an outlier detection strategy is defined based on the topological distribution of neurons. The algorithm utilises a two-step procedure that first uses clustering to classify the input dataset and then uses a detection strategy to detect outliers. This SOM-based algorithm has proven to be efficient

for multidimensional spatial series data.

The iterative r and iterative z algorithms are based on clustering techniques. Both algorithms detect outliers and normalise the data. The K nearest neighbour algorithm is used in both of the algorithms. The neighbourhood function g and comparison function h are the essence behind these algorithms. The neighbourhood function g is evaluated at a spatial point x and is considered the average of all k nearest neighbours of x, while the comparison function h(x) is considered the ratio of f(x) and g(x), where f(x) is an attribute function of x. The value of h(x) decides, being very large or small, whether the given point x is an outlier or not[74].

Classical outlier detection considers global outliers. Local outlier detection is a complex process. The spatial local outlier measure (SLOM) algorithm is used to detect spatial local outliers [75]. SLOM utilises attribute partition and considers self co-relation and heterogeneity in spatial data. But the limitation with SLOM is that the computation of the fluctuation factor is limited, which sometimes causes degradation. The pros and cons of SLOF and SLOM algorithms led to the development of new algorithms using spatial outlying degree factor (SODF) [76]. Which utilises self co-rrelation from SLOM and spatial neighbourhood concepts from SLOF to reduce the computational complexity and produce better results.

3.3.2 Suitability of Outlier Detection Techniques for Spatial-Temporal Data

The effectiveness of outlier detection techniques for spatial-temporal datasets largely relies on their ability to adeptly manage both spatial correlations and temporal fluctuations. Conventional statistical methods, like Moran's I and Geary's C, proficiently quantify spatial autocorrelation by assessing the extent to which data points demonstrate similarity or dissimilarity in regard to spatial proximity [77]. These methods are beneficial for detecting local spatial clusters or isolated outliers by evaluating variations in spatial patterns. Nonetheless, their princi-

pal shortcoming resides in their inability to account for temporal dynamics, as they often evaluate data as static snapshots [78]. Consequently, when addressing time-series data or dynamic phenomena captured via satellite photography, their efficacy significantly declines.

Distance-based techniques, such as k-nearest neighbours (k-NN), provide clear and direct approaches for detecting geographical outliers by analysing the proximity of each data point to its neighbours. Spatially separated places, with distances that substantially above conventional neighbourhood thresholds, are designated as anomalies. Although these methods consistently identify discrete spatial abnormalities, they naturally neglect the temporal continuity of spatial events. As a result, individuals may incorrectly interpret transient variations in environmental conditions or typical temporal oscillations as anomalies [79, 78]. This limitation renders them less appropriate for ongoing, long-term remote sensing data analysis, where temporal correlations critically affect the accuracy of anomaly detection.

Density-based clustering methods, particularly DBSCAN (Density-Based Spatial Clustering of Applications with Noise), proficiently locate outliers by recognizing clusters in high-density areas and designating isolated data points as anomalies [80]. These approaches offer resilience to noise and irregular cluster configurations, especially beneficial for heterogeneous geographical information. Density-based approaches exhibit significant sensitivity to parameter selection, including neighbourhood radius (epsilon) and the minimal number of points necessary for cluster formation. In dynamic spatial-temporal situations, these parameters may fluctuate over time, necessitating meticulous and ongoing calibration . This sensitivity presents considerable practical difficulties, especially when utilized with large remote sensing datasets marked by intricate and changing environmental circumstances [81].

Recent advancements in machine learning and deep learning techniques, such as neural networks, isolated forests, convolutional neural networks (CNN), and autoencoders, have exhibited formidable proficiency in concurrently modeling in-

tricate spatial-temporal connections [82]. Neural networks, particularly recurrent neural networks (RNN) and Long Short-Term Memory (LSTM) networks, are proficient in recognizing temporal patterns and forecasting anticipated behaviors, hence effectively detecting anomalies [83]. Isolation forests offer computational efficiency and scalability for anomaly identification, making them appropriate for extensive datasets. Deep learning architectures such as CNNs and autoencoders proficiently utilize spatial and temporal connections by obtaining hierarchical and latent representations from the input. These sophisticated models can autonomously adjust to diverse data patterns without costly manual parameter adjustments, markedly improving anomaly detection precision and resilience. However, its intricacy frequently necessitates considerable processing resources and vast labeled training datasets, presenting practical constraints in real-world remote sensing applications [84].

Considering these strengths and limitations, methodologies that incorporate spatial-temporal dimensions exhibit enhanced efficacy in addressing the complexities associated with remote sensing datasets, including intricate environmental interconnections and dynamic temporal patterns. Although statistical, distance-based, and density-based methods provide significant insights, their intrinsic constraints in temporal modeling underscore the benefits of utilizing modern machine learning and deep learning techniques. Thus, the selection of outlier detection methods for spatial-temporal data must emphasize techniques that can effectively identify intricate interactions across space and time, while balancing computational efficiency with detection robustness to meet the practical requirements of modern remote sensing analyses.

3.4 Quantitative Remote Sensing

Satellite remote sensing (SRS) is vital for weather predictions and the examination of the impacts of global warming. It provides key data for climate change research and Numerical Weather Prediction (NWP) models [85] [86] [87]. Satellites

such as INSAT, IRS, and others provide significant data on many environmental factors like sea surface temperature (SST), cloud motion vectors, and vegetation development. This data helps in monitoring changes connected to climate. The progress in satellite technology, including the use of distributed remote sensing satellite systems and image fusion techniques, enables more effective and precise gathering of remote sensing data. This, in turn, enhances weather analysis and assessments of climate change. The correlation between ground surface temperature (GST) and satellite-based land surface temperature (LST) has been investigated, demonstrating the possibility of connecting these variables to enhance comprehension and utilisation in many disciplines [3]. The progress in algorithms for estimating land surface temperature (LST) using thermal infrared (TIR) has resulted in the creation of LST products of excellent quality. These products are valuable for studying surface evapotranspiration, estimating soil moisture, and investigating climatic change [88]. Furthermore, the comprehensive analysis of TIR LST satellite data applications emphasises the growing significance of satellites such as MODIS in the surveillance of LST for various research objectives [89]. Satellite remote sensing offers unique insights into weather patterns and the dynamics of land surface temperature, thereby improving our capacity to predict and comprehend environmental processes.

Satellite remote sensing (SRS) is an important aspect of the study of weather predictions and the effects of global warming[90]. Remote sensing data provides land surface properties and variables, which are important for understanding the earth system. Quantitative remote sensing involves the extraction of many different parameters[91].

3.4.1 Errors in Quantitative Remote Sensing

Quantitative remote sensing errors originate from multiple factors, such as topography influences, atmospheric circumstances, sensor constraints, and data processing methodologies. The topography has a considerable effect on the re-

flectance that sensors monitor. To account for this, digital elevation models (DEMs) are used to compensate for the effects of topography. Nevertheless, global digital elevation models (DEMs) frequently underestimate the cosine of the sun angle and inadequately depict shadows, resulting in mistakes in areas with mountains [92, 93]. Furthermore, the process of obtaining information about aerosol and surface properties is filled with uncertainties caused by both random and systematic errors. These errors are assessed in real-time using algorithms such as GRASP [94]. The presence of errors can also be attributed to the representativeness error produced by scale transformation (REST), which is inherent in data assimilation processes that involve many sources and scales of data [95]. The measurement of precipitation events is made more complex by temporal and geographical sample mistakes. These errors are more pronounced on land, particularly in hilly areas, compared to oceans [93]. The geolocation calibration of CubeSats introduces positional inaccuracies, which in turn cause further interpolation mistakes. These flaws may be accurately modelled analytically in order to assess their impact on remote sensing products [96]. Furthermore, the transmission of inaccuracies in data via analytical processes can result in substantial uncertainty in measurements of vegetation productivity. This highlights the need for reliable anomaly detection techniques to identify and rectify outliers before the data is widely utilised [97]. In order to enhance the precision and dependability of remote sensing data, it is essential to comprehend and address these many sources of mistake.

Satellite remote sensing is the only suitable option to obtain land surface properties effectively [98]. Currently, MODIS is considered the most trusted satellite for providing data for land surface variables by supporting 44 different products [99]. The information retrieved or observed from satellites is subject to errors caused by sensor malfunctions or efficient retrieval algorithms. Even the accuracy of the MODIS is different at the global level as compared to the local or regional level [100].

3.4.2 Remote Sensing Information Reconstruction

Reconstructing remote sensing information is a crucial procedure for overcoming the difficulties caused by incomplete or degraded satellite data resulting from variables such as noise, dead pixels, cloud cover, and low resolution. Several sophisticated methods have been devised to address these problems. At present, based on the classification proposed by [36, 9, 101] the algorithms used for missing information reconstruction are mainly classified into four categories:

- Spatial-based methods
- Spectral-based methods
- Temporal-based methods
- Hybrid methods

These methods are discussed in detail in the upcoming sections.

Spatial Based Methods

Early approaches primarily relied on spatial reconstruction methods, leveraging the spatial continuity and autocorrelation of remote sensing data. Image in-painting is a fundamental technique in spatial-based methods, representing the traditional methods used for image restoration in remote sensing and computer vision. Image in-painting is based on the idea that areas with missing data share similarities in geographical aspects with their surrounding environment, using this likeness to fill in the gaps in information [102]. Spatial methods do not require additional imagery to assist in the reconstruction process, unlike other techniques [103]. They carefully examine how local and global data in the image interact, using this connection to smoothly incorporate missing content. Spatial approaches are more effective in revitalising places with limited data gaps. These models are limited by their dependence on a small amount of reference material, making them less capable of reconstructing big or complex landscapes.

Spatial approaches are effective for filling tiny gaps but less reliable for recreating complicated terrains due to the uncertainty in predicting geographical features. A generative adversarial network (GAN) has been used recently for image inpainting. Due to its instability in training and producing results that are far from reality, a deep convolutional generative adversarial network (DCGAN) [104] was used to perform image inpainting for sea surface occlusion due to clouds. A two-stage procedure that first involves training DCGAN to generate close-to-real SST images. Second, the encoding of the corrupted image is formed using the inpainting loss function, and this encoding is passed to DCGAN to produce the missing content [105]. Although GAN-based inpainting improves detail realism, the training instability and high computational overhead pose serious barriers for consistent deployment in operational EO pipelines. Additionally, their dependence on large training datasets and parameter sensitivity reduces generalizability across scenes with different land cover types.

Spatial methods have also been termed self-complementation methods [106]. Spatial methods can be further categorised into four types:

- Interpolation methods
- Propagated diffusion methods
- Variation-based methods
- Exemplar-based methods
- Learning-based methods

Most image-painting algorithms are used in digital image processing and can be used with remote sensing images as well. [107, 108, 109, 110] Interpolation methods are the most basic algorithms in spatial-based methods. Interpolation methods share a general rule of weighted averages of sample values, as given in equation 3.1.

$$\hat{I}(x_0) = \sum_{i=1}^{N} w_i I(x_i) \tag{3.1}$$

P_x0 is the point of interest, and $\hat{I}(x_0)$ is the estimated value of the point of interest. $I(x_i)$ is the observed value, w_i is the weight of the sample values $P_x i$, and N is the number of sampled points used for interpolation. [36]. Typical examples of interpolation methods are [111] and [112]. Both methods utilise the Kriging algorithm to interpolate the missing pixels. Interpolation methods are not suitable for complex ground features because the spatial information of the remote sensing image is not fully utilised to fill the gap. [113, 107, 114, 115, 116] are examples of image inpainting algorithms based on interpolation. The basic idea of [117] is to input the missing area to the algorithm, and it fills the missing information starting from the edges and moving inwards, but with [113], the area that needs to be filled is selected by the user, and the algorithm then determines which area in the image to be used to fill the gap.

The scan-line corrector-off problem is common in Landsat imagery, which causes information gaps caused by missing pixels due to sensor failure. Missing pixels make up 22% of the data [118]. Neighbourhood similar interpolator (NSPI) was developed by [119] to fill the information gaps due to missing pixels in Landsat ETM+ SLC imagery. It is a three step procedure. The first difference of similar pixels is determined between the target and input image using the radiometric difference. Second, NSPI selects a certain sample size, which makes the estimation of pixels statistically reliable. Third, a spectral similarity approach is applied to identify pixels that belong to similar land cover features. The chosen pixel replaces the target pixel. Similar to this functional concurrent linear model (FCLM), it is proposed to fill the SLC-off Landsat 7 information gaps [120].

Propagated diffusion methods are another type of spatial-based method that follows the idea of image inpainting. Diffusion methods start filling the edges with information and propagate inward until the gap is filled. This is performed using partial differential equations; therefore, these methods are also known as

partial differential equation (PDE) methods, for example, [121, 122]. Diffusion-based methods use isotropic and anisotropic partial differential equation models. [123] and [124] are examples of isotropic and anisotropic diffusion-based models for image inpainting. Diffusion models are not suitable when the missing area is large, as they end up causing blurring.

Variational-based image inpainting algorithms focus on maintaining the geometrical integrity of the image by filling in small, missing areas in the image. It considers the image to be composed of objects and shapes, with sharp edges and objects being considered smooth on their own. Due to this limitation, variational-based methods are only suitable for small inpainting or retouching problems [125]. Image regularisation is used to formulate variational-based inpainting problems. There are four types of image regularisation algorithms mainly in use, named as follows:

- ℓ^2 Norm Regularisation
- Variation (TV) Regularisation
- ℓ^1 - ell^2 Norm Regularisation
- Nonlocal Regularisation

Laplacian regularisation is representative of ℓ^2 norm regularization. Its goal is to apply smoothness to the image by minimising the high-frequency components of the image. Therefore, Laplacian models are suitable for flat surfaces or low-resolution images because their performance degrades with detailed areas of information reconstruction and results in blurring [126]. One of the other regularisation techniques is Gauss-Markov regularisation [127].

Variation TV regularisation is suitable where sharp edges are to be recovered. However, if the missing area is larger than the object, its performance degrades. TV regularisation is used to smooth the remote sensing image by using tensor ring completion. For example, [128] used total variation tensor ring completion for missing information reconstruction in remote sensing images.

ℓ^2 regularisation was only suitable for images with flat regions, but as an image is a combination of flat and detailed regions, ℓ^1 and ℓ^2 are combined to form ℓ^1 - $ell^2\ell^2$ norm regularization. They update the cost function in the manner given in 2. [129]

$$Costfunction = loss(say, binarycrossentropy) + regularisationterm \tag{3.2}$$

ℓ^1 and ℓ^2 differ in regularisation terms. The ℓ^1 regularisation term is given as

$$CostFunction = Loss + \frac{\lambda}{2m} \times \sum ||w|| \tag{3.3}$$

and the ℓ^2 regularisation term is given as

$$CostFunction = Loss + \frac{\lambda}{2m} \times \sum ||w||^2 \tag{3.4}$$

All the regularisation methods discussed before perform local regularisation by using local information to reconstruct the image. Nonlocal regularisation is another technique that uses information about the complete image and tries to fill up the missing area. [130, 131, 132, 133] are some nonlocal image regularisation techniques. But nonlocal regularisation is still subject to blurring when the missing area is large, complex terrain.

Exemplar-based methods follow a greedy strategy to recover the missing area. These methods fill the missing pixel by copying another pixel that is most similar to it, so they use a pixel-by-pixel approach to fill the missing area. Exemplary methods mainly preserve the texture information of the digital image. Criminisi proposed an exemplar-based image inpainting algorithm that can remove objects from the image and recover the texture of the actual image behind it, as shown in the figure 3.2. [134]

More relevantly, several other methods have been studied and proposed following the idea of image inpainting. Some of them are cokriging interpolation

a b

Figure 3.2: Exemplar-based Region Filling and Object Removal

[135], structure-preserving global optimisation [136] compressed sensing [137] and sparse dictionary learning [138].

While spatial methods are intuitive and require minimal auxiliary data, they are inherently limited in scenarios with extensive or irregular missing regions. Their reliance on local neighborhood information makes them unsuitable for reconstructing structurally complex scenes or capturing large-scale spatial variations. Moreover, methods like interpolation and variation-based regularization suffer from blurring and oversmoothing when applied to highly textured regions or large gaps.

Spectral Based Methods

Intermediate advancements in remote sensing technology introduced spectral reconstruction methods, exploiting spectral band correlations for improved reconstruction accuracy. The information from one band can be used to reconstruct the information that is missing, which allows spatial-based algorithms to circumvent the problem of a lack of prior information. When hyper-spectral or multi-spectral images are lacking information, they both feature bands that have complete information as well as bands that contain missing information. Therefore, the goal is to use the bands that contain complete information to reconstruct missing information by building a correlation between the bands. Different names for spectrum methods include multi-spectral complementation methods and spectral

approaches. [106].

Due to a sensor fault in band 6, for instance, Aqua MODIS's imagery is characterised by various patterns of black strips that appear repeatedly. There was an initial suggestion made by [32] regarding the solution to this problem. With coefficients of 0.9821 and 0.9777, Aqua MODIS bands 6 and 7 are associated with one another, according to the fundamental concept that underpinned this. It is therefore possible to retrieve information that is lost from band 6 by using band 7, which is highly linked. Through the utilisation of quadratic polynomials, the following polynomials were discovered by [32]:

$$K^6(r) = 1.6032(K^7(r))^3 - 1.9458(K^7(r))^2 + 1.7948K^7(r) + 0.012396 \quad (3.5)$$

or

$$K^6(r) = -0.70472(K^7(x))^2 + 1.5369K^7(x) + 0.025409 \quad (3.6)$$

where $K^6(r)$ and $K^7(r)$ are top of atmosphere (TOA) coefficients. But the coefficient of correlation for the bands is subject to the terrain and subject to change for different scenes. These coefficients presented by [32] are for snow-covered terrain and work best in this particular scenario. Following the difference in coefficient between scene types,

[139] developed a class local fitting (WCLF) algorithm that first classifies the image based on scene type, then recovers missing pixels based on scene type, and after the recovery, outliers are removed by a smoothing procedure. This method is strongly dependent on the classification procedure, specifically for pixels at the joints of different scene types.

[140] recovered the missing band 6 information by using a two-step procedure: first, histogram matching is used to rectify the information errors in working detectors, and second, local least squares fitting is used to fill in the missing information from faulty detectors. Although this method introduces distortion

in the image, unlike the method proposed by [32], it is not dependent on scene type to process the missing information.

There are seven spectral bands, and the methods discussed so far only discuss the utilisation of band 7 information to reconstruct missing information in band 6. [141] proposed a way to utilise the information of other bands as well as reconstruct the information. Based on the correlation between working detection of different bands, [142] presented the M-estimator multiregression method to recover the missing information in band 6.

A multi-scale segmentation approach was used by [143] to fill the gaps.

Spectral methods are better at information recovery as compared to spatial methods, and the results are closer to the actual scenario. Spectral methods can remove thin clouds and haze efficiently, but if clouds are thick, which causes all the bands to be contaminated, spectral methods lose accuracy because information in every band is affected to some extent [106]. Spectral methods offer the advantage of internal redundancy by utilizing band correlation, particularly in multispectral or hyperspectral images. However, their effectiveness deteriorates when all spectral bands are affected simultaneously—such as under dense cloud cover or sensor-wide anomalies. Moreover, many methods (e.g., polynomial regression or spectral fitting) are scene-specific and require frequent recalibration, limiting scalability. Some methods, like band 6 reconstruction using band 7, perform well under specific terrain types (e.g., snow cover) but generalize poorly to heterogeneous scenes. Scene dependence significantly limits their utility in global-scale environmental applications.

Temporal Based Methods

Temporal reconstruction methods became prominent with continuous accumulation of satellite data, enabling the effective reconstruction of missing data by exploiting historical patterns and temporal dynamics The dense cloud cover causes all of the spectral bands to be contaminated and to have missing information

in them. Additionally, the malfunctioning sensors can result in missing information in all of the bands for a certain spectral band. Because of this, spectral approaches are rendered ineffective because they are founded on spectral correlation, which is lost once all of the bands have information that is missing. The temporal-based approaches come into play at this point in the process. Due to the fact that the clouds are constantly moving, it is possible to collect data for the same location at a different time interval. The determination of the time interval is a challenging aspect of this method. If the time interval is large, it will be influenced by changes in the land cover, but if the land cover is minimal, it will have clouds that overlap in two different time slots. Significant efforts have been made by researchers in the field of temporal-based approaches. For instance, [144] described a method that utilised local linear histogram matching (LLHM). This method necessitated the utilisation of high-quality data in order to operate properly, but ultimately yielded unsatisfactory outcomes when applied to heterogeneous landscapes. Algorithms were proposed by [145, 146, 147] in order to enhance the radiometric consistency of multi-temporal pictures for heterogeneous land. Temporal approaches leverage historical observations to restore missing values, making them powerful in dynamic environments. However, their reliability hinges on temporal consistency. In rapidly evolving landscapes or those with frequent land use change, these models often introduce bias or fail to capture abrupt shifts.

Temporal-based models are also known as auxiliary-sensor-complementation-based approaches. Temporal methods are divided into the following categories:

- Temporal replacement methods
- Temporal filter methods
- Temporal learning models

Temporal replacement methods are also referred to as mosaicing by [148]. The basic idea behind temporal-based methods is to simply replace the missing area with

another complete image at some other time. [149, 150, 151] are some temporal replacement methods. The idea of temporal replacement leads to a multi-temporal cloud removal strategy. Multi-temporal cloud removal (MCR) is used for radiometric correction before using clear reference images to reconstruct the missing values in the image. The method not only removes clouds but also gets rid of cloud shadows to produce a clear image [152]. Temporal replacement can be performed in two different ways: direct replacement and indirect replacement. Direct replacement selects the replacement patch by using the optimal value in the time series, while indirect replacement first reduces the temporal difference and then replaces the missing information following the direct replacement procedure [153]. Many cloud removal approaches use a reference or auxiliary image to reconstruct the missing information gap. [106] proposed a different approach to produce continuous cloud-free images based on spatial temporal weighted regression (STWR). The spatial-temporal weighted regression model uses pixel information from both the self-target and reference images.

Multi-temporal cloud removal and information reconstruction techniques also follow the idea of temporal replacement or patch-based information reconstruction. [154] showed an information cloning technique to remove clouds and recover lost data. The method considers the fact that land cover change is insignificant over a short period of time. The cloning algorithms reconstruct the information gap using cloud-free regions from different multi-temporal images. Global optimisation is done after cloning to ensure radiometric consistency. The problem of radiometric inconsistency has been critical for information reconstruction algorithms following multi-temporal or patch-based approaches. [155] attained consistent brightness by using the mean, standard deviation, and linear transform of brightness values for reference and the target image. Considering the radiometric difference between the reference and target image, [101] used the digital number value of the target image and reference image and performed linear regression modelling (LRM) to propose a multi-temporal cloud removal algorithm. The

result shows good spectral compatibility and radiometric consistency between reference and target images. Similar to this, [9] also focused on the radiometric consistency between the target and reference image and proposed a seam determination technique to determine good boundary conditions and a clustering algorithm to cluster the contaminated image patch with similar cloud-free patches to clone the information by maintaining consistent temporal intensity. The performance of this method, just like other multi-temporal-based methods, degrades as land cover changes significantly over a period of time.

Time series data is subject to fluctuations and thus contains noise. Temporal filter methods are used to remove the noise in one-dimensional time series data. [156] is a typical example of a temporal filter-based method. Temporal learning models focus on establishing a relationship between erroneous data and consistent data in the temporal domain. [157, 158] are examples of temporal learning-based models. Dictionary learning is another aspect of image restoration. Dictionary learning is used in sparse reconstruction techniques for signal restoration. These methods use a small set of data to form a dictionary and try to search for the dataset that best recovers the signal. Dictionary learning has a number of applications, including the extraction and classification of image features ([159]), image denoising [160] and face recognition [161].

The satellite image time series comprises a set of images obtained from different satellites or sensors over a similar area at different time slots. Clouds are a major source of obscuration in these images. As discussed earlier, spatial, temporal, and hybrid methods have been widely used to remove the clouds and reconstruct the underlying image scene. Dictionary learning has also established its roots in the cloud obscuration problem in terms of sparse reconstruction. A two-step procedure based on sparse reconstruction starts with masking the clouds and their shadows first and then forms a dictionary of cloud-free pixels from the time series image. The dictionary is used in the reconstruction of the cloud-contaminated pixels [162]. In the brief overview presented in [36] sparse reconstruction tech-

niques perform better as compared to spatial or temporal-based techniques. [163] established a variant of dictionary learning for recovering missing information in remote sensing images and presented two multi-temporal dictionary learning algorithms utilising KSVD (K-Singular Value Decomposition) and Bayesian dictionary learning. KSVD uses temporal transformation to establish the temporal correlation, while the Bayesian framework adaptively forms the temporal correlation based on weights. [164] is an improved version of this method in which dictionary learning group (DGL) is formed by low-resolution (LR) and synthetic aperture radar images (SAR), which are used as auxiliary images. The non-local similarity between cloud-contaminated patches in HR, LR, and SAR images is determined, which plays a key role in reconstructing the cloud-contaminated patches in HR images. Multi-temporal dictionary learning (MDL) makes two dictionaries of the target image and the reference image. The removal of clouds takes place by combining the coefficients of reconstruction obtained from two dictionaries. Results show that MDL gives better performance as compared to MNSPI [165].

[138, 153] are some of the typical sparse reconstruction techniques. Similarly, interpolation, matrix completion, and robust matrix completion have been used as reconstruction models for remote sensing data recovery. [166] used a variation of robust matrix completion with a two-step procedure that first detects the cloud, generates a mask to remove the cloud, and then reconstructs the underlying image using the Augmented Lagrange method. It uses a sequence of non-cloudy images for the reconstruction of lost data.

Techniques like temporal replacement are simple and computationally efficient, but they often result in radiometric inconsistencies. While filtering models reduce temporal noise, they may suppress relevant temporal signals, leading to the loss of significant changes in EO data. Advanced methods, such as multi-temporal dictionary learning and patch-based reconstruction, demonstrate improved accuracy, yet they require large, well-aligned datasets and intensive preprocessing. Their reliance on temporal continuity limits their effectiveness in regions with

sparse observations or frequent occlusions.

Spatial-Temporal-Spectral Based Methods

Contemporary hybrid approaches have emerged to overcome limitations inherent to individual spatial, spectral, or temporal techniques. These advanced methods integrate multiple dimensions, combining the advantages of each to improve accuracy and robustness The performance of temporal methods is superior to that of all other approaches; nevertheless, temporal methods are also inferior because of the restrictions placed on the amount of land cover change. Exploring temporal techniques for filling in missing data shows notable progress in managing time-series data, especially in scenarios with intricate variables like land subsidence, which are influenced by several circumstances. Utilising a multi-factorial approach with principal component analysis (PCA) highlights the need of pinpointing and prioritising the most impactful elements to enhance reconstruction precision [167]. This method excels in simplifying intricate, multidimensional data into primary components that encapsulate the highest variation, therefore improving the predictability and comprehensibility of the reconstruction process
.

The systematic theory that utilises spatiotemporal memory using artificial neural networks is a significant advancement in temporal data reconstruction methods. By guaranteeing Lipschitz continuity [168], this method ensures the stability of the reconstruction process and offers a more dependable and resilient framework for managing data with temporal dependencies. Ensuring the integrity and continuity of data points is especially important in time-series data for precise analysis and forecasting.

Temporal polynomial interpolation methods [169], particularly quadratic interpolation, provide a straightforward yet efficient option for filling in missing data. Their superior performance compared to previous polynomial methods in terms of accuracy demonstrates the promise of mathematical interpolation techniques in

situations with randomly missing data points. This method is advantageous due to its simplicity and the few assumptions it necessitates regarding the underlying data distribution, rendering it extensively adaptable across diverse industries.

Introducing attention-based architectures [170] for spatial and temporal information reconstruction is a major advancement in addressing missing data in spatiotemporal graphs. By utilising a spatiotemporal propagation architecture, these techniques may efficiently capture spatial and temporal relationships, resulting in more precise imputation outcomes. This method is especially important in the age of big data, because datasets are not just extensive but also intricate, encompassing complex connections between spatial and temporal aspects.

Despite progress, there is an increasing demand for hybrid methods that may integrate the advantages of many approaches to overcome the inherent limits of individual methods. Hybrid approaches can combine the detailed factor analysis of PCA, the predictive capabilities of neural networks, the simplicity of polynomial interpolation, and the nuanced knowledge of spatial-temporal correlations provided by attention-based architectures. Integrated techniques could provide more flexibility, adaptability, and accuracy when reconstructing missing data in various applications and datasets.

The spatiotemporal fusion-based techniques utilises data fusion from many sources to address this problem. The study by [171] utilised two reference photographs that were taken close in time to the target hazy image, instead of only one. This idea prevents errors resulting from temporal methods induced by substantial land cover changes. This method employs a residual correction procedure to enhance spectral similarity between the restored area and the unaffected cloud-free zone.

The spatial, temporal, and spectral (STS) techniques outlined earlier each have unique strengths and weaknesses. STS approaches are developed by combining various strategies to recover lost data more effectively and accurately. The STS model, created by [35], is a comprehensive model that employs a deep convolutional neural network. The model tackles deadline issues in MODIS Band 6 and

fixes the corrector-off problem in Landsat Enhanced Thematic Mapper Imaging. It can remove thick clouds and shadows by using multiple data sources. This method links incomplete data with complete data by using auxiliary data in a deep Convolutional Neural Network (CNN). The model uses a residual output to analyse the relationship between different auxiliary data. These techniques are also known as hybrid methods [36]. Hybrid methods combining spatial, temporal, and spectral information address the weaknesses of individual domains and provide state-of-the-art results in reconstruction tasks. However, their complexity is non-trivial—requiring substantial computing resources, advanced model design, and large auxiliary datasets. Deep learning-based hybrid models, like CNNs and attention-based architectures, offer strong performance but often at the cost of interpretability and reproducibility. Many are trained on custom datasets and lack generalization across different environmental conditions.

[172] presented an adaptive weighted tensor completion method for missing data reconstruction in remote sensing using data from spatial, temporal, and spectral dimensions. The idea behind this is to compute the weights by considering the information from the spatial, spectral, and temporal domains. Compared to using threshold weights for tensor completion [173], the adaptive method has better performance and accuracy for recovering missing information. Similarly, [153] also presented a spatial-temporal method using group sparse information that utilises correlation between local and nonlocal regions by extending single patch matching to multi-patch matching. This method utilises spatial as well as temporal correlations to minimise the difference between target and auxiliary images before searching the matching patch. [36] categorised hybrid methods into two categories:

- Spatio-temporal methods
- Spectral-temporal methods

Spatio-temporal methods utilise the best of spatial and temporal methods for the completion of missing information. [174, 112, 33] are typical examples of

spatio-temporal methods. Similarly, as spectral and temporal correlations cannot give effective results on their own, joint spectral-temporal methods are used. [175] is a typical example of the joint spectral-temporal method. Following the spatial-temporal characteristics, a four-step procedure for predicting the missing values was proposed [176]. The first algorithm selects the neighbourhood pixel, considering spatial-temporal characteristics, in the form of a submit. From these subsets, the ranking of the sub-images is done based on the similarity score. From the subset of images, an estimate for the quantile is completed, followed by quantile regression that regresses all the predicted values on the associate image ranks. [174] presented an improved Markov Random Field technique to remove clouds and reconstruct the missing information by constructing an optimal offset map and selecting the most appropriate pixel from the reference map. STMRF (Spatial-Temporal Markov Random Field) models a complex global relationship using the local neighbourhood pixel approach. The STMRF optimal offset map is generated using the following equation:.

$$M(K) = \sum_{p \in \omega} M_d(K(p)) + \alpha \sum_{p \in \omega} M_t(K(p), K(p')) + \beta \sum_{(p,p") \in N} M_s(K(p), K(p")) \tag{3.7}$$

Here, ω is the missing region. P (x,y) represents a pixel in the target image, which is cloud-contaminated in this case. P' is a pixel of the reference image in a similar position as p, and N represents a spatially connected neighbourhood system. P" is a 4-connected neighbour of pixel p.

The composite image method can be used for temporal high-resolution images such as AVHRR (Advanced Very High Resolution Radiometer) and MODIS (Moderate Resolution Imaging Spectroradiometer) to fill the missing information gap. But due to noise in the composite image, the normalised difference vegetation index (NDVI) is used to remove the noise and obtain relatively better-quality composite images [177]. This way of obtaining composite images has

been utilised to get 8-d or 16-d NDVI products with high-resolution images. But we need multiple observations with clouds or any other form of noise and one clear image of that particular scene to get a noise-free composite image, which is difficult in rainy or cloudy regions. Which causes a certain level of noise to be present in the 18-d and 16-d composite images. To remove this noise, previously available mathematical models such as Fourier and Gaussian models can be used. The missing values can be obtained by estimation using predicted values of the model, which results in a cloud-free image [178, 179].

Spectral-temporal methods follow the idea of the mosaicing effect to reconstruct the missing information due to cloud or cloud shadows. The mosaicing method follows the idea that a similar region may be captured at some other time that will not be contaminated with clouds. But due to temporal differences, radiometric differences may occur, which can introduce salt and pepper noise into the resulting image. Therefore, radiometric normalisation may be used to avoid the noise by removing the radiometric difference [180, 181]. To better overcome the noisy effect, a spectral-temporal patch-based technique is utilised to reconstruct the missing data in the time series image and avoid the noisy effects as well by preserving the textual information from a clear spectral-temporal patch from adjacent images. [182] followed a three-step procedure to obtain the cloud- and noise-free result. First, the pixels are separated based on their spectral and temporal properties to create a spectral-temporal patch. Next, the reference spectral-temporal patch is used to reconstruct missing information. Finally, contextual information is taken from the reference temporal image and added to the missing area to create a cloudy and noise-free image.

Multi-spectral and multi-temporal remote sensing data should be used together to reconstruct the missing information. One of the examples is the tempo-spectral angle model. It measures the similarity of pixels using temporal and spectral dimensions. Missing pixels are then replaced by temporal-spectral angle mapping

(TSAM). The The temporal and spectral angles of mapping are given as follows:

$$TSAM = \frac{1}{2H} \times \sum_{j=1}^{H} \times \arccos \left\{ \frac{\sum_{i=1}^{M} r_i, j s_i, j}{\sqrt{\sum_{i=1}^{M} r_{i,j}^2} \sqrt{\sum_{i=1}^{M} s_{i,j}^2}} \right\} + \frac{1}{2M} \times \sum_{i=1}^{M} \times \arccos \left\{ \frac{\sum_{j=1}^{H} r_{i,j} s_{i,j}}{\sqrt{\sum_{j=1}^{H} r_{i,j}^2} \sqrt{\sum_{j=1}^{H} s_{i,j}^2}} \right\} \tag{3.8}$$

where $r_{i,j} \in \mathbf{R}_{MXH}$ and $s_{i,j} \in \mathbf{S}_{MXH}$ and $\mathbf{R}$ and $\mathbf{S}$ are two pixels which are defined using spectral and temporal perspective, and both of these pixels have H temporal images and M spectral bands. The flow chart of TSAM is shown in figure 3.3 [183].

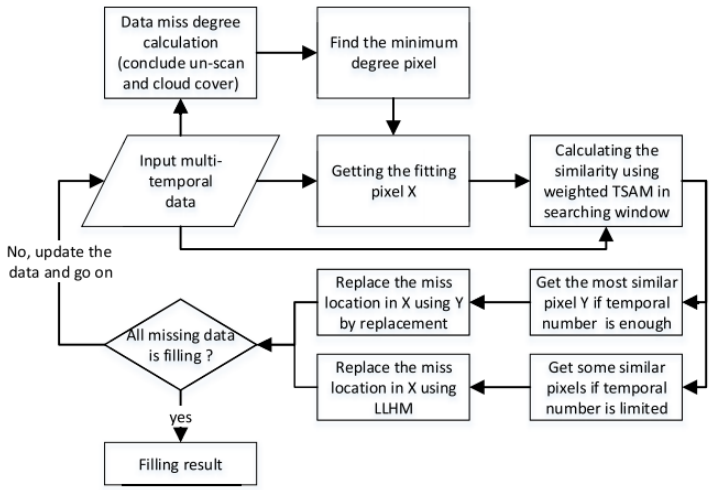

Figure 3.3: Flowchart of multi-temporal similar replacement-based TSAM

The neighbourhood-similar-pixel approach has been previously used to solve the Scan Line Corrector (SLC)-off problem in remote sensing imagery [119]. The similar NSPI approach has been modified to develop a hybrid thick cloud removal algorithm. The MNSPI uses one cloud-free image and one cloudy image, tries to find the similarity of cloud-free pixels first, and computes the weight of each similar pixel. After that, it predicts spectro-temporal and spectro-spatial information. Based on pixel distance from the cloud centre, it uses spectro-temporal information if the pixel is near the centre of the cloud and spectro-spatial information if the pixel to be replaced is near the edges and far from the centre of the

cloud [184].

Utilising spatial and temporal information at the same time to recover lost data has been popular among researchers. An extensive amount of research has been done in order to remove clouds and reconstruct the missing information in remote sensing imagery. Mostly, the images are of low resolution. A similar problem is also persistent in high-resolution images. [103] used adjacent temporal images as complementary information to create a mosaicing effect for removing clouds and recovering lost data for high-resolution images. A cloud mask is generated, and its boundaries are improved for effective recovery of missing data. Finally, missing information is reconstructed starting from the outer edge of the cloud and gradually moving to the centre point of the cloudy pixel. The radiometric difference is critical in high-resolution imagery, so After completely filling the information gap, a residual correction based on global optimisation is performed to reduce the radiometric difference between the recovered and cloud-free regions. There are two conditions for the reconstruction of LST data [185].

- Clear Sky
- Cloudy Sky

Reconstructing missing LST data for clear sky conditions may lead to an overestimation relative to the reconstructed data in cloudy conditions [186]. Despite limitations, there remains a significant research gap to generate high-quality reconstructed land surface temperature data. The author suggested a resilient gap filling technique by combining MODIS and VIIRS LST data in [187]. The method suggests utilising a selection of photos to fill in missing data in a target image while the sky is clear.

Hybrid or combined spatial-temporal-spectral methods represent the most recent and advanced class of reconstruction techniques. These methods integrate the strengths of individual spatial, temporal, and spectral approaches to achieve comprehensive data recovery and superior accuracy. Purely spatial methods, such

as interpolation and kriging, rely extensively on spatial correlations and continuity, providing excellent results in uniform areas but struggling in heterogeneous or dynamic environments. Similarly, purely spectral methods leverage spectral band correlations effectively but are susceptible to contamination when correlated bands are simultaneously affected. Temporal methods, including time-series forecasting and filtering approaches, efficiently handle predictable temporal changes but often neglect detailed spatial variability.

Recognizing these limitations, combined methods explicitly incorporate multiple dimensions to overcome the deficiencies inherent to individual approaches. Spatial-temporal-spectral methods often utilize advanced machine learning and deep learning models, such as multidimensional convolutional neural networks (CNNs), adaptive weighted tensor factorization, and transformer-based deep learning models. For example, CNN-based methods efficiently capture spatial dependencies while integrating spectral and temporal information to improve reconstruction performance. Tensor-based methods explicitly model multidimensional correlations, achieving high accuracy but often at the cost of increased computational complexity.

Despite their clear advantages, combined methods can be computationally intensive, requiring significant computing resources and careful parameter tuning. Therefore, practical application often demands a balance between reconstruction accuracy, computational efficiency, and dataset characteristics. Critical evaluation of these hybrid methods indicates their substantial potential in addressing complex, large-scale remote sensing data challenges, justifying their adoption for sophisticated remote sensing analyses

3.5 Conclusion

The review shows that there's a clear gap between traditional methods' efficiency and modern methods' scalability and robustness. The limitations in generalization, complexity, or computational cost across existing techniques call for models

that can dynamically adapt to diverse spatial-temporal patterns while remaining efficient and interpretable. This forms the basis for adopting TabNet, a deep learning model specifically optimized for tabular, structured, and sparse data. It allows for learning complex spatial-temporal features through an attention-based architecture with reduced training costs. Complemented by a lightweight outlier detection strategy (OSR), the proposed approach balances performance and scalability, offering a feasible solution for operational EO systems.

The examined methods encompass a variety of statistical, machine learning, and deep learning techniques utilizing image based dataset. It is essential to comprehend the relative performance of different strategies utilized by these methods. The table (3.1) and 3.2 provides a concise summary of the outlier detection and data reconstruction techniques.

Table 3.1: Comparison of Outlier Detection Techniques

Method	**Advantages**	**Disadvantages**
Statistical [50, 52, 53]	- Mathematically robust	- Assumes specific distributions
Distance-Based [54, 55, 57, 58]	- Simple and intuitive	- Struggles with varying densities
Density-Based [59, 52]	- Identifies complex patterns	- Computationally complex
Depth-Based [60, 188]	- No specific model assumption	- Computationally intensive
Clustering [61, 63, 74]	- Summarizes data well	- Optimal cluster number is subjective
Neural Networks [63, 168]	- Models complex relationships	- Needs large data; risk of overfitting
SVM [64, 63]	- Good for linear/nonlinear data	- Complex kernel/parameter tuning

Table 3.2: Comparison of Data Reconstruction Techniques

Method	**Advantages**	**Disadvantages**
Image Inpainting [107, 109, 110]	- High detail preservation	- Not for large areas or complex scenes
Interpolation [112, 111, 117]	- Applicable for small gaps	- Less effective for complex terrains
Spectral Correlation [142, 139, 143]	- Utilizes band correlations	- Limited by band contamination
Temporal Replacement [148, 149]	- Uses clear observations over time	- Affected by rapid changes
Temporal Filtering [156, 158]	- Improves signal clarity	- May remove important variations
Dictionary Learning [162, 153]	- Captures complex patterns	- Requires significant resources
Spatio-Temporal Fusion [174, 112]	- Integrates multiple data sources	- Complex and resource-intensive
Cloud Removal [70, 39]	- Targets cloud cover effectively	- May introduce artifacts
Exemplar-Based [134, 136]	- Preserves texture and structure	- Can create repetitive patterns
Propagated Diffusion [123, 124]	- Preserves edges and details	- Risk of blurring in large gaps
Variation-Based [128, 129]	- Maintains image integrity	- Limited for large missing areas
Learning-Based [138]	- Learns spatial relationships	- Requires extensive training data
Spectral-Temporal [181, 182]	- Uses spectral and temporal info	- Less effective with cloud cover
Adaptive Weighted Tensor [153, 106]	- Incorporates multi-dimensional info	- High complexity and demands
Hybrid Methods [167, 168]	- Comprehensive recovery	- Requires advanced modeling

Chapter 4

Research Methodology

4.1 Introduction

This chapter outlines the comprehensive methodological framework adopted to address the dual challenge of outlier detection and the reconstruction of missing Land Surface Temperature (LST) data in remote sensing imagery. Given the growing reliance on high-resolution satellite observations for environmental analysis, ensuring the accuracy and completeness of such data is critical. This research adopts a two-pronged strategy combining traditional rule-based approaches with modern self-supervised learning methods to address these challenges effectively.

In the first part of this methodology, a temporal-spatial algorithm named Outlier Search and Replace (OSR) is proposed. This method uses Dynamic Time Warping (DTW) to detect inconsistencies across time-series satellite images and leverages spatial context for robust reconstruction of anomalous or missing values.

In the second part, the methodology transitions to a deep learning-based framework using TabNet, a self-supervised model specifically designed for structured tabular data. To enable this, raster-based satellite imagery is transformed into a structured tabular format comprising spatial coordinates and LST values. TabNet is then trained to learn latent spatial-temporal patterns for reconstructing missing values using masked self-supervised learning.

Together, these complementary approaches provide a complete and scalable solution for improving data quality in Earth observation datasets. The evaluation is based on both traditional statistical metrics (e.g., MSE, RMSE) and visual agreement techniques (e.g., Bland–Altman plots) to ensure practical and scien-

tific reliability.

4.2 Dataset Overview

The research focuses on the Beijing-Tianjin-Hebei region for data collection, which is a highly urbanised area located in the northern part of China. This area includes the capital city of the nation, Beijing, the harbour city of Tianjin, and the neighbouring province of Hebei. The region consists of 11 local administrative units covering Beijing, Tianjin, and Hebei Province, encompassing a total area of approximately 216,000 km^2. The Beijing-Tianjin-Hebei region is located between 36°03′N to 42°32′N latitude and 113°30′E to 119°15′E longitude. 4.1

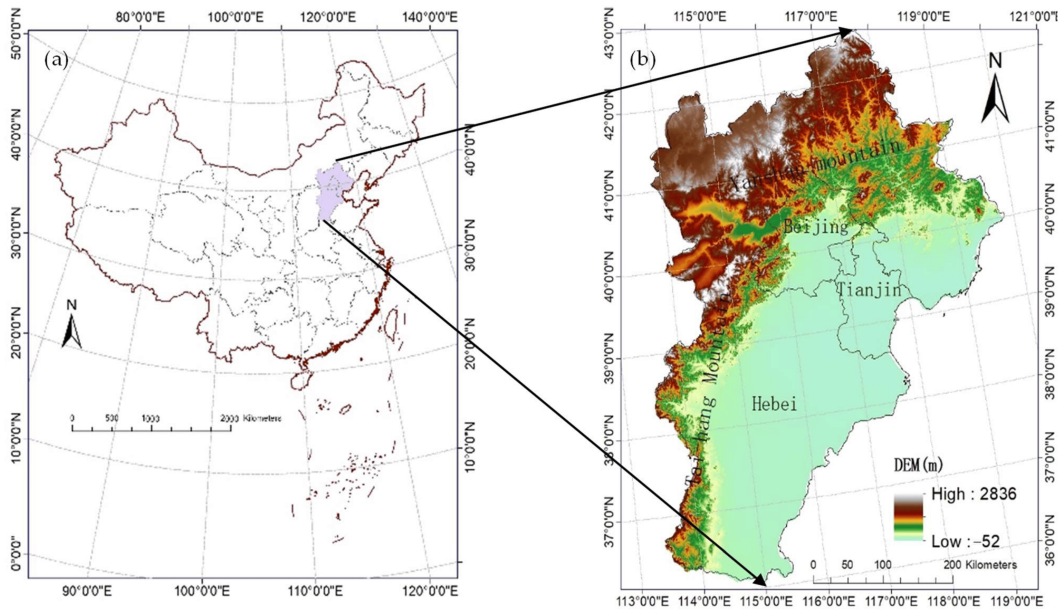

Figure 4.1: Study Area of Beijing-Tianjin-Hebei region

The Beijing-Tianjin-Hebei (BTH) region exhibits a varied geographical composition, encompassing highly crowded urban hubs such as Beijing and Tianjin, undulating terrains, fertile agricultural plains, and industrial zones. The presence of several elements contributes to the complexity and versatility of this field, specifically when studying the temperature of the Earth's surface. [189]. Hence the selection of the Beijing-Tianjin-Hebei region is motivated by its complex urban-rural transitions, high variability in LST, and frequent cloud coverage, offering a

suitable testbed for evaluating outlier detection and reconstruction techniques.

4.3 Exploratory Data Analysis

4.3.1 Spatial and temporal trends

Scatter plots were created for the first six months of 2017 (see Figure 4.2). These scatter plots helped explain the land surface temperature's spatial and temporal trends. Each study subplot shows data for a month, from January to June 2017. The subplots show land surface temperature dispersion using a colour map. The gradual shift in land surface temperature (LST) readings during the chosen months shows seasonality in this analysis's scatter plots. With seasonality and time-series analysis to observe persistent annual trends, land surface temperature's geographical and temporal behaviour can be better understood. This tabular method is notable given the paper's focus on structuring image-centric Earth observation data.

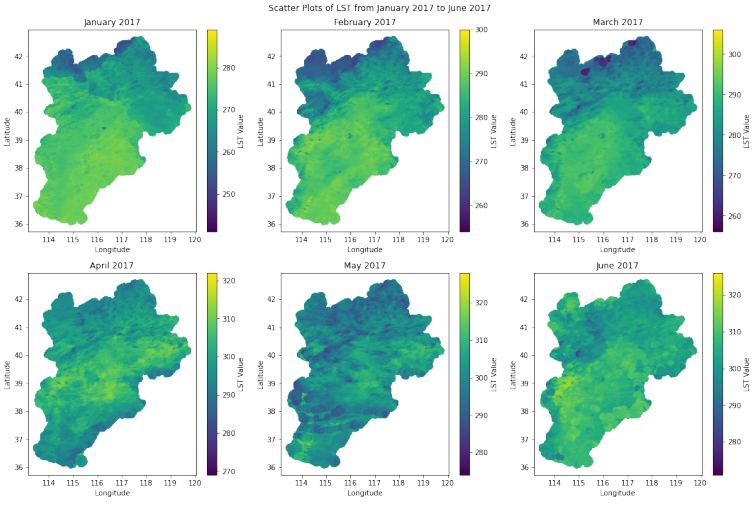

Figure 4.2: Scatter Plots of LST from January 2017 to June 201

4.3.2 Outliers in the Data

A box plot was created to analyse the distribution of land surface temperature (LST) values within the dataset. A box plot is a standardised method for vi-

sually representing the distribution of data using a summary of five values: the minimum, the first quartile (Q1), the median, the third quartile (Q3), and the maximum. (Figure 4.3)

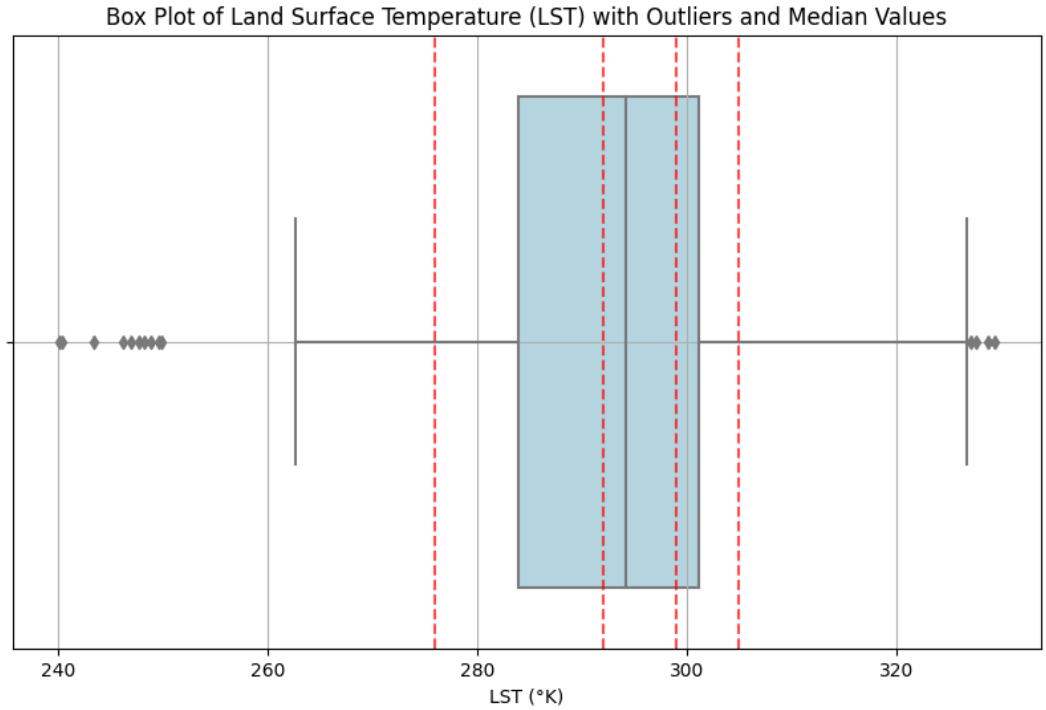

Figure 4.3: Box Plot of Land Surface Temperature (LST) with Outliers and Median Values

- Central Tendency and Variability: The median LST, depicted as the horizontal line inside the box, represents the centre inclination of the data. The interquartile range (IQR), represented by the box, offers information on the dispersion of the LST values and covers the range from the first quartile (Q1) to the third quartile (Q3).

- Outliers Identification: Outliers refer to individual data points that fall outside the range defined by the whiskers, which are extended to 1.5 times the Interquartile Range (IQR) from the quartiles. The spots, represented as dots outside the whiskers, indicate LST values that are much lower or higher than the rest of the range.

- Seasonal Median Indicators: The seasonal median indicators display the median values of land surface temperature (LST) for each season, namely

winter, spring, summer, and autumn. These values are shown by red dashed lines overlaid on the graph. These lines facilitate a comparative investigation of how median temperatures fluctuate throughout the seasons.

- Symmetry and Skewness: The box's symmetry and the positioning of the median suggest a distribution of LST values that is relatively symmetric. The existence of outliers at both extremes indicates that although the data is generally centred, there are particular cases where the LST (Land Surface Temperature) drastically deviates from the central values that are being addressed in this research.

4.3.3 Seasonal LST Distribution

The density plot, as shown in Figure 4.4, clearly exhibits a seasonal distribution of LST values, providing a precise representation of the probability density of temperatures. Every season, including winter, spring, summer, and autumn, is associated with a distinct colour, enabling a quick visual assessment of their distinctive temperature patterns.

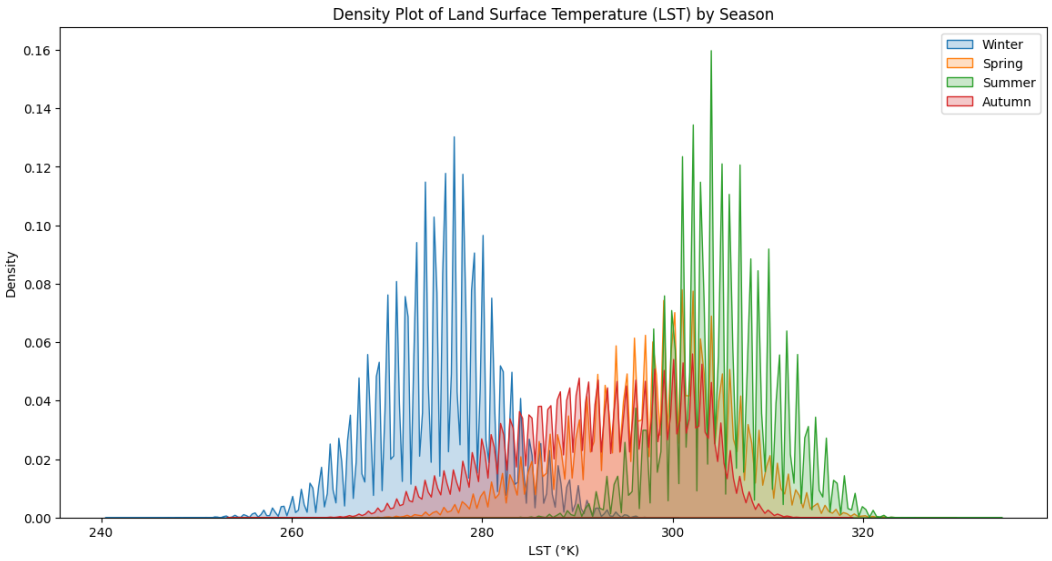

Figure 4.4: Density Plot of Land Surface Temperature (LST) by Season

It is observed that:

- The LST readings during the summer season display a distribution characterized by a higher average temperature, as anticipated given the warmer weather during this time of year.
- The winter season exhibits a distribution characterized by lower Land Surface Temperature (LST) values, which correspond to the colder environment during this time of year.
- Spring and autumn exhibit transitional patterns that connect the temperature disparity between the harsh seasons of winter and summer.

4.4 Evaluation Metrics

This section outlines the evaluation framework employed to examine the efficacy of the suggested models for outlier detection and reconstruction of lost big earth data. The proposed methodology features a two-stage structure, wherein the Outlier Search and Replace (OSR) algorithm conducts both outlier detection and reconstruction, while the TabNet model emphasises learning-based reconstruction from tabular data; consequently, different sets of metrics are employed to represent their specific objectives.

4.4.1 OSR Evaluation Metrics

The OSR algorithm performs two sequential tasks: (a) detection of outlying data points in satellite imagery using temporal-spatial comparisons, and (b) reconstruction of those detected outliers based on local neighborhood information.

A. Outlier Detection

The accuracy of outlier detection is evaluated using standard classification metrics:

- **Precision**

Measures the proportion of correctly identified outliers among all detected outliers.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**

  Measures the proportion of actual outliers that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score**

  Provides a balanced measure that considers both precision and recall.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are computed over different experimental setups, including varying window sizes and simulated outlier densities, to evaluate the sensitivity and generalization of the OSR detection mechanism.

B. Reconstruction Evaluation (within OSR)

For pixels identified as outliers, OSR reconstructs their values using spatial and temporal context. The reconstruction quality is assessed using:

- **Mean Squared Error (MSE)**

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Bland–Altman Plot**

  A graphical method to analyze the agreement between reconstructed and true values. It helps visualize systematic deviations or bias in reconstructed pixel values.

These evaluation tools help assess both numerical accuracy and statistical agreement, which are essential in remote sensing tasks where minor temperature deviations may have critical implications.

4.4.2 TabNet Evaluation Metrics

The TabNet model reconstructs missing values in tabular LST datasets derived from raster sources. It is trained to learn latent spatial-temporal patterns that support accurate prediction of missing data points. The following evaluation metrics are used to assess its performance:

- **Mean Squared Error (MSE)**

  Measures the average of the squared differences between predicted and actual LST values.

  $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  Lower MSE values indicate better prediction performance.

- **Root Mean Squared Error (RMSE)**

  Provides an interpretable error in the same unit as the target variable (Kelvin), offering insight into typical prediction deviations.

  $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Bland–Altman Plot**

  The Bland–Altman (BA) plot visually compares predicted LST values with actual values, plotting the mean of each pair against their difference. This method helps assess the level of agreement and any systematic bias. A good reconstruction model should show differences clustered around zero with narrow limits of agreement.

These metrics are applied across varying levels of missingness in the input data,

allowing evaluation of TabNet's robustness and generalization under different reconstruction challenges.

4.5 Conclusion

This chapter presented the dual-methodological design adopted in this study, offering both rule-based and machine learning-driven strategies for handling outliers in LST datasets. The OSR algorithm, grounded in temporal distance metrics and local spatial correction, provided an interpretable and robust approach for outlier detection and reconstruction in raster imagery. It was tested under multiple experimental scenarios, confirming its reliability across varying window sizes and anomaly densities.

In parallel, the TabNet model was introduced as a deep learning alternative capable of learning spatial-temporal patterns directly from structured tabular data derived from MODIS LST products. Through careful preprocessing and dataset restructuring, the model was trained using a masked self-supervised learning strategy, achieving low reconstruction errors and narrow agreement in Bland–Altman evaluations.

The integration of both techniques not only enabled comprehensive outlier detection and reconstructin but also validated the effectiveness of the tabular dataset structure itself. This methodological foundation forms the basis for subsequent results and discussions and supports the thesis's overarching aim of enhancing EO data reliability through hybrid, scalable solutions.

Chapter 5

Outlier Detection and Reconstruction of Lost Land Surface Temperature data in Remote Sensing imagery

5.1 Introduction

This chapter focuses on a crucial component of environmental monitoring: the precise measurement and analysis of the temperature of the Earth's surface. The terrestrial surface temperature data holds great importance in multiple environmental domains, particularly in agriculture, where it acts as a crucial indicator of atmospheric conditions. The collection of this data heavily relies on remote sensing instruments, which, despite being widely used, frequently face issues such as missing data and outliers. The main causes of these problems can be traced to the inherent limitations of remote sensing technologies and the unpredictable changes in meteorological conditions. In order to overcome these challenges, this chapter presents and investigates the Outlier-Search-and-Replace (OSR) algorithm. This novel approach is characterised by its utilisation of spatial and temporal data, allowing for the effective identification and reconstruction of absent data points in land surface temperature datasets. The OSR method plays a crucial role in improving the accuracy and comprehensiveness of temperature data, leading to enhanced quality in environmental evaluations and decision-making. The efficacy

of the OSR method is showcased by its utilisation on a dataset of land surface temperatures obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS) in January 2018. This case study demonstrates the algorithm's ability to locate and fill in missing data, highlighting its practical importance in environmental research. The knowledge acquired from this chapter is crucial for the overarching goals of this thesis, which aim to enhance the techniques for outlier detection and reconstruction of lost big earth data.

5.2 Proposed Model

Pseudo-code for the proposed model for outlier detection and reconstruction of missing data is shown in 1 and 2 respectively. Section 5.2.1 and 5.2.2 explains the working of both parts of algorithm respectively. The following section presents a detailed evaluation of the OSR algorithm under different experimental settings to validate its effectiveness in both outlier detection and data reconstruction.

5.2.1 Outlier Detection

The process involves identifying a Region of Interest (ROI) with dimensions $M \times N$. Within this ROI, the missing data are anticipated to be found and subsequently reconstructed. A similar region of interest (ROI), which will be indicated as ROI_w, will be chosen for each of the input photos in this early stage. Selecting the corresponding ROI_w refers to extracting a spatial window of fixed size $M \times N$ centered at the same pixel coordinates across each image I_i in the temporal sequence. This ensures that temporal comparisons using DTW are made between aligned regions across time, allowing consistent pixel-level tracking. The position of this ROI can either be moved across the image (in a sliding window fashion) or fixed based on a region of interest selected by the user.

Figure 5.1 illustrates the temporal mosaic created by stacking spatially aligned ROIs across multiple days. Within this mosaic, the value of each pixel loca-

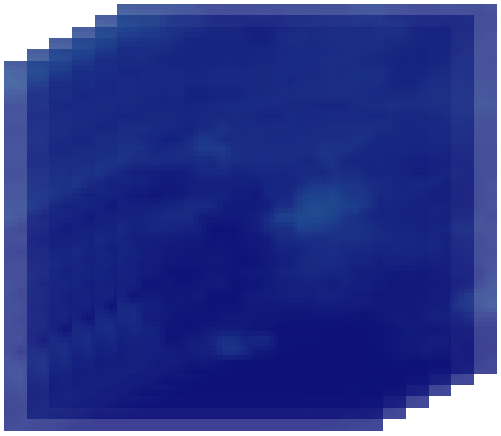

Figure 5.1: Mosaic image constructed by stacking $M \times N$ windowed views of Input Images

tion $P(j,k)$ can be tracked as a sequence through time. Dynamic Time Warping (DTW) is applied to compare such sequences between two consecutive slices (e.g., I_i and I_{i+1}) at the same pixel location using equation 5.1. Unlike Euclidean distance, which assumes linear alignment, DTW flexibly aligns sequences that may be non-uniform due to sensor noise or environmental fluctuations, thus making it particularly effective in detecting subtle dissimilarities caused by outliers or missing data. This ability to measure non-linear temporal dissimilarity ensures robust detection even when values shift slightly in time, which is common in satellite-based measurements.

$$C_d, D_d \;=\; DTW\ (I_{i\,P(j,k),}\, I_{i+1\,P(j,k),}) \tag{5.1}$$

The pixel value at location j,k in each succeeding day's input image area of interest ROI_w is denoted by the expression $P(j,k)$ respectively. In addition to storing the coordinates of each pixel that is being compared, the D_d variable stores the distance values that exist between the pixels of each subsequent image slice. In addition to the spatial information of data from the same satellite as well as numerous satellites, there is a high connection between the temporal LST values of the same area in consecutive days. This correlation is also present between the LST values of several satellites. When the pixel values are accurate,

a linear distance curve is generated; however, anytime there is a missing value in the image, the distance value is extremely high, which indicates that there is an abnormality. The following section will recreate these values after they have been detected and located as outliers based on a threshold value denoted by the capital letter D_d. At the moment, the threshold value, denoted by T_h, is being determined through experimentation by virtue of a comparison between real pixels and known outliers.

Algorithm 1 Outlier Detection using DTW

Require: A set of images $I = \{I_1, I_2, \ldots, I_n\}$, Region of Interest (ROI) size $M \times N$, Threshold T_h
Ensure: Set of detected outlier coordinates C_{do}
1: Define ROI_w of size $M \times N$ within each image in I
2: Initialize set of outlier coordinates $C_{do} \leftarrow \emptyset$
3: **for** each image slice I_i in I **do**
4: Select the corresponding ROI_w in I_i
5: Form a mosaic of temporal information using ROI_w from all I
6: **end for**
7: **for** each pixel $P(j,k)$ in ROI_w **do**
8: Calculate DTW distance between $P(j,k)$ in I_i and I_{i+1} using:

$$C_d, D_d = DTW(I_{i_{P(j,k)}}, I_{(i+1)_{P(j,k)}})$$

9: **if** $D_d > T_h$ **then**
10: Mark $P(j,k)$ as an outlier
11: Add $P(j,k)$ to C_{do}
12: **end if**
13: **end for**
14: **return** C_{do}

5.2.2 Missing Data Reconstruction

Following the identification of the specific coordinates of outlier values, which are labeled as C_{d_o} , the reconstruction process begins.

In order to restore these anomalies, the algorithm must first go through a process that is extremely precise. During this step, an exhaustive search is conducted throughout all the distance values of D_{d_o} in temporal slices contained within the selected image mosaic, with a particular focus on the coordinates of the outliers

C_{d_o}. By isolating a pixel value that displays the least amount of variance in terms of distance from the norm, the goal is to ensure that the pixel value is compatible with the data context that is surrounding it. The equation 5.2 that governs this selection procedure is as follows:

$$D_d \quad = Min(C_D, D_D\ ) \qquad for\ \ 1 \leqq\ R\ \leqq\ N \tag{5.2}$$

C_D in 5.2 gives the location of the pixel P_{C_D} with the lowest distance D_d between pixel of image slices in mosaic at location similar to outlier C_{d_o}. The pixel P_{C_D} is taken as a reference pixel to reconstruct the outlier. Now a $3X3$ window W_o is taken around the location of the outlier, and the reference pixel P_{C_D} is compared using 5.2 with the neighbouring pixels of the outlying value in W_o. The pixel that is at the lowest distance from the reference pixel is copied at the outlier location. This algorithm is similar to the spatial reconstruction of missing values, but in this case, temporal information is also being utilised to identify the outliers.

Algorithm 2 Reconstruction of Outliers using Spatial Neighborhood

Require: Set of outlier coordinates C_{do}, Temporal slices $I = \{I_1, I_2, \ldots, I_n\}$
Ensure: Reconstructed image data

for each outlier coordinate C_{do} in C_{do} **do**
    Find pixel P_{C_D} with minimum DTW distance in temporal slices:
    $D_d = \min(\{C_D, D_D\})$ for $1 \leq r \leq N$
    Set reference pixel $P_{ref} = P_{C_D}$
end for
for each outlier coordinate in C_{do} **do**
    Create a 3×3 window W_o around the outlier location
    for each pixel P' in W_o **do**
        Calculate distance to P_{ref}
        if distance $<$ min_distance **then**
            Set min_distance = distance
            Set replacement_pixel = P'
        end if
    end for
    Replace the outlier pixel in C_{do} with replacement_pixel
end for
return the reconstructed image data

5.3 Experimental Results

This section presents the experimental evaluation of the OSR algorithm. The dataset used is obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS), focusing on the Beijing-Tianjin-Hebei (BTH) region, as previously introduced in Section 4.2.

To simulate missing or anomalous data, random outliers were introduced into fixed-size input windows $M \times N$. The effectiveness of the algorithm is assessed by comparing the reconstructed images to the original, uncorrupted ones.

Figure 5.2 illustrates sequential LST images captured over four consecutive days. These temporally aligned images form a mosaic that serves as the input for OSR, enabling pixel-wise temporal comparisons. The temporal behavior of each pixel location is analyzed using DTW to identify outliers and reconstruct missing values.

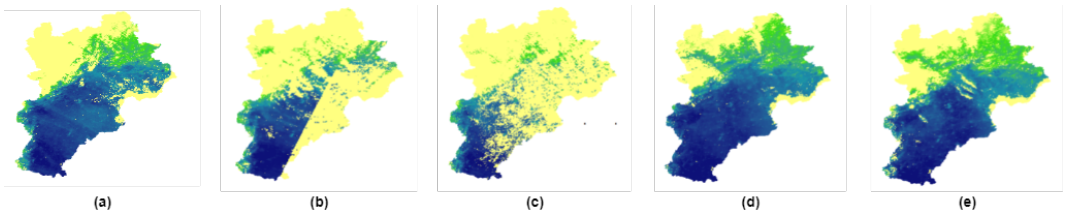

Figure 5.2: Sequential Daily MODIS Images Across a Four-Day Period

Three evaluation scenarios were explored to analyze the effect of varying conditions:

- **Scenario 1:** Different window sizes with varying numbers of random outliers.
- **Scenario 2:** Different window sizes with a fixed number (30) of random outliers.
- **Scenario 3:** Fixed window size (10×10) with an increasing number of outliers.

For each scenario, two performance aspects are analyzed:

1. **Outlier detection accuracy** — evaluated using precision, recall, and F1 score.

2. **Reconstruction quality** — measured using Mean Squared Error (MSE).

5.3.1 Outlier Detection Evaluation

Table 5.1 shows the results for Scenario 1. The 10×10 window provides the best balance, achieving the highest F1 score. Smaller windows struggled to identify patterns due to limited spatial context, while larger windows suffered from over-smoothing.

Table 5.1: Scenario 1 — Varying Window Sizes and Outlier Counts

Window Size	Outliers	Precision	Recall	F1 Score
4×4	5	0.333	0.200	0.250
6×6	10	0.000	0.000	0.000
8×8	15	0.778	0.467	0.583
10×10	32	0.857	0.750	0.800
15×15	62	0.808	0.339	0.478

Table 5.2 evaluates detection when a fixed number of outliers (30) is injected. The 10×10 configuration again performs best across all metrics.

Table 5.2: Scenario 2 — Fixed 30 Outliers with Varying Window Sizes

Window Size	Precision	Recall	F1 Score
4×4	0.400	0.267	0.320
6×6	0.480	0.433	0.455
8×8	0.667	0.700	0.684
10×10	0.857	0.767	0.810
15×15	0.722	0.567	0.635

Table 5.3 presents the results for Scenario 3, showing how detection accuracy evolves as the number of outliers increases for a fixed 10×10 window. While precision remains stable, recall slightly drops, reflecting the increasing challenge in detecting more anomalies.

Table 5.3: Scenario 3 — Increasing Outliers with Fixed 10×10 Window

Outliers Injected	Precision	Recall	F1 Score
5	0.833	0.800	0.816
10	0.875	0.800	0.836
15	0.857	0.733	0.790
20	0.857	0.750	0.800
25	0.857	0.720	0.782
30	0.857	0.767	0.810
35	0.840	0.743	0.788
40	0.833	0.725	0.776
45	0.820	0.689	0.748
50	0.810	0.660	0.727

5.3.2 Reconstruction Evaluation

Following the identification of outliers, the OSR algorithm proceeds with reconstruction by leveraging spatial and temporal cues around the detected coordinates. The accuracy of reconstruction is evaluated using the Mean Squared Error (MSE), representing the average squared difference between original and reconstructed LST values.

The reconstruction is tested across the same three scenarios described earlier:

- **Scenario 1:** Different window sizes and different numbers of outliers.
- **Scenario 2:** Fixed number of outliers (30) with different window sizes.
- **Scenario 3:** Increasing outlier count with fixed 10×10 window.

Scenario 1: The results in Table 5.4 show that 10×10 again provides the lowest MSE (2.25 K^2), confirming its effectiveness for both detection and reconstruction.

Table 5.4: Scenario 1 — Different Window Sizes with Varying Outlier Counts

Window Size	MSE (K^2)
4×4	9.30
6×6	11.90
8×8	6.76
10×10	2.25
15×15	4.84

Scenario 2: When 30 random outliers are introduced across different window sizes (Table 5.5), the 10×10 window again exhibits the best reconstruction accuracy. This consistency highlights the robustness of the window size across varying conditions.

Table 5.5: Scenario 2 — Fixed 30 Outliers with Different Window Sizes

Window Size	MSE (K^2)
4×4	5.95
6×6	4.83
8×8	3.80
10×10	2.25
15×15	3.46

Scenario 3: Table 5.6 shows reconstruction performance when the number of injected outliers increases gradually while keeping the window size fixed at 10×10. The MSE increases steadily, reflecting a logical performance degradation as more outliers make reconstruction increasingly difficult.

Table 5.6: Scenario 3 — Fixed 10×10 Window with Increasing Number of Outliers

Outliers Injected	MSE (K^2)
5	2.20
10	2.50
15	2.90
20	3.10
25	3.80
30	4.00
35	4.30
40	4.60
45	5.00
50	5.30

Visual Example: A side-by-side image comparison is presented in Figure 5.3, showing artificially introduced outliers (top row) and the corresponding reconstructed results (bottom row).

As the visual comparison and MSE values indicate, the 10×10 configuration consistently outperforms other window sizes, striking the best trade-off between

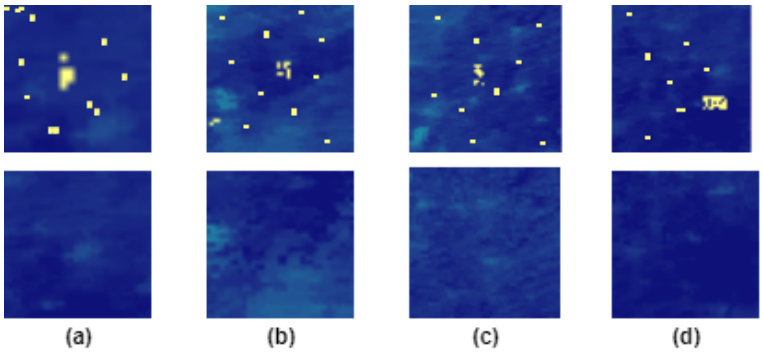

Figure 5.3: Outlier and Reconstructed Image Pair

context and localization. Too small windows (e.g., 4×4) lack sufficient spatial information, whereas overly large windows (e.g., 15×15) dilute important local variations, increasing reconstruction error.

5.3.3 Bland–Altman Plot Across Scenarios

To further assess the accuracy of the OSR reconstruction, Bland–Altman (BA) plots are used. These plots compare the difference between original and reconstructed Land Surface Temperature (LST) values against their average, providing insight into bias, spread, and consistency.

BA plots are generated for all three experimental scenarios. In each case, the performance of different window sizes is visualized, while the 10×10 configuration is highlighted due to its consistent performance across all metrics.

Scenario 1: Varying Window Sizes with Varying Outlier Counts Figure 5.4 shows the BA plot for the first scenario. The 10×10 configuration yields the narrowest limits of agreement and the smallest mean difference, aligning with its lowest MSE in Table 5.4. The spread for other window sizes is wider, indicating less reliable reconstruction.

Scenario 2: Fixed 30 Outliers with Varying Window Sizes In Figure 5.5, the 10×10 window again demonstrates stable reconstruction, with differences tightly clustered around the zero line and narrower agreement bounds compared

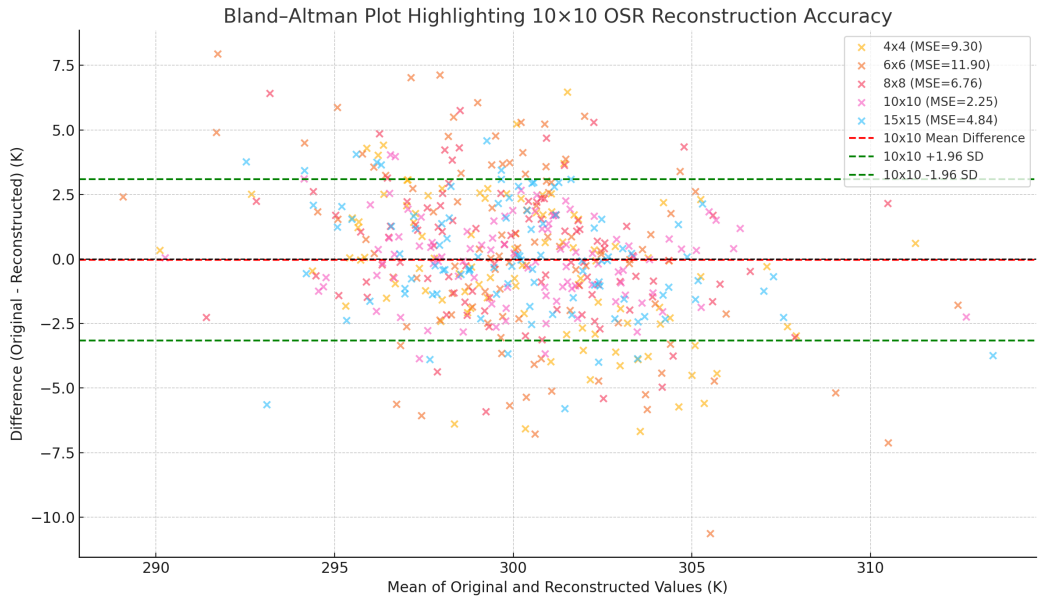

Figure 5.4: Bland–Altman Plot for Scenario 1: Varying window sizes and outlier counts. 10×10 window exhibits minimal bias and narrow agreement range.

to other windows. This aligns with the lowest MSE reported in Table 5.5.

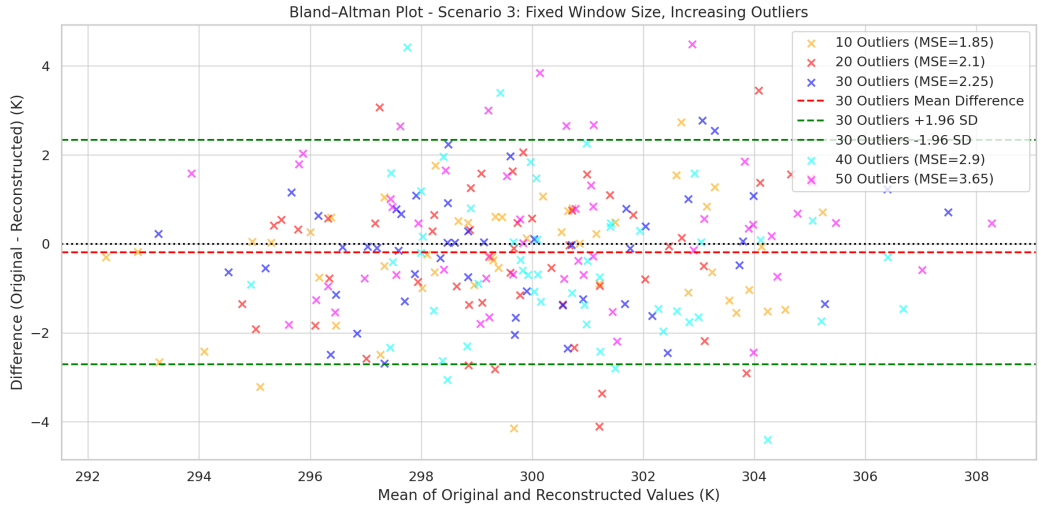

Figure 5.5: Bland–Altman Plot for Scenario 2: Fixed 30 outliers. 10×10 window shows superior consistency and reduced bias.

Scenario 3: Fixed 10×10 Window with Increasing Outliers Figure 5.6 visualizes how reconstruction error grows as more outliers are introduced. The spread in differences increases with higher outlier counts, confirming the trend in Table 5.6. This supports the sensitivity of the algorithm to the number of anomalies in the ROI.

Figure 5.6: Bland–Altman Plot for Scenario 3: Fixed 10×10 window with increasing outlier count. Spread increases with more outliers, reducing accuracy.

Overall, these plots offer visual confirmation of the numerical results and provide an intuitive understanding of the OSR algorithm's reconstruction reliability under various conditions. The 10×10 window size consistently yields the most stable and least biased reconstructions.

5.4 Conclusion

This chapter presented the Outlier Search and Replace (OSR) algorithm for detecting and reconstructing missing or anomalous values in Land Surface Temperature (LST) imagery obtained through remote sensing. The algorithm leverages both spatial and temporal dependencies to locate abnormal patterns across image sequences and then reconstructs them using neighborhood-based correction informed by dynamic time-based comparisons.

To evaluate the performance of the proposed algorithm, three experimental scenarios were tested: (i) varying window sizes with varying outlier counts, (ii) fixed outlier count across varying window sizes, and (iii) increasing outlier counts with a fixed 10×10 window. Each scenario was designed to assess how different factors influence the precision and reliability of both detection and reconstruction stages. In the outlier detection phase, results consistently showed that the 10×10 window

size offers the best balance between precision and recall, with the highest F1 score across all scenarios. Smaller window sizes lacked sufficient spatial context, while larger windows led to over-smoothing and reduced sensitivity.

For reconstruction, the algorithm was evaluated using Mean Squared Error (MSE) and Bland–Altman plots. In all three scenarios, the 10×10 window configuration yielded the lowest MSE and the most stable reconstruction behavior, with narrow limits of agreement and low bias. The degradation in accuracy as the number of outliers increased further confirmed the algorithm's sensitivity to anomaly density, emphasizing the need for optimal window tuning.

These results demonstrate the robustness of the OSR method under diverse conditions, and validate its effectiveness for practical applications where data quality is impacted by sensor noise or atmospheric disruptions. This lays the foundation for more scalable approaches, which are explored in the next chapter using learning-based models for advanced reconstruction across broader datasets.

Chapter 6

Outlier Detection and Reconstruction in Big Earth Data Using Tabular Self-Supervised Learning

6.1 Introduction

Satellite-based remote sensing offers large-scale and frequent observations of land surface temperature (LST), a critical environmental variable for understanding climate behavior, vegetation health, and hydrological processes [190]. However, the inherent limitations of satellite instruments, such as cloud contamination, sensor noise, and missing acquisitions, often result in incomplete or corrupted datasets [37, 40].

While traditional anomaly detection and reconstruction methods (e.g., the OSR algorithm from Chapter 5) have shown strong performance in spatially local and temporally short contexts [101], their effectiveness diminishes when faced with widespread or complex data gaps.

To address this challenge, this chapter explores a deep learning-based solution leveraging the TabNet architecture. TabNet is a recent innovation designed specifically for tabular data [191]. Its ability to perform sparse feature selection using attention-based decision steps enables it to learn meaningful patterns from large geospatial datasets without requiring intensive feature engineering. Unlike standard fully-connected networks, TabNet preserves interpretability and is naturally suited to structured environmental data.

This chapter investigates how TabNet can be self-supervised to reconstruct missing LST values across large spatiotemporal domains. We focus on building a tabular representation of the MODIS LST product for the Beijing-Tianjin-Hebei region and training TabNet to learn the underlying temperature patterns from partial observations. This approach complements the OSR algorithm by providing a more scalable, learning-driven reconstruction method suitable for long-range data completion.

6.2 Data Collection and Processing

6.2.1 Data Collection

The Land Surface Temperature (LST) dataset used in this chapter was collected from the MODIS (Moderate Resolution Imaging Spectroradiometer) satellite product for the Beijing-Tianjin-Hebei (BTH) region. MODIS LST products provide global coverage with a spatial resolution of 1 km and temporal frequency of four observations per day (day/night for both Terra and Aqua satellites), making them well-suited for time-series temperature reconstruction tasks https://ladsweb.modaps.eosdis.nasa.gov accessed on 10 June 2023. Each MODIS granule was clipped to the BTH region and stored as georeferenced raster imagery. The dataset was obtained during a period of three years, specifically from January 1, 2017, to December 31, 2019. In instances where data is not obtainable for a certain time period, said period is duly recorded within a compilation of time periods characterized by the absence of data.

6.2.2 Data Processing

The data preprocessing pipeline, shown in Figure **??**, includes five key steps: projection, gap detection, value assignment, unit conversion, and tabular restructuring.

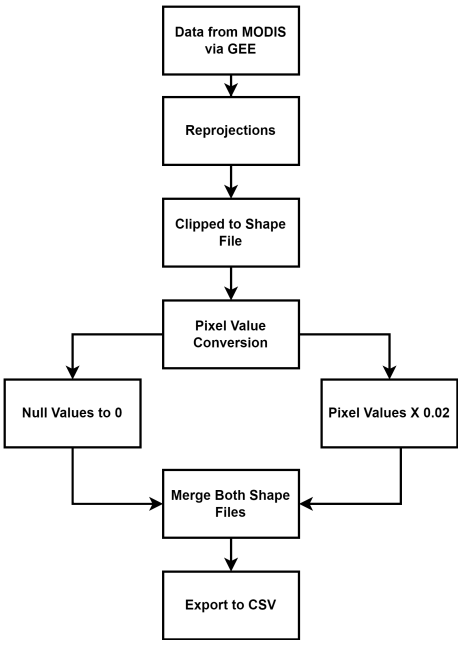

Figure 6.1: Data Collection & Processing

The first two steps—geo-referencing/projection and gap detection—were adopted directly from standard MODIS processing guidelines [192].

The next three steps were developed as part of this research to make the data suitable for deep learning-based modeling:

- **Value Assignment:** All missing values were temporarily replaced with zeros to maintain tensor shape during training. While zero-filling is simplistic, it was used in conjunction with a masking layer in TabNet to prevent the model from learning from placeholders.

- **Temperature Conversion:** LST values, originally in Celsius, were rescaled to Kelvin units for consistency with geoscientific standards and to ensure non-negativity of temperature inputs.

- **Tabular Restructuring:** The most significant processing step involved transforming raster images into tabular format. Each row represents a unique pixel, with columns for spatial coordinates, and LST value. This flattening allows TabNet to process the data as structured sequences with partial observations.

While some operations (e.g., zero-filling and Kelvin scaling) are standard preprocessing tasks, the design and implementation of the tabular transformation—particularly

the alignment of spatial-temporal features across images—form a key technical contribution of this chapter.

This processed data enables pixel-level learning in a self-supervised manner, where the model is trained to predict masked LST values from their spatial and temporal context without requiring external labels.

6.3 Experimental Results

The dataset captures essential geographical and meteorological parameters relevant for the study of land surface temperature. It consists of four key parameters. Table **??** shows the glimpse data and parameters.

Table 6.1: Tabular LST data of BTH Region

Latitude	**Longitude**	**LST* (K)**	**Digital Number (DN)**
114.5487	42.1265	0	NULL
114.4229	40.81495	302	15093
114.2343	40.80597	305	15263
114.2702	40.80597	303	15167

The Beijing-Tianjin-Hebei dataset records latitude and longitude in degrees for precise geographic identification. Land Surface Temperature (LST) readings in Kelvin are essential for local climate analysis, vegetative health studies, and heat stress assessment in urban and rural locations. A Digital Number (DN) values are compressed LST values that can be translated to Kelvin using a scaling factor of 0.02. This conversion approach optimises data efficiency and allows more extensive analytics.

6.3.1 Exploratory Data Analysis of Generated Dataset

The accuracy and reliability of the unique tabular dataset is confirmed through a thorough investigation and rigorous analysis. Subsequent sections will establish this argument with the help of different analysis.

6.3.2 Temporal Trend Analysis

The organised tabular format of the dataset for thorough and systematic computations of seasonal temperature fluctuations, which played a vital role in verifying the dataset's precision and dependability. Figures 6.2 and 6.4 present a comprehensive analysis of the temporal fluctuations in Land Surface Temperature (LST) throughout a wide geographic region of Beijing-Tianjin-Hebei region situated within China, covering the period from January 2017 to end of 2019. Figure 6.2 depicts the average Land Surface Temperature (LST) on a daily basis, while Figure 6.4 provides a supplementary perspective by showing the weekly averages. Both representations consist of data points that represent the average temperature values derived from a large number of XY coordinate positions. The daily time series graph (Figure 6.2) displays the rapid changes in tempera-

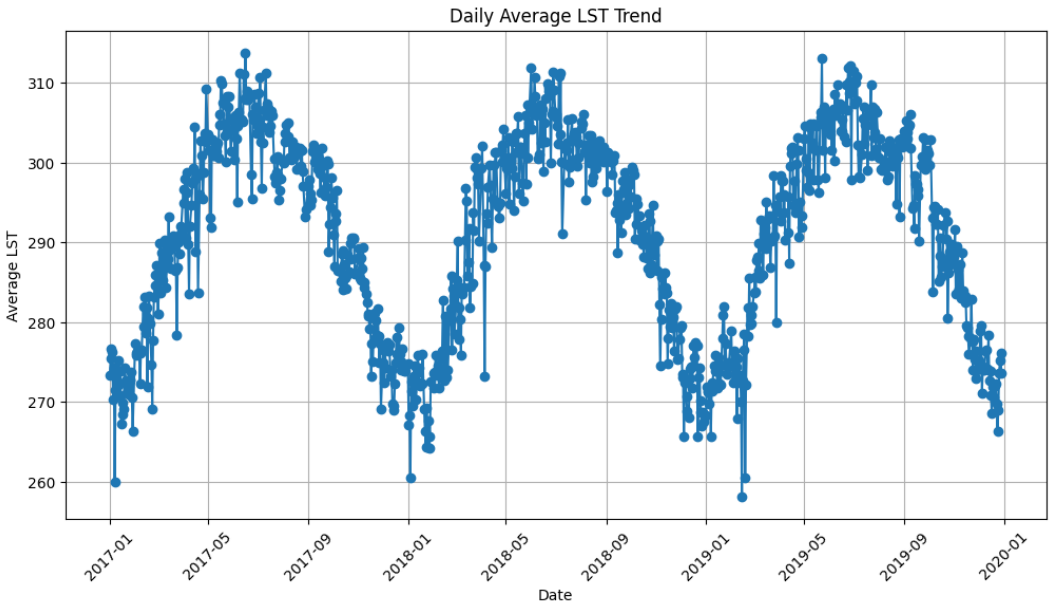

Figure 6.2: Daily LST Average Trend

ture, providing a detailed view of the day-to-day variations. This high-resolution representation enables the examination of brief irregularities and severe weather occurrences that might be concealed in data that is more consolidated. The weekly average graph (Figure 6.4) mitigates the daily fluctuations to provide a more distinct depiction of the long-term trends and patterns. Both graphs display

a prominent cyclical pattern that aligns with the seasonal variations in temperature. The peaks, which usually occur in the middle of the year, indicate the summer season, while the troughs at the end and beginning of the year indicate the winter season.

The plots display inter-annual changes, indicating a constant seasonal effect but also slight year-to-year differences in temperature. These variations may arise from natural climate oscillations, such as El Niño or La Niña phenomena [193, 194], or human-induced factors, such as alterations in land use or urbanisation [195].

Figure 6.3 shows daily average LST from 2017 to 2019, overlaid with periods of El Niño (in red) and La Niña (in blue) based on NOAA's ENSO classification. As evident, the early 2017 La Niña phase corresponds to a cooler period, while the El Niño event between late 2018 and mid-2019 shows higher LST values, suggesting a potential linkage between ENSO phases and LST fluctuations [196].

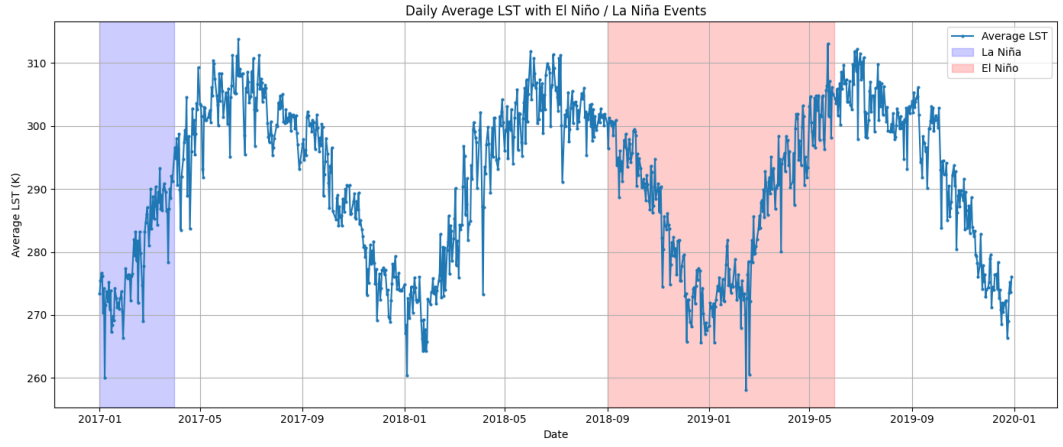

Figure 6.3: Daily average Land Surface Temperature (LST) with El Niño (red) and La Niña (blue)

The scatter of daily data points in Figure 6.2 emphasises the fluctuation in Land Surface Temperature (LST) that may result from daily temperature changes, localised micro-environments, or discrepancies in data collection. Meanwhile, the more streamlined weekly graph depicted in Figure 6.4 may help emphasise overarching patterns and changes that could suggest alterations in climatic patterns or the effectiveness of long-term environmental initiatives.

The dataset's robustness and the data collection process are emphasised by the consistent seasonal trends shown in both panels. Nevertheless, the existence of exceptional data points and yearly variations emphasised by the daily graph are the outliers which are essentially addressed in this research.

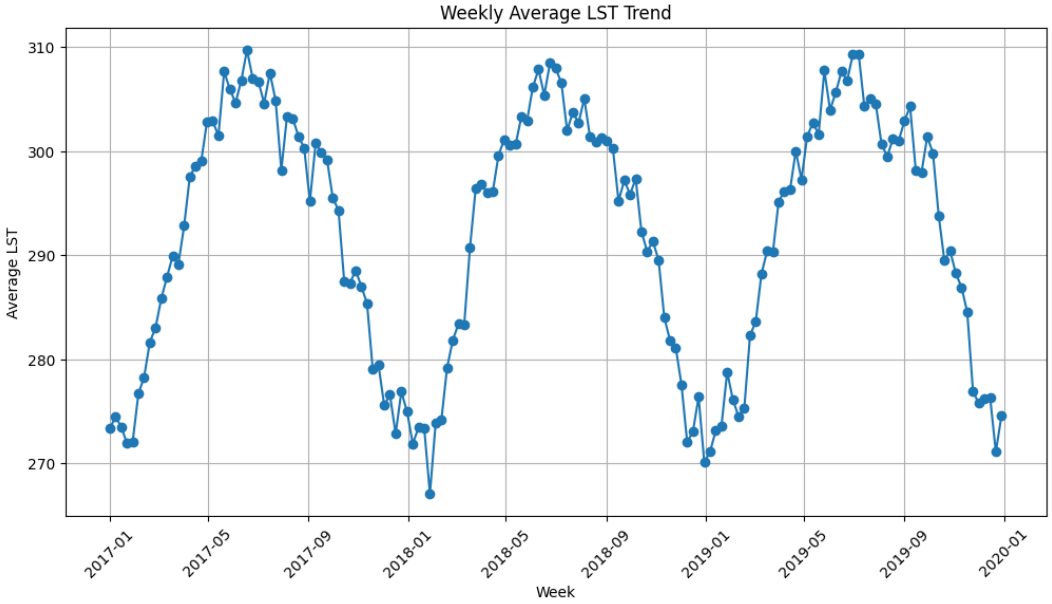

Figure 6.4: Weekly Average LST Trend

The findings shown in Figure 6.6, illustrates the average fluctuations in Land Surface Temperature (LST) on both a monthly and quarterly basis. The graph effectively demonstrates a coherent trend in Land Surface Temperature (LST) over the duration of the three-year investigation.

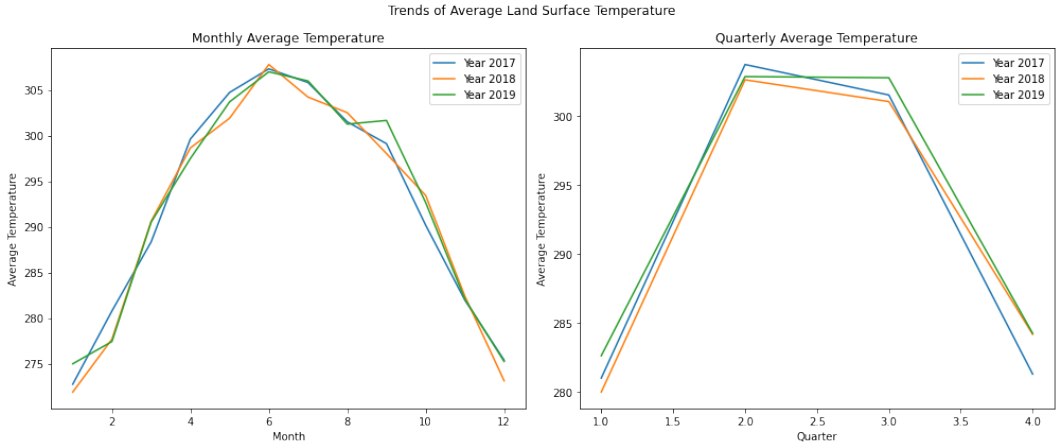

Figure 6.5: Trend of Average Land Surface Temperature

6.3.3 Seasonal Trend Analysis

The seasonal average Land Surface Temperature (LST) trends, as depicted in Figure 6.6 and seasonal mean and median values are shown in table 6.2.The results from a thorough investigation and provide a clear representation of the temperature changes across the seasons.

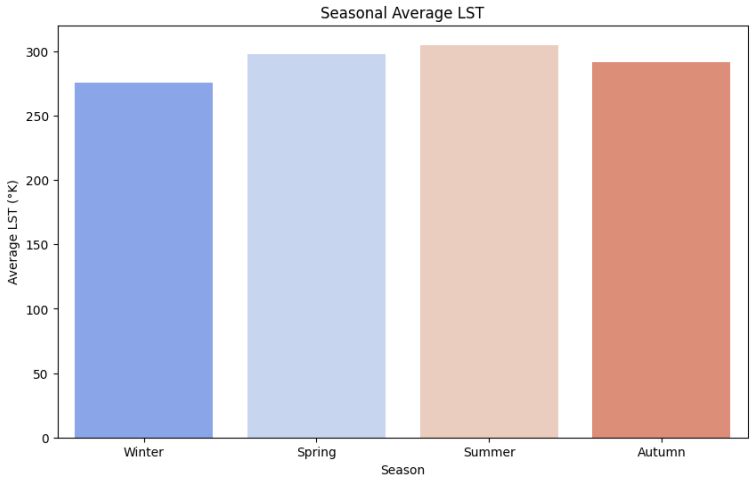

Figure 6.6: Seasonal Average LST Trend Analysis

Table 6.2: Seasonal Average Temperatures

No.	Season	Mean (K)	Median (K)
1	Winter	275.44	276.00
2	Spring	297.77	299.00
3	Summer	305.02	305.00
4	Autumn	291.88	292.00

- Winter: The dataset precisely reflects the anticipated decrease in average Land Surface Temperature (LST) to 275.44°K, which corresponds to the reduced amount of sunlight throughout the winter season. The median value corresponds to the mean, confirming the dataset's coherence its reliability in depicting winter temperatures.(Table 6.2)

- Spring: The dataset clearly shows a significant rise in average Land Surface Temperature (LST) to 297.77°K during the spring season, indicating a shift

towards higher temperatures. The dataset's ability to accurately record the beginning of higher temperatures is confirmed by the slightly higher median value, which is consistent with expected seasonal warming. (Table 6.2)

- Summer: The highest average Land Surface Temperature (LST) of 305.02°K exactly matches the usual conditions throughout summer, providing more evidence of the dataset's accuracy. The precision of the median value at 305.0°K demonstrates the dataset's resilience in continuously capturing the highest point of seasonal warmth. (Table 6.2)

- Autumn: The dataset accurately documents a decline in average Land Surface Temperature (LST) to 291.88°K, in line with the anticipated cooling over the fall season. The close proximity of the median to the mean confirms the dataset's trustworthiness in accurately following the progressive decrease in seasonal temperatures. (Table 6.2)

6.4 Proposed Model

TabNet, introduced in 2019 by [191], represents a significant advancement in the field of deep neural networks (DNN) for handling tabular data.
The application of TabNet as self-supervised learning (SSL) model offers a new and promising approach in the field of large-scale Earth data processing . This research relies on TabNet regressor, a transformer-based model that is widely recognised for its exceptional performance in handling tabular data. TabNet's distinctive features render it very suitable for addressing the specific issues posed by the dataset employed in this work. The use of TabNet is motivated by its ability to handle missing values in tabular form without requiring heavy preprocessing or imputation. Unlike conventional dense neural networks or tree-based models, TabNet's sparse feature selection ensures that only the most informative features are considered during training. This is highly beneficial for LST datasets, where missing data is widespread and patterns are often nonlinear and

region-specific. Moreover, its built-in interpretability supports transparency in environmental modeling, which is critical in scientific and policy applications.

To simulate missing data scenarios and assess model performance, a self-supervised learning approach is adopted. In this approach, random subsets of the data are masked during training, and the model learns to reconstruct the missing values. This strategy helps the model generalize better for real-world missing data and aligns with the problem context defined in previous chapters.

The following are the primary characteristics of TabNet that have been utilised:

- **Direct Input of Raw Data:** TabNet performs well in directly handling unprocessed tabular data, reducing the requirement for substantial preparation. This capacity is essential for managing the varied and unprocessed attributes of the Earth data in our dataset.

- **Sequential Attention Mechanism:** TabNet utilises a hierarchical structure of attention modules, which take feature samples as input and selectively focus on a subset of features at each stage using a trainable mask represented by a matrix $M \in \mathbb{R}^{n \times d}$. This method is crucial for the model's capacity to focus exclusively on essential characteristics, hence improving both accuracy and efficiency.

- **Feature Selection and Transformation:** The model employs a mask, generated using an attentive transformer, to perform gentle feature selection at each iteration, taking into account the utilisation of features in prior iterations. The masked features are efficiently encoded by undergoing transformation through fully connected (FC) layers, batch normalisation (BN), and gated linear units (GLUs.

- **Self-Supervised Learning for Anomaly Detection:** By undergoing training to forecast added anomalies and missing data, the model acquires a more profound comprehension of dataset patterns and irregularities. The

SSL technique is exceptionally efficient at identifying outliers and reconstructing data in large tabular dataset.

The sequential attention mechanism and sparse feature selection capabilities of TabNet are crucial for effectively managing the intricacies of Earth data analysis. Following several iterations of feature masking and transformation, the ultimate encoding is employed to achieve precise prediction in subsequent tasks, while back propagation fine-tunes the network's weights.

6.4.1 TabNet Architecture

TabNet is an interpretable deep learning architecture specifically designed for tabular data [191]. It combines the power of gradient boosting-like feature selection with the flexibility of deep neural networks. Unlike traditional dense networks that use all features uniformly, TabNet employs sequential attention mechanisms that enable the model to focus only on the most relevant features at each decision step. This allows it to learn sparse and diverse feature masks over time, improving both efficiency and interpretability. The architecture consists of a shared feature transformer followed by multiple decision steps. Each decision step contains an attentive transformer that learns a mask over input features and a feature transformer block that processes selected features to make partial predictions. These partial predictions are then aggregated across decision steps to produce the final output. A key advantage of TabNet is its built-in interpretability. The learned feature masks allow practitioners to inspect which input features contribute most to the prediction. Additionally, TabNet supports self-supervised learning by masking inputs and reconstructing them, making it suitable for tasks like missing data reconstruction, as demonstrated in this chapter. Compared to conventional fully connected networks or tree-based ensembles, TabNet maintains end-to-end differentiability and requires minimal preprocessing, which is ideal for structured environmental datasets such as LST values, where patterns vary temporally and spatially.

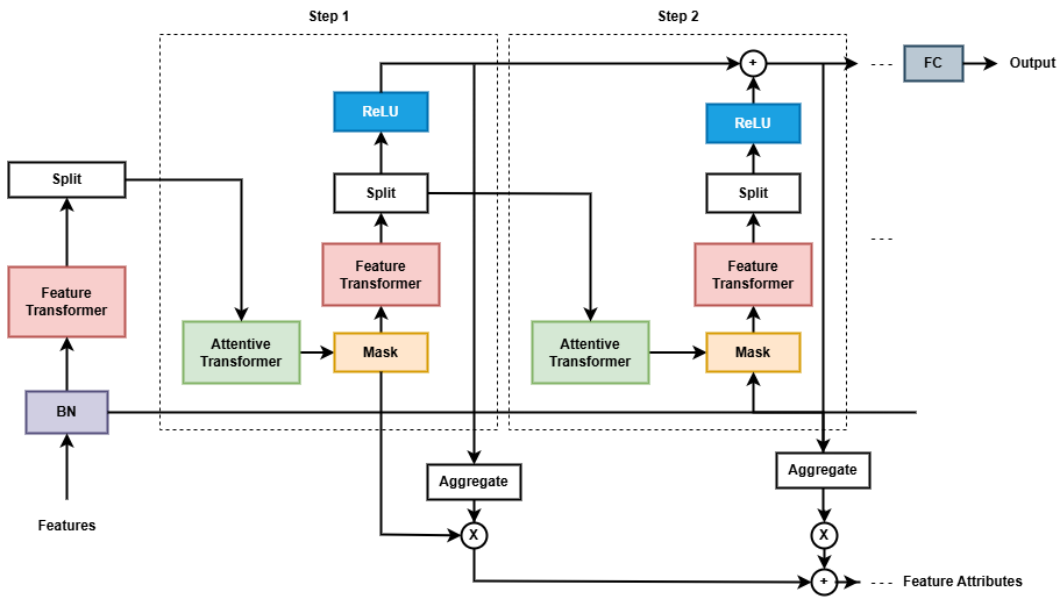

Figure 6.7: : TabNet encoder architecture

The TabNet encoder comprises multiple sequential steps, each containing three main components:

- **Feature Transformer:** This comprises of a sequence of fully connected (FC) layers, batch normalisation (BN), and a Gated Linear Unit (GLU). The feature transformer encodes the features and divides the output into two components: one component directly feeds into the decision-making process, while the other is forwarded to the next stage. This finally links to a residual connection, incorporating normalisation with a value of $\sqrt{0.5}$. Refer to Figure 6.12 for a concise visual representation. The selection of this specific normalisation method is based on the fact that it decreases the variability in the network. Once the features are successfully encoded using the feature transformer, the output is divided into two embeddings. The first embedding, $d_i \epsilon R^{N \times M_d}$, directly influences the decision-making process. The second embedding, $a_i \epsilon R^{N \times M_a}$, is passed on to the subsequent phase. The hyperparameters M_d and M_a can be adjusted.

- **Feature Selection:** An acquirable mask is employed to choose prominent characteristics at every stage. The mask is applied by performing element-wise multiplication on the tabular data. The computation involves utilising

an attentive transformer that merges the processed characteristics from the preceding stage with a prior, and subsequently undergoes a function known as sparsemax as shown in equation 6.1. The prior is a quantification of the extent to which a specific feature has been utilised in preceding stages, with a gamma parameter (γ) regulating the focus on various features.

$$M_i = \text{sparsemax}(P_{i-1} \cdot h_i(a_{i-1})) \tag{6.1}$$

where h_i is a trainable function, a fully-connected network in this case. The prior P_i is a measure of how much a particular feature has been used in equation 6.2

$$P_i = \prod_{j=1}^{i} (\gamma - M_j) \tag{6.2}$$

Given an initial value of $P_0 = 1^{MXN}$, we have a grid of size $N \times M$. Given that all elements in M are inside the range of zero and one, if γ is equal to 1, the subtraction of the mask will result in certain features approaching zero, thereby excluding them from the decision-making process. However, when γ is significantly larger than 1, eliminating M will result in comparable values in the previous, so promoting a more evenly distributed focus on the various features. This elucidates the rationale behind the definition.

- **Decision Making:** Similar to a decision tree, decision embeddings from each step are passed through a ReLU activation function, aggregated, and then passed through a fully connected layer to make the final decision. (Equation 6.3)

$$y = FC\left(\sum_{i=1}^{N_{\text{steps}}} \text{ReLU}(d_i)\right) \tag{6.3}$$

- **Decoder:**The decoder architecture is specifically tailored to enhance SSL (Self-Supervised Learning) inside the framework of TabNet. The purpose of this decoder is to restore the missing columns in tabular data. The technique entails generating a feature mask, denoted as matrix S, which is

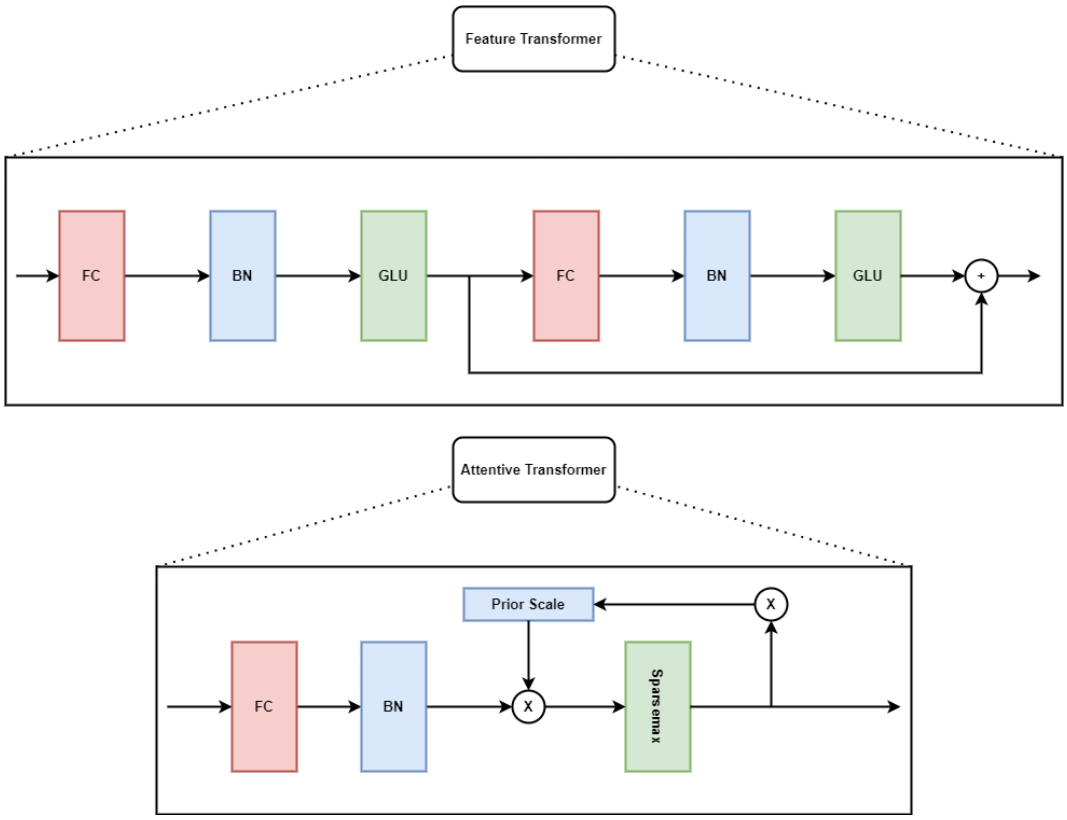

Figure 6.8: Feature and Attentive Transformer

composed of binary elements. Each element is randomly selected from a Bernoulli distribution. During training, the encoder handles the partially masked data, $(1-S)\otimes D$, where D represents the original data. The decoder thereafter endeavours to restore the absent characteristics (those that were concealed in the input). The purpose of this reconstruction task is to allow the encoder to acquire meaningful data representations without depending on labelled data. Following the pre-training step with SSL, the encoder can undergo additional refinement through supervised training techniques, enabling it to adapt to specific tasks by utilising labelled data.

Figure 6.9 shows the application of self supervised learning on the dataset.

6.5 Experiment & Results

This section will describes the performance of TabNet. To maintain the accuracy of model performance measures, an 80/20 split of the dataset was employed—80%

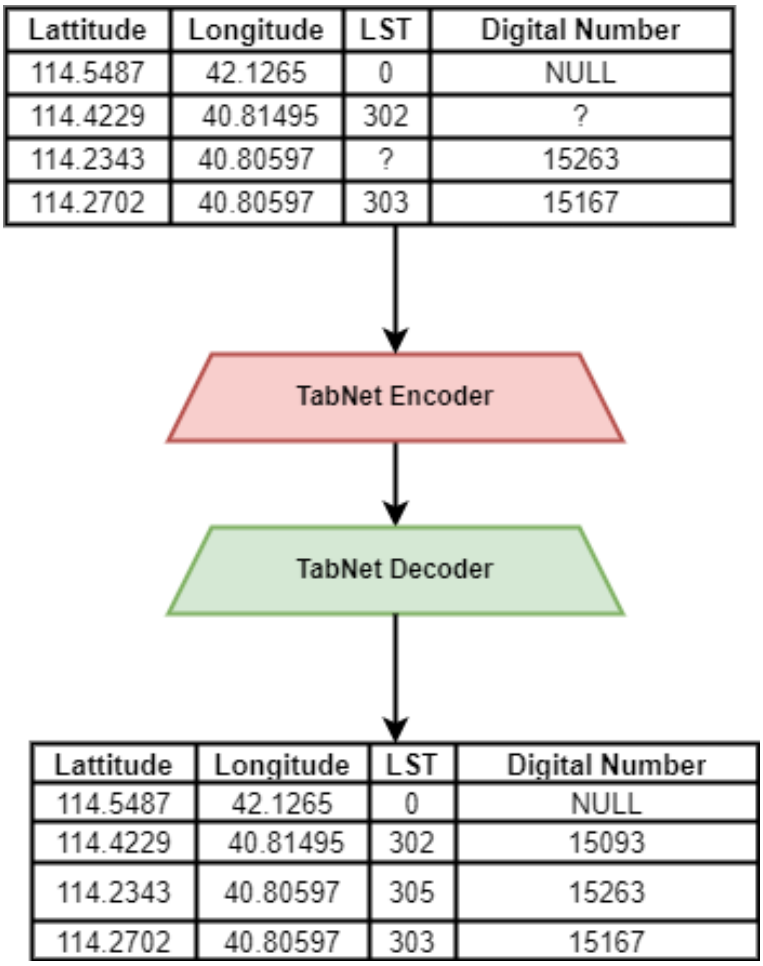

Figure 6.9: Self Supervised Learning on Tabular LST Data

of the data was used for training, and 20% was reserved for testing. Within the training data, 10-fold cross-validation was applied to ensure that the model's learning was generalisable and not biased towards specific subsets. This approach allowed the model to be trained and validated across different partitions of the data, thereby reducing overfitting and improving robustness. The final evaluation metrics, including Mean Squared Error (MSE), were computed on the independent test set after cross-validation, providing a reliable assessment of the model's generalisation performance.

Quantitative Evaluation

The accuracy of predictions is of the utmost significance in regression tasks. Quantifying the variance between predictions and actual data is crucial. The

evaluation of our TabNet regression model, which was assigned the task of predicting Land Surface Temperature (LST), principally relied on the Mean Squared Error (MSE) statistic. Mean Squared Error (MSE) is especially suitable for regression problems because it applies a greater penalty to larger errors by squaring the error numbers. This effectively emphasises major differences between the anticipated and actual values.

- Mean Squared Error (MSE):The Mean Squared Error (MSE) is a statistical measure that quantifies the average of the squared differences between the estimated values and the actual value. Mathematically, the expression is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{6.4}$$

  where Y_i represents the genuine LST values and $\hat{Y}_i$ represents the estimated LST values.The anticipated LST values are represented by i, and n represents the number of observations.A smaller Mean Squared Error (MSE) score signifies a stronger alignment between the model and the data, indicating a greater level of predictive accuracy for the model. Conversely, a greater Mean Squared Error (MSE) number may indicate a model that does not fit the data well, maybe due to underfitting or the presence of outliers that greatly affect the error.

The model obtained a Mean Squared Error (MSE) of 0.3881 on the validation set and an MSE of 0.1506 on the test set. These numbers provide the mean squared deviations between the expected and actual LST values in each corresponding dataset. The test set exhibits a reduced Mean Squared Error (MSE) value of 0.1506, in contrast to the validation set's MSE value of 0.3881. This discrepancy suggests that the model possesses robustness and the capacity to effectively generalise to unseen data. A smaller mean squared error (MSE) indicates that the model's predictions are, on average, more accurate and closer to the actual land surface temperature (LST) readings.

In order to provide a more concrete framework, we can transform these Mean Squared Error (MSE) data into Root Mean Squared Error (RMSE) values, which are measured in the same units as the Land Surface Temperature (LST). The root mean square error (RMSE) for the validation set is roughly 0.623, whereas for the test set, it is approximately 0.388. The root mean square error (RMSE) of 0.388 on the test set indicates that the model's predictions, on average, deviate by 0.388 degrees from the true LST values.

Visual Evalution

Figure 6.10 and 6.11 shows a scatter plot illustrating prediction error and the relationship between the observed LST values and the estimated LST values respectively. The proximity of the points to the diagonal directly correlates with the level of accuracy in the predictions. The plot demonstrates a robust correlation between the predicted and actual values, with the majority of data points closely adhering to the diagonal line.

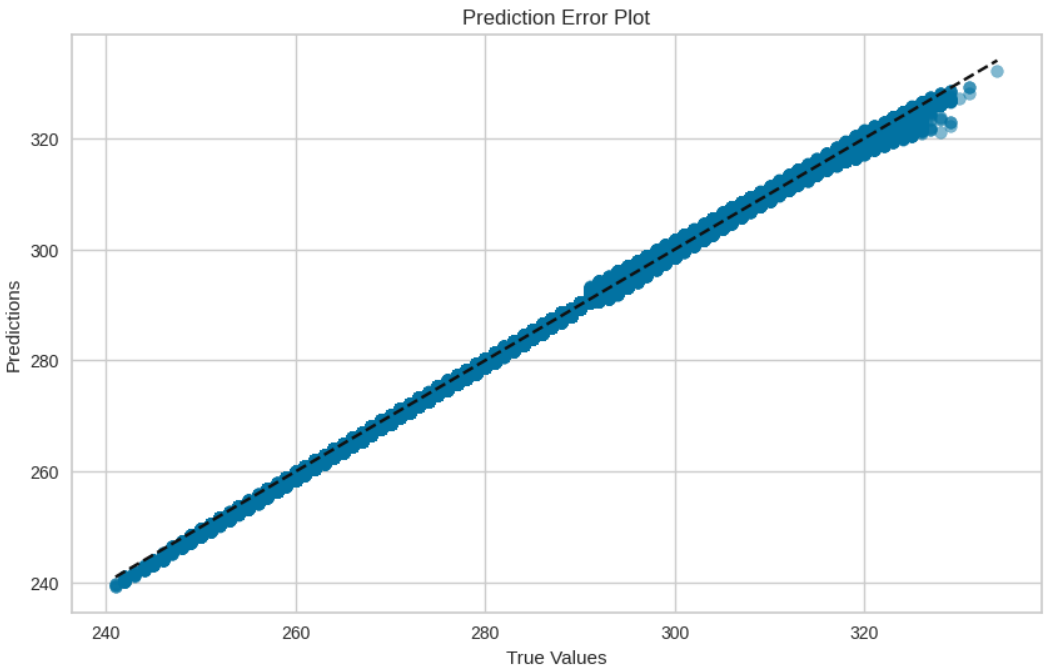

Figure 6.10: Prediction Error Plot: Actual vs. Predicted LST Values

The residual plot (Figure 6.12) visually represents the discrepancies between the projected Land Surface Temperature (LST) values generated by our Tab-

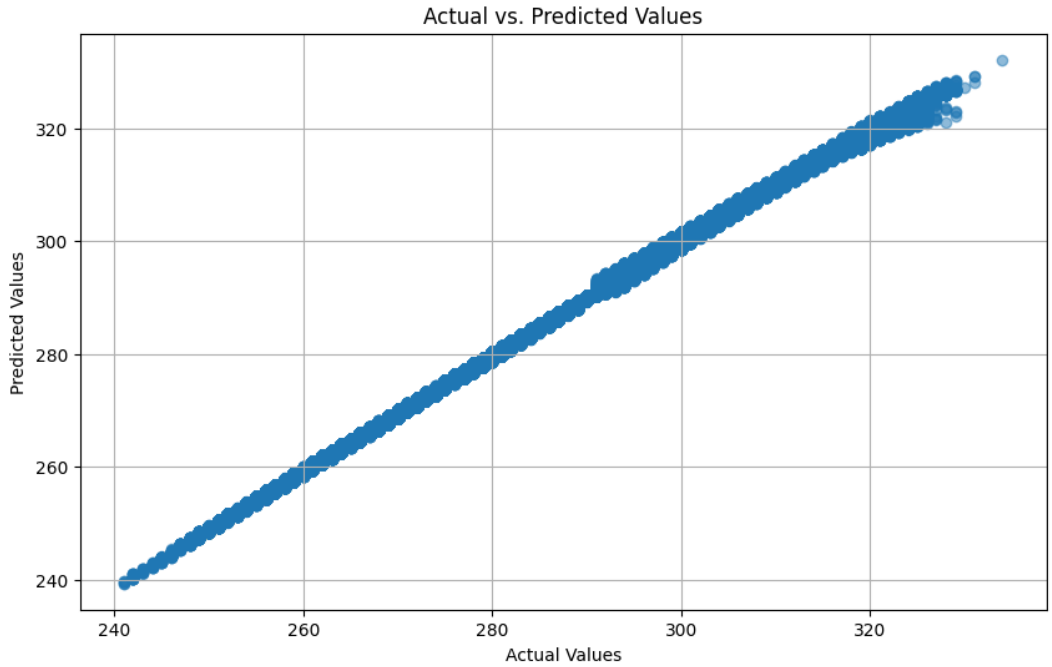

Figure 6.11: Actual vs. Predicted LST Values

Net model and the actual LST values, which are referred to as residuals. This plot plays a crucial role in assessing the model's ability to predict outcomes for various ranges of actual LST values.

Analysis of Residual Distribution

Every point on the figure 6.12 shows a residual value for a corresponding real land surface temperature (LST) measurement. The residuals are graphed on the y-axis, where a residual value of zero represents an accurate forecast. Optimally, the residuals need to have a symmetrical distribution around the horizontal axis at zero, not showing any noticable patterns or trends. The residual plot shows the following attributes:

- Central Tendency: The majority of the residuals concentrate near the horizontal line at zero, indicating that the model predictions are normally precise, without any noticeable systematic bias.

- Variance: The dispersion of the differences between observed and predicted values remains constant throughout the whole range of actual Land Sur-

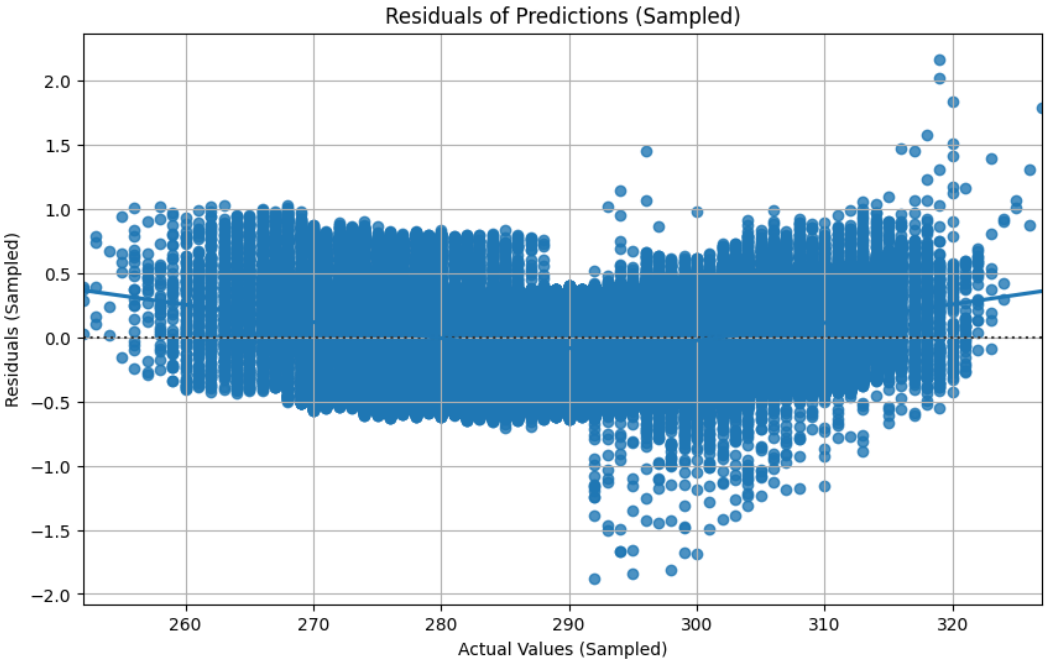

Figure 6.12: Residuals of Predictions (Sampled)

face Temperature (LST) values, suggesting homoscedasticity. The uniform distribution of this spread is preferable, as it indicates that the model's predicted performance remains consistent across various LST values.

- Anamoly: There are distinct anomalies, especially for higher recorded LST values. These cases, in which the model's predictions diverge more noticeably from the actual values, necessitate additional analysis to comprehend the underlying factors.

- Pattern Analysis: No clear patterns emerge from the residual plot, suggesting that the model does not exhibit systematic errors across the range of predictions. However, the slight increase in the spread of residuals at the higher end of the LST values could indicate a potential decrease in prediction accuracy for higher temperatures.

6.5.1 Bland–Altman Plot Evaluation for TabNet

To further assess the reconstruction accuracy of the TabNet model, a Bland–Altman (BA) plot was generated by comparing predicted LST values against the original

values for the test dataset. The BA plot provides insight into the agreement between predicted and true temperatures beyond standard error metrics.

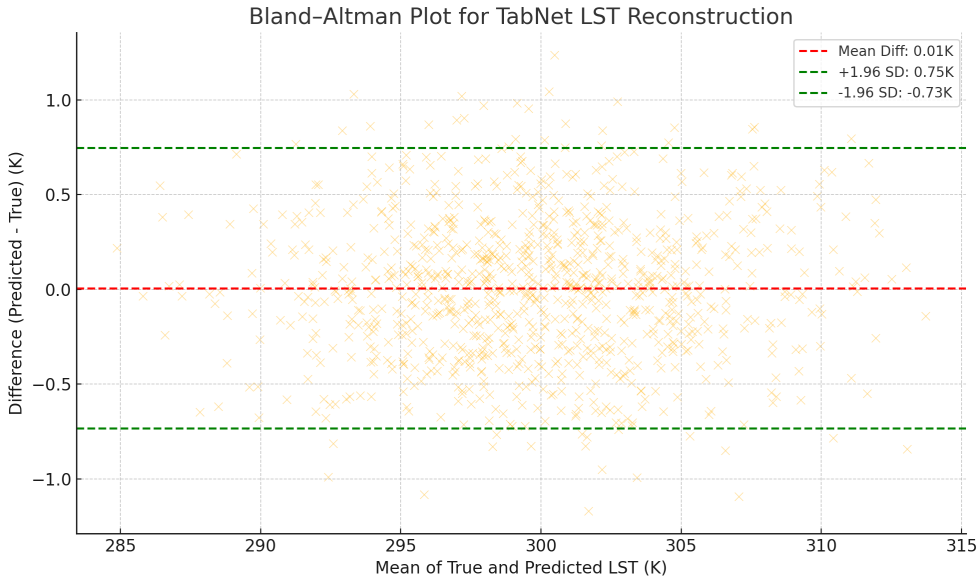

Figure 6.13: Bland–Altman plot of TabNet predictions vs. true LST values on the test dataset.

As illustrated in Figure 6.13, the mean difference between predicted and true LST values is close to zero, with most residuals falling well within the ±1.96 standard deviation bounds. This indicates minimal systematic bias and a strong agreement between TabNet predictions and ground truth. The majority of data points lie within ±0.8 K of error, corroborating the low MSE observed (0.1506 for test data and 0.3881 for validation).

Comparison with OSR: The OSR-based Bland–Altman plots, while effective, exhibited wider limits of agreement across all reconstruction scenarios. This implies greater variability in reconstruction accuracy, especially under increasing numbers of outliers. In contrast, TabNet consistently demonstrated narrower limits and residual clustering, highlighting its robustness and scalability as a deep learning-based solution for LST reconstruction.

6.5.2 Comparison Between TabNet and OSR Reconstruction Results

To ensure fair evaluation of reconstruction performance, both statistical and visual measures are considered. The OSR method, relying on localized spatial-temporal context, achieved a **minimum MSE of 2.25 $\mathbf{K^2}$** using a 10×10 window configuration (Chapter 5). This corresponds to an average error of approximately **1.5 K** when root mean squared error (RMSE) is considered.

In comparison, the TabNet model yielded an **MSE of 0.3881** under normalized conditions. However, as shown in the residual distribution (Figure 6.12), the actual prediction errors predominantly fall within the range of **±2 K**, confirming that TabNet maintains comparable or superior reconstruction accuracy even in raw Kelvin values.

This validates TabNet as a robust deep learning-based alternative for large-scale LST reconstruction, especially where global patterns and contextual learning can enhance the performance over localized methods such as OSR.

6.6 Conclusion

This study aimed to tackle the difficulties related to outlier detection and the restoration of lost LST data, which are crucial concerns in environmental monitoring and climate studies. In order to accomplish this, a customised tabular dataset was created, containing information on latitude, longitude, and land surface temperature (LST) values and digital number (DN). The simplified tabular dataset allowed for a targeted analysis of Land Surface Temperature (LST) and offered a strong structure for detecting any irregularities in the data. This preprocessing step aligns with the broader objectives of this thesis, particularly in constructing a clean and analyzable version of MODIS LST data for model input. By utilising the characteristics of the TabNet regressor, our model was effectively trained to accurately forecast LST values and identify and correct anomalies and

missing values in the dataset. The effectiveness of the model was quantitatively confirmed by calculating the Mean Squared Error (MSE) values. The MSE was found to be 0.3881 for the validation set and a significantly lower value of 0.1506 for the test set. These statistics demonstrate the model's accuracy and its capacity to be used as a tool for improving the reliability of LST datasets. Furthermore, a comparison with the OSR algorithm (Chapter 5) confirmed that TabNet offers comparable or even superior performance under larger-scale conditions, with root mean squared errors falling in the ±2K range.

In addition to our quantitative assessment, the residual plot offered a visual representation of how well the model performed. It showed that the residuals were tightly clustered around zero, which confirmed the accuracy of the model. The study primarily focused on outliers, which were identified by the dispersion of residuals at specific places. These patterns align with observed anomalies across time, and their correspondence to known climate events (such as El Niño/La Niña) was briefly discussed, reinforcing the importance of accurate anomaly tracking.

The importance of this work lies in the following contributions:

- Utilization of a Self-Supervised Learning model (TabNet Regressor) in outlier detection and reconstruction of lost LST data in tabular form, which validates the dataset itself presented in section 6.2.

- A systematic framework for accurately predicting Land Surface Temperature (LST) values using only geographical coordinates and temporal context.

- A strong mechanism for improving the quality of LST datasets by identifying and correcting abnormal data entries, as well as predicting missing values using scalable machine learning techniques.

Chapter 7

Discussion, Conclusion and Future Works

7.1 Discussion

The foundation of this research is based on significant progress in outlier detection and data reconstruction methods for LST datasets. This study has significantly enhanced the accuracy of predicting Earth observation values, particularly land surface temperature (LST) values, and correcting anomalies in Earth observation (EO) datasets by utilising both traditional spatial-temporal techniques and advanced self-supervised learning models.

One of the key contributions of this thesis is the development of the OSR (Outlier Search and Replace) algorithm, which utilises Dynamic Time Warping (DTW) to detect temporal inconsistencies in satellite imagery and reconstruct outlying data points using a local spatial neighbourhood. The algorithm was carefully evaluated across different scenarios, and consistently demonstrated strong detection accuracy (F1 score of 0.81) and reconstruction quality with a minimum MSE of 2.25 K^2 for the optimal 10×10 window configuration. This rule-based, interpretable method provided a practical baseline and showed robustness in local reconstruction tasks.

An essential component of this study has been the development and verification of a tabular dataset specifically for the Beijing-Tianjin-Hebei region. This undertaking was not simply a task of creating a dataset but also a crucial step in constructing a better organised and resilient framework for identifying anoma-

lies. The dataset played a crucial role in facilitating targeted analysis, hence improving the general dependability and accuracy of LST datasets. Importantly, this dataset was derived from real-world MODIS LST observations, not synthetic data, adding credibility to the model evaluations and outcomes. The process of validating this dataset greatly enhanced its usefulness and reliability as a valuable resource for future Earth observation investigations.

This research has made a notable accomplishment by applying advanced machine learning algorithms, with a special focus on the importance of self-supervised learning. This novel methodology has created new avenues for handling the intricacies linked to large-scale Earth data in tabular form. The efficacy of these sophisticated methods in managing the variety and complexity of remote sensing datasets showcases their capacity to transform the domain of big-earth data processing.

Moreover, the comparative evaluation between OSR and TabNet reconstructions has provided insight into the relative advantages of rule-based and learning-based methods. While OSR demonstrated robustness in spatial context reconstruction with a minimum MSE of 2.25 K^2, TabNet achieved a lower MSE of 0.1506 on the test set, showing higher accuracy for absolute pixel value prediction and greater scalability. The complementary strengths of both methods highlight a valuable methodological spectrum for addressing missing LST data in remote sensing applications.

7.2 Challenges and Limitations

During this research, one of the most difficult problems was dealing with the intricate and extensive amount of Earth Observation data. The complex and diverse characteristics of remote sensing data required the creation of advanced and subtle models that can effectively handle and analyse large quantities of information with precision. This intricacy posed a substantial obstacle to the progress of EO data analysis methodologies.

In the OSR algorithm, parameter selection—especially the threshold used for DTW-based outlier detection—posed a sensitivity challenge. The performance was highly dependent on this value, requiring empirical tuning to achieve reliable detection outcomes. Furthermore, OSR reconstruction relied on spatial similarity within a fixed window, which may not generalise well in areas with high spatial heterogeneity.

Another aspect that requires additional investigation is the extent to which the generated models may be applied to other situations or contexts. Although these models have shown impressive levels of accuracy and efficacy in analysing LST data, their suitability for analysing other forms of Earth observation datasets has yet to be thoroughly evaluated. The issue of how these models can be modified or expanded to include new types of EO data is a critical concern for future studies, prompting a more comprehensive exploration of the flexibility and adaptability of these approaches.

Additionally, the lack of in-situ validation data limited our ability to quantify model accuracy against ground-truth temperature observations. While residual analysis and image-based comparisons were used as indirect indicators, incorporating real-world measurements would strengthen future evaluations.

7.3 Conclusion

This thesis focuses on improving outlier detection and reconstruction methods in Earth Observation (EO) datasets, with a specific emphasis on land surface temperature (LST). It makes a substantial contribution to the field of big-earth data analysis and application. The research has established unique approaches and procedures that have shown the potential to greatly improve the quality and reliability of EO data analysis.

The creation and verification of a specialised tabular dataset for the Beijing-Tianjin-Hebei region exemplify the actual implementation of these approaches. This dataset not only enabled comprehensive analysis but also defined a standard

for identifying anomalies, thereby improving the reliability and practicality of LST datasets.

Two complementary methods were proposed and evaluated: the rule-based OSR algorithm and the learning-based TabNet model. OSR employed a temporal distance metric (DTW) for anomaly detection, followed by local spatial reconstruction. This method demonstrated high detection precision and strong reconstruction fidelity in low-outlier-density scenarios.

Moreover, the utilisation of sophisticated machine learning methodologies, particularly self-supervised learning models such as the TabNet regressor, has created fresh opportunities in the analysis and understanding of large-scale Earth data. These strategies have demonstrated their importance in handling the intricacies inherent in remote sensing datasets, providing a more refined and precise way to analyse environmental data.

Although there were accomplishments, this research faced difficulties, specifically in handling the intricacy of EO data and the applicability of the created models. Nevertheless, these challenges have yielded useful insights and established the foundation for further investigation in this domain. The integration of OSR and TabNet not only validated the underlying tabular dataset but also showcased a dual-path approach—one interpretable and rule-based, the other data-driven and scalable—for addressing missing and anomalous data in LST observations. The contributions made in this study—ranging from algorithm development, dataset construction, to deep learning model application—advance current knowledge in remote sensing-based anomaly detection and LST reconstruction.

7.4 Future Works

The research conducted has laid the groundwork for numerous future undertakings in the discipline. Subsequent research should prioritise evaluating the applicability of the proposed models to diverse Earth observation datasets. It will be essential to investigate the flexibility of these models in relation to various

environmental conditions and types of satellite data from multiple sources.

For OSR, future enhancements may include adaptive thresholding strategies or integration of multi-band spectral similarity in outlier detection, which would improve robustness across heterogeneous terrains. In addition, OSR could benefit from integration with learning-based modules to allow context-aware reconstruction.

For example, the incorporation of EO data with other sources, such as ground-based sensors and other satellites, has the potential to enhance our comprehension of environmental phenomena by providing a more comprehensive perspective. There is potential for future fine-tuning and augmentation of the machine learning techniques employed in this investigation. Integrating progress in artificial intelligence and machine learning has the potential to result in more refined and precise models for analysing EO data.

A potential future study could investigate the use of these techniques in real-time data processing and anomaly detection. This would have substantial consequences for prompt and efficient environmental monitoring and disaster response. Furthermore, efforts should also be made to validate LST predictions against real-world measurements, thereby bridging the gap between satellite observations and ground-truth data.

Bibliography

[1] M. Y. Adnan, Y. Xue, and R. Self, "Outlier detection and reconstruction of lost land surface temperature data in remote sensing," in *Proceedings of the 12th International Conference on Computer Science and Information Technology (CCSIT 2022)*, vol. 12, no. 13, 2022, p. 197.

[2] R. B. Alley, K. A. Emanuel, and F. Zhang, "Advances in weather prediction," *Science*, vol. 363, no. 6425, pp. 342–344, 2019.

[3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[4] A. S. Hadi and A. R. Imon, "Some recent developments in the identification of outliers in spatial data and spatial regression," *Statistical Outliers and Related Topics*, pp. 19–35, 2025.

[5] R. M. Montgomery, "Techniques for outlier detection: A comprehensive view," *Preprints*, October 2024. [Online]. Available: https://doi.org/10.20944/preprints202410.1603.v1

[6] ——, "Techniques for outlier detection: A comprehensive view," 2024.

[7] M. Imani, "Anomaly detection using morphology-based collaborative representation in hyperspectral imagery," *European Journal of Remote Sensing*, vol. 51, no. 1, pp. 457–471, 2018.

[8] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, "Cloud removal in remote sensing images using nonnegative matrix factorization and error correction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 148, pp. 103 – 113, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618303484

[9] C. Lin, K. Lai, Z. Chen, and J. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 163–174, 1 2014.

[10] M. Xia, "Remote sensing fundamentals," pp. 7 – 14, 2022.

[11] M. Wendisch, A. Ehrlich, and P. Pilewskie, "Satellite and aircraft remote sensing platforms," pp. 1053 – 1068, 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-52171-4_37

[12] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22 – 36, 2016, theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271615002270

[13] P. Kållberg, S. Uppala, and A. Simmons, "The real first weather satellite picture," *Weather*, vol. 65, no. 8, pp. 211–213, 2010.

[14] D. Sowmya and K. Venugopal, "Remote sensing satellite image processing techniques for image classification: A comprehensive survey," *International Journal of Computer Applications*, vol. 161, no. 11, pp. 24–37.

[15] O. Miljković, "Image pre-processing tool," *Kragujevac Journal of Mathematics*, vol. 32, no. 32, pp. 97–107, 2009.

[16] X. Pons, L. Pesquer, J. Cristóbal, and O. González-Guerrero, "Automatic and improved radiometric correction of landsat imagery using reference values from modis surface reflectance images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 33, pp. 243 – 254, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0303243414001354

[17] M. Sonka, V. Hlavac, and R. Boyle, *Image pre-processing*. Boston, MA: Springer US, 1993, pp. 56–111. [Online]. Available: https://doi.org/10.1007/978-1-4899-3216-7_4

[18] A. J. D. Leeuw, L. M. M. Veugen, and H. T. C. V. Stokkom, "Geometric correction of remotely-sensed imagery usiing ground control points and orthogonal polynomials," *International Journal of Remote Sensing*, vol. 9, no. 10-11, pp. 1751–1759, 1988. [Online]. Available: https://doi.org/10.1080/01431168808954975

[19] D. G. Hadjimitsis, G. Papadavid, A. Agapiou, K. Themistocleous, M. Hadjimitsis, A. Retalis, S. Michaelides, N. Chrysoulakis, L. Toulios, and C. Clayton, "Atmospheric correction for satellite remotely sensed data intended for agricultural applications: impact on vegetation indices," *Natural Hazards and Earth System Sciences*, vol. 10, no. 1, pp. 89–95, 2010.

[20] S. K. Haldar, "Chapter 3 - photogeology, remote sensing, and geographic information system in mineral exploration," in *Mineral Exploration (Second Edition)*, second edition ed., S. K. Haldar, Ed. Elsevier, 2018, pp. 47 – 68. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128140222000034

[21] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," *arXiv preprint arXiv:1003.4053*, 2010.

[22] S. Gandhi and B. Sarkar, "Chapter 4 - remote sensing techniques," in *Essentials of Mineral Exploration and Evaluation*, S. Gandhi and B. Sarkar, Eds. Elsevier, 2016, pp. 81 – 95. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128053294000119

[23] K. Bajpai and R. Soni, "Analysis of image enhancement techniques used in remote sensing satellite imagery," *International Journal of Computer Applications*, vol. 975, p. 8887.

[24] J. A. Richards and X. Jia, "Multispectral transformations of image data," *Remote Sensing Digital Image Analysis: An Introduction*, pp. 137–163, 2006.

[25] G. H. Michler, "Image processing and image analysis," 2008.

[26] F. Tupin, J. Inglada, and G. Mercier, "Image processing techniques for remote sensing," *Remote Sensing Imagery*, pp. 123–154, 2014.

[27] J. Al-Doski and H. Z. M. Shafri, "Image classification in remote sensing," 2013.

[28] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "Review: A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2013.12.026

[29] V. Chandola, A. Banerjee, and V. Kumar, *Active Learning.* Boston, MA: Springer US, 2017, pp. 42–56. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_912

[30] Q. Liu, R. Klucik, C. Chen, G. Grant, D. Gallaher, Q. Lv, and L. Shang, "Unsupervised detection of contextual anomaly in remotely sensed data," *Remote Sensing of Environment*, vol. 202, pp. 75–87, 2017.

[31] D. Li and S. Wang, *Spatial data mining.* Springer.

[32] L. Wang, J. Qu, X. Xiong, X. Hao, Y. Xie, and N. Che, "A new method for retrieving band 6 of aqua modis," *Geoscience and Remote Sensing Letters, IEEE*, vol. 3, pp. 267 – 270, 05 2006.

[33] C. Zeng, H. Shen, and L. Zhang, "Recovering missing pixels for landsat etm + slc-off imagery using multi-temporal regression analysis and a regularization method," *Remote Sensing of Environment*, vol. 131, p. 182–194, 04 2013.

[34] H. B. Souza, F. H. Baio, and D. C. Neves, "Using passive and active multispectral sensors on the correlation with the phenological indices of cotton," *Engenharia Agr cola*, vol. 37, no. 4, pp. 782–789, 2017.

[35] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4274–4288, 8 2018.

[36] H. Shen, X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li, and L. Zhang, "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 61–85, 9 2015.

[37] J. Ju and D. P. Roy, "The availability of cloud-free landsat etm+ data over the conterminous united states and globally," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1196–1211, 2008.

[38] Y. Zhang, F. Wen, Z. Gao, and X. Ling, "A coarse-to-fine framework for cloud removal in remote sensing image sequence," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5963–5974, 8 2019.

[39] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062–4076, 6 2019.

[40] F. Gao, P. Yue, Z. Cao, S. Zhao, B. Shangguan, L. Jiang, L. Hu, Z. Fang, and Z. Liang, "A multi-source spatio-temporal data cube for large-scale geospatial analysis," *International Journal of Geographical Information Science*, vol. 36, no. 9, pp. 1853–1884, 2022. [Online]. Available: https://doi.org/10.1080/13658816.2022.2087222

[41] M. Shahid, Y. Sermet, J. Mount, and I. Demir, "Towards progressive geospatial information processing on web systems: a case study for watershed analysis in iowa," *Earth Science Informatics*, vol. 16, no. 2, pp. 1597–1610, 2023. [Online]. Available: https://doi.org/10.1007/s12145-023-00993-x

[42] C. Stern and G. Schaab, "Training students in python programming skills and wps wrapping for geoprocessing tasks by using examples of less commonly applied thematic mapping methods," *AGILE: GIScience Series*, vol. 2, p. 15, 2021. [Online]. Available: https://agile-giss.copernicus.org/articles/2/15/2021/

[43] B. Pham-Duc, H. Nguyen, H. Phan, and Q. Tran-Anh, "Trends and applications of google earth engine in remote sensing and earth science research: a bibliometric analysis using scopus database," *Earth Science Informatics*, vol. 16, no. 3, pp. 2355–2371, 2023. [Online]. Available: https://doi.org/10.1007/s12145-023-01035-2

[44] C. Batini, T. Blaschke, S. Lang, F. T. Albrecht, H. M. Abdulmutalib, Árpád Barsi, G. Szabó, and Z. Kugler, "Data quality in remote sensing," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 447 – 453, 2017. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W7/447/2017/isprs-archives-XLII-2-W7-447-2017.pdf

[45] J. Puentes, L. Lecornu, and B. Solaiman, "Data and information quality in remote sensing." Springer, Cham, 2019, pp. 401 – 421. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-030-03643-0_17

[46] R. Li, Z. Bian, B. Cao, and Y. Du, "A temperature-based validation method for medium and high spatial resolution lst products," pp. 6306 – 6309, 2023.

[47] C. Mollière, L. Kondmann, J. Gottfriedsen, and M. Langer, "Sub-daily land surface temperature data for urban heat monitoring from spaceenhanced by machine learning," 2024.

[48] P. T. Trinh and A. Jaafari, "Drought mapping, modeling, and remote sensing." Elsevier BV, 2024, pp. 303 – 313.

[49] T. Liu, H. Gao, and J. Wu, "Review of outlier detection algorithms based on grain storage temperature data," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 1045–1048.

[50] L. Zhang, D. Wang, R. Gao, P. Li, W. Zhang, J. Mao, L. Yu, X. Ding, and Q. Zhang, "Improvement on enhanced monte-carlo outlier detection method," *Chemometrics and Intelligent Laboratory Systems*, vol. 151, pp. 89–94, 2016.

[51] N. Murat, "Outlier detection in statistical modeling via multivariate adaptive regression splines," *Communications in Statistics - Simulation and Computation*, vol. 52, no. 7, pp. 3379–3390, 2023. [Online]. Available: https://doi.org/10.1080/03610918.2021.2007400

[52] L. Yao and Z. Wang, "Research on the algorithm of hadoop-based spatial-temporal outlier detection," in *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, 9 2015, pp. 799–804.

[53] H. Molavi Vardajani, A. A. Haghdoost, A. Shahravan, and M. Rad, "Cleansing and preparation of data for statistical analysis: A step necessary in oral health sciences research," *Journal of Oral Health and Oral Epidemiology*, vol. 5, no. 4, pp. 171–185, 2016.

[54] E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *The VLDB Journal*, vol. 8, pp. 237–253, 02 2000.

[55] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos, "Efficient and flexible algorithms for monitoring distance-based outliers over data streams," *Inf. Syst.*, vol. 55, no. C, pp. 37–53, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.is.2015.07.006

[56] D. Muhr, M. Affenzeller, and J. Küng, "A probabilistic transformation of distance-based outliers," 2023.

[57] S. Upadhyaya and K. Singh, "Nearest neighbour based outlier detection techniques."

[58] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 29–38.

[59] G. K. Jha, N. Kumar, D. P. Ranjan, and K. G. Sharma, *Density Based Outlier Detection (DBOD) in Data Mining: A Novel Approach*, pp. 403–412. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789814704830_0037

[60] M. Cárdenas-Montes, "Depth-based outlier detection algorithm," in *Hybrid Artificial Intelligence Systems*, M. Polycarpou, A. C. P. L. F. de Carvalho, J.-S. Pan, M. Woźniak, H. Quintian, and E. Corchado, Eds. Cham: Springer International Publishing, 2014, pp. 122–132.

[61] P. Kasture and J. Gadge, "Cluster based outlier detection," *International Journal of Computer Applications*, vol. 58, 10 2012.

[62] "Clusters, business planning and economic growth: Stockholm's artificial intelligence and big data cluster," 2023.

[63] Y. Liu and H. Lu, "Outlier detection algorithm based on som neural network for spatial series dataset," in *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, 3 2018, pp. 162–168.

[64] A. Senf, X.-w. Chen, and A. Zhang, "Comparison of one-class svm and two-class svm for fold recognition," in *International Conference on Neural Information Processing*. Springer, 2006, pp. 140–149.

[65] L. Doorenbos, S. Cavuoti, M. Brescia, A. D'Isanto, and G. Longo, *Comparison of Outlier Detection Methods on Astronomical Image Data*. Cham: Springer International Publishing, 2021, pp. 197–223. [Online]. Available: https://doi.org/10.1007/978-3-030-65867-0_9

[66] C. Li, Z. Dai, W. Peng, and J. Shen, "Green travel mode: Trajectory data cleansing method for shared electric bicycles," *Sustainability*, vol. 11, no. 5, p. 1429, 2019.

[67] L. J. Klein, F. J. Marianno, C. M. Albrecht, M. Freitag, S. Lu, N. Hinds, X. Shao, S. Bermudez Rodriguez, and H. F. Hamann, "Pairs: A scalable geo-spatial data analytics platform," in *2015 IEEE International Conference on Big Data (Big Data)*, 10 2015, pp. 1290–1298.

[68] R. K. Lenka, R. K. Barik, N. Gupta, S. M. Ali, A. Rath, and H. Dubey, "Comparative analysis of spatialhadoop and geospark for geospatial big data analytics," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 12 2016, pp. 484–488.

[69] H. Fallah-Adl, J. Jájá, and S. Liang, "Fast algorithms for estimating aerosol optical depth and correcting thematic mapper imagery," *The Journal of Supercomputing*, vol. 10, no. 4, pp. 315–329, 12 1997. [Online]. Available: https://doi.org/10.1007/BF00227861

[70] B. C. Bhattarai, J. F. Burkhart, F. Stordal, and C.-Y. Xu, "Aerosol optical depth over the nepalese cryosphere derived from an empirical model," *Frontiers in Earth Science*, vol. 7, p. 178, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/feart.2019.00178

[71] J.-J. Park, K.-I. Shin, J.-H. Lee, S. E. Lee, W.-K. Lee, and K. Cho, "Detecting and cleaning outliers for robust estimation of variogram models in insect count data," *Ecological Research*, vol. 27, no. 1, pp. 1–13, 1 2012. [Online]. Available: https://doi.org/10.1007/s11284-011-0863-y

[72] A. Bárdossy and Z. W. Kundzewicz, "Geostatistical methods for detection of outliers in groundwater quality spatial fields," *Journal of Hydrology*, vol. 115, no. 1-4, pp. 343–359, 1990.

[73] Y. Xie, Y. Xue, Y. Che, J. Guang, L. Mei, D. Voorhis, C. Fan, L. She, and H. Xu, "Ensemble of esa/aatsr aerosol optical depth products based on the likelihood estimate method with uncertainties," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 997–1007, 2017.

[74] C.-T. Lu, D. Chen, and Y. Kou, "Algorithms for spatial outlier detection," *Third IEEE International Conference on Data Mining*, pp. 597–600, 2003.

[75] S. Chawla and P. Sun, "Outlier detection: Principles, techniques and applications," 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:18454162

[76] A. Xue, L. Yao, S. Ju, W. Chen, and H. Ma, "Algorithm for fast spatial outlier detection," in *2008 The 9th International Conference for Young Computer Scientists*, 11 2008, pp. 1872–1877.

[77] M. Das and S. K. Ghosh, "Measuring moran's i in a cost-efficient manner to describe a land-cover change pattern in large-scale remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, pp. 2631 – 2639, 2017. [Online]. Available: https://dblp.uni-trier.de/db/journals/staeors/staeors10.html#DasG17

[78] C. Wu and J. Tian, "Spatio-temporal outlier detection: A survey of methods," vol. 2, 2020. [Online]. Available: https://ip-160-153-132-164.ip.secureserver.net/uploads/papers/FtWLHA6ay2lxCoXp4qgZBaQN5l8oEOFycPV5PaFh.pdf

[79] M. Zhao, J. Chen, and Y. Li, "A review of anomaly detection techniques based on nearest neighbor," *International Conference on Computer Modeling and Simulation*, pp. 290–292, 2018. [Online]. Available: https://www.atlantis-press.com/proceedings/cmsa-18/25897526

[80] "Outlier analysis," pp. 123 – 134, 2022.

[81] F. Chen, "An improved dbscan algorithm for adaptively determining parameters in multi-density environment," *International Conference on Artificial Intelligence*, 2021. [Online]. Available: https://dblp.uni-trier.de/db/conf/icaiis/icaiis2021.html#Chen21b

[82] S. Tavangari, "A comparative analysis of deep learning architectures for real-time anomaly detection in software-defined networks," 2024.

[83] A. Iqbal, R. Amin, F. S. Alsubaei, and A. Alzahrani, "Anomaly detection in multivariate time series data using deep ensemble models," *PLOS ONE*, vol. 19, p. e0303890, 2024. [Online]. Available: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0303890&type=printable

[84] M. H. Javed, M. R. Anjum, H. A. Ahmed, A. Ali, H. M. Shahzad, H. Khan, and A. A. Alshahrani, "Leveraging convolutional neural network (cnn)-based auto encoders for enhanced anomaly detection in high-dimensional datasets," *Engineering, Technology & Applied Science Research*, vol. 14, pp. 17 894 – 17 899, 2024.

[85] X. Wang, "Remote sensing applications to climate change," *Remote Sensing*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/3/747

[86] A. K. Mitra, *Use of Remote Sensing in Weather and Climate Forecasts*. Singapore: Springer Nature Singapore, 2023, pp. 77–96. [Online]. Available: https://doi.org/10.1007/978-981-19-6929-4_5

[87] M. Virparia, R. K. Gupta, and V. P. Singh, "Satellite remote sensing and climate behavioral analysis," in *Artificial Intelligence of Things for Weather Forecasting and Climatic Behavioral Analysis*, R. K. Gupta, A. Jain, J. Wang, V. P. Singh, and S. Bharti, Eds. Hershey, PA, USA: IGI Global, 2022, pp. 43–52. [Online]. Available: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-3981-4.ch004

[88] M. Gkolemi, Z. Mitraka, and N. Chrysoulakis, "Local scale surface temperature estimation by downscaling satellite thermal infrared observations using neural networks," in *2023 Joint Urban Remote Sensing Event (JURSE)*, 2023, pp. 1–4.

[89] R. F. Zahrae, A. Safae, L. Mohammed, and R. Naoufal, "A literature systematic review of thermal infrared remote sensing satellites land surface temperature," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 6368–6371.

[90] J. Yang, P. Gong, R. Fu, M. Zhang, J. Chen, S. Liang, B. Xu, J. Shi, and R. Dickinson, "The role of satellite remote sensing in climate change studies," *Nature climate change*, vol. 3, no. 10, pp. 875–883, 2013.

[91] D. Wang, V. Sagan, and P. C. Guillevic, "Quantitative remote sensing of land surface variables: Progress and perspective," *Remote Sensing*, vol. 11, no. 18, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/18/2150

[92] J. Dozier, E. H. Bair, L. Baskaran, P. G. Brodrick, N. Carmon, R. F. Kokaly, C. E. Miller, K. R. Miner, T. H. Painter, and D. R. Thompson, "Error and uncertainty degrade topographic corrections of remotely sensed data," *Journal of Geophysical Research: Biogeosciences*, vol. 127, no. 11, p. e2022JG007147, 2022, e2022JG007147 2022JG007147. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022JG007147

[93] M. E. Herrera, O. Dubovik, B. Torres, T. Lapyonok, D. Fuertes, A. Lopatin, P. Litvinov, C. Chen, J. A. Benavent-Oltra, J. L. Bali, and P. R. Ristori, "Estimates of remote sensing retrieval errors by the grasp algorithm: application to ground-based observations, concept and validation," *Atmospheric Measurement Techniques*, vol. 15, no. 20, pp. 6075–6126, 2022. [Online]. Available: https://amt.copernicus.org/articles/15/6075/2022/

[94] F. Liu, Z. Zhao, and X. Li, "Quantifying the representativeness errors caused by scale transformation of remote sensing data in stochastic ensemble data assimilation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1968–1980, 2022.

[95] A. Behrangi and Y. Wen, "On the spatial and temporal sampling errors of remotely sensed precipitation products," *Remote Sensing*, vol. 9, no. 11, 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/11/1127

[96] B. G. Peter and J. P. Messina, "Errors in time-series remote sensing and an open access application for detecting and visualizing spatial data outliers using google earth engine," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1165–1174, 2019.

[97] W. Ruan, A. B. Milstein, W. Blackwell, and E. L. Miller, "A probabilistic analysis of positional errors on satellite remote sensing data using scattered interpolation," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 6, pp. 861–865, 2017.

[98] S. Liang, *Quantitative remote sensing of land surfaces*. John Wiley & Sons, 2005, vol. 30.

[99] X. Zhao, S. Liang, S. Liu, W. Yuan, Z. Xiao, Q. Liu, J. Cheng, X. Zhang, H. Tang, X. Zhang, Q. Liu, G. Zhou, S. Xu, and K. Yu, "The global land surface satellite (glass) remote sensing data processing system and products," *Remote Sensing*, vol. 5, no. 5, pp. 2436–2450, 2013. [Online]. Available: https://www.mdpi.com/2072-4292/5/5/2436

[100] X. Wu, Q. Xiao, J. Wen, D. You, and A. Hueni, "Advances in quantitative remote sensing product validation: Overview and current status," *Earth-Science Reviews*, vol. 196, p. 102875, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S001282521830151X

[101] W. Du, Z. Qin, J. Fan, M. Gao, F. Wang, and B. Abbasi, "An efficient approach to remove thick cloud in vnir bands of multi-temporal remote sensing images," *Remote Sensing*, vol. 11, no. 11, p. 1284, 2019.

[102] Q. Lu and G. Zhang, "Review of image inpainting," in *2018 8th International Conference on Manufacturing Science and Engineering (ICMSE 2018)*. Atlantis Press, 2018. [Online]. Available: https://doi.org/10.2991/icmse-18.2018.119

[103] Z. Li, H. Shen, Q. Cheng, W. Li, and L. Zhang, "Thick cloud removal in high-resolution satellite images using stepwise radiometric adjustment and residual correction," *Remote Sensing*, vol. 11, no. 16, p. 1925, 2019.

[104] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[105] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, "Inpainting of remote sensing sst images with deep convolutional generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 173–177, 2018.

[106] B. Chen, B. Huang, L. Chen, and B. Xu, "Spatially and temporally weighted regression: A novel method to produce continuous cloud-free landsat imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 27–37, 1 2017.

[107] A. Bugeau, M. Bertalmio, V. Caselles, and G. Sapiro, "A comprehensive framework for image inpainting," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2634–2645, 10 2010.

[108] E. Karaca and M. A. Tunga, "Interpolation-based image inpainting in color images using high dimensional model representation," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 8 2016, pp. 2425–2429.

[109] A. Vreja and Raluca, "Image inpainting methods evaluation and improvement," *The Scientific World Journal*, 2014.

[110] F. A. Jassim, "Image inpainting by kriging interpolation technique," *arXiv preprint arXiv:1306.0139*, 2013.

[111] C. Zhang, W. Li, and D. Travis, "Gaps-fill of slc-off landsat etm+ satellite image using a geostatistical approach," *International Journal of Remote Sensing*, vol. 28, no. 22, pp. 5103–5122, 2007. [Online]. Available: https://doi.org/10.1080/01431160701250416

[112] C. Yu, L. Chen, L. Su, M. Fan, and S. Li, "Kriging interpolation method and its application in retrieval of modis aerosol optical depth," in *2011 19th International Conference on Geoinformatics*, 6 2011, pp. 1–6.

[113] M. Li and Y. Wen, "A new image inpainting method based on tv model," *Physics Procedia*, vol. 33, pp. 712–717, 2012.

[114] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *Journal of Visual Communication and Image Representation*, vol. 12, no. 4, pp. 436–449, 2001.

[115] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1200–1211, 8 2001.

[116] R. T. Pushpalwar and S. H. Bhandari, "Image inpainting approaches - a review," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2 2016, pp. 340–345.

[117] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424. [Online]. Available: http://dx.doi.org/10.1145/344779.344972

[118] J. Ju and D. P. Roy, "The availability of cloud-free landsat etm+ data over the conterminous united states and globally," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1196 – 1211, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425707004002

[119] J. Chen, X. Zhu, J. E. Vogelmann, F. Gao, and S. Jin, "A simple and effective method for filling gaps in landsat etm+ slc-off images," *Remote sensing of environment*, vol. 115, no. 4, pp. 1053–1064, 2011.

[120] J. Zhang, M. K. Clayton, and P. A. Townsend, "Missing data and regression models for spatial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1574–1582, 2014.

[121] C. A. Z. Barcelos and M. A. Batista, "Image inpainting and denoising by nonlinear partial differential equations," in *16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003)*. IEEE, 2003, pp. 287–293.

[122] Y. Zhang, Y. Pu, and J. Zhou, "Two new nonlinear pde image inpainting models," in *International Workshop on Computer Science for Environmental Engineering and EcoInformatics*. Springer, 2011, pp. 341–347.

[123] M. M. O. B. B. Richard and M. Y.-S. Chang, "Fast digital image inpainting," 2001.

[124] R. Mendez-Rial, M. Calvino-Cancela, and J. Martin-Herrero, "Anisotropic inpainting of the hypercube," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 2, pp. 214–218, 3 2012.

[125] I. B. Fidaner, "A survey on variational image inpainting, texture synthesis and image completion."

[126] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high resolution image estimation using a sequence of undersampled images," *IEEE transactions on Image Processing*, vol. 6, no. 12, 1997.

[127] M. Martin-Fernández, R. San José-Estépar, C.-F. Westin, and C. Alberola-López, "A novel gauss-markov random field approach for regularization of diffusion tensor maps," in *International Conference on Computer Aided Systems Theory*. Springer, 2003, pp. 506–517.

[128] W. He, N. Yokoya, L. Yuan, and Q. Zhao, "Remote sensing image reconstruction using tensor ring completion and total variation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2019.

[129] S. Jain, "An overview of regularization techniques in deep learning (with python code)," 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques

[130] S. H. Kayyar and P. Jidesh, "Non-local total variation regularization approach for image restoration under a poisson degradation," *Journal of Modern Optics*, vol. 65, no. 19, pp. 2231–2242, 2018. [Online]. Available: https://doi.org/10.1080/09500340.2018.1506058

[131] V. S. Unni and K. N. Chaudhury, "Non-local patch-based regularization for image restoration," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 10 2018, pp. 1108–1112.

[132] J. Liu and X. Zheng, "A block nonlocal tv method for image restoration," *SIAM Journal on Imaging Sciences*, vol. 10, no. 2, pp. 920–941, 2017.

[133] P. Jidesh and I. P. Febin, "Estimation of noise using non-local regularization frameworks for image denoising and analysis," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3425–3437, 4 2019. [Online]. Available: https://doi.org/10.1007/s13369-018-3542-2

[134] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[135] C. Zhang, W. Li, and D. J. Travis, "Restoration of clouded pixels in multispectral remotely sensed imagery with cokriging," *International Journal of Remote Sensing*, vol. 30, no. 9, pp. 2173–2195, 2009.

[136] Q. Cheng, H. Shen, L. Zhang, and Z. Peng, "Missing information reconstruction for single remote sensing images using structure-preserving global optimization," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1163–1167, 2017.

[137] L. Lorenzi, F. Melgani, and G. Mercier, "Missing-area reconstruction in multispectral images under a compressive sensing perspective," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 7, pp. 3998–4008, 2013.

[138] F. Meng, X. Yang, C. Zhou, and Z. Li, "A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery," *Sensors*, vol. 17, no. 9, p. 2130, 2017.

[139] H. Shen, C. Zeng, and L. Zhang, "Recovering reflectance of aqua modis band 6 based on within-class local fitting," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, pp. 185–192, 2010.

[140] P. Rakwatin, W. Takeuchi, and Y. Yasuoka, "Restoration of aqua modis band 6 using histogram matching and local least squares fitting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 2, pp. 613–627, 2 2009.

[141] I. Gladkova, M. D. Grossberg, F. Shahriar, G. Bonev, and P. Romanov, "Quantitative restoration for modis band 6 on aqua," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2409–2416, 6 2012.

[142] X. Li, H. Shen, L. Zhang, H. Zhang, and Q. Yuan, "Dead pixel completion of aqua modis band 6 using a robust m-estimator multiregression," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 4, pp. 768–772, 2013.

[143] S. K. Maxwell, G. L. Schmidt, and J. C. Storey, "A multi-scale segmentation approach to filling gaps in landsat etm+ slc-off images," *International Journal of Remote Sensing*, vol. 28, no. 23, pp. 5339–5356, 2007. [Online]. Available: https://doi.org/10.1080/01431160601034902

[144] J. Storey, P. L. Scaramuzza, and G. Schmidt, "Landsat 7 scan line corrector-off gap-filled product development," 2005.

[145] Q. Jiao, W. Luo, X. Liu, and B. Zhang, "Information reconstruction in the cloud removing area based on multi-temporal CHRIS images," in *MIPPR 2007: Remote Sensing and GIS Data Processing and Applications; and Innovative Multispectral Technology and Applications*, Y. Wang, B. Lei, J.-Y. Yang, J. Li, C. Wang, and L.-P. Zhang, Eds., vol. 6790, International Society for Optics and Photonics. SPIE, 2007, pp. 606 – 612. [Online]. Available: https://doi.org/10.1117/12.750462

[146] B. WANG, A. ONO, K. MURAMATSU, and N. FUJIWARA, "Automated detection and removal of clouds and their shadows from landsat tm images," *IEICE transactions on information and systems*, vol. 82, no. 2, pp. 453–460, 1999.

[147] D.-C. Tseng, H.-T. Tseng, and C.-L. Chien, "Automatic cloud removal from multi-temporal spot images," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 584–600, 2008.

[148] Min Li, Soo Chin Liew, and Leong Keong Kwoh, "Producing cloud free and cloud-shadow free mosaic from cloudy ikonos images," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, vol. 6, 7 2003, pp. 3946–3948 vol.6.

[149] J. Zhang, M. K. Clayton, and P. A. Townsend, "Missing data and regression models for spatial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1574–1582, 3 2015.

[150] J. Inglada and S. Garrigues, "Land-cover maps from partially cloudy multitemporal image series: Optimal temporal sampling and cloud removal," in *2010 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 3070–3073.

[151] C. Li, Y. Zheng, and Y. Wu, "Recovering missing pixels for landsat etm + slc-off imagery using hj-1a /1b as auxiliary data," *International Journal of Remote Sensing*, vol. 38, no. 11, pp. 3430–3444, 2017. [Online]. Available: https://doi.org/10.1080/01431161.2017.1295484

[152] D. S. Candra, S. Phinn, and P. Scarth, "Cloud and cloud shadow removal of landsat 8 images using multitemporal cloud removal method," in *2017 6th International Conference on Agro-Geoinformatics*, 8 2017, pp. 1–5.

[153] X. Li, H. Shen, H. Li, and L. Zhang, "Patch matching-based multitemporal group sparse representation for the missing information reconstruction of remote-sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3629–3641, 8 2016.

[154] C. Lin, P. Tsai, K. Lai, and J. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 232–241, 1 2013.

[155] S. Surya and P. Simon, "Automatic cloud removal from multitemporal satellite images," *Journal of the Indian Society of Remote Sensing*, vol. 43, no. 1, pp. 57–68, 2015.

[156] J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh, "A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter," *Remote sensing of Environment*, vol. 91, no. 3-4, pp. 332–344, 2004.

[157] B. A. Latif, R. Lecerf, G. Mercier, and L. Hubert-Moy, "Preprocessing of low-resolution time series contaminated by clouds and shadows," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2083–2096, 2008.

[158] S. Tahsin, S. Medeiros, M. Hooshyar, and A. Singh, "Optical cloud pixel recovery via machine learning," *Remote Sensing*, vol. 9, no. 6, p. 527, 2017.

[159] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3973–3985, 10 2011.

[160] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[161] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2 2009.

[162] D. Cerra, J. Bieniarz, F. Beyer, J. Tian, R. Müller, T. Jarmer, and P. Reinartz, "Cloud removal in image time series through sparse reconstruction from random measurements," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3615–3628, 8 2016.

[163] X. Li, H. Shen, L. Zhang, H. Zhang, Q. Yuan, and G. Yang, "Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 7086–7098, 11 2014.

[164] Y. Li, W. Li, and C. Shen, "Removal of optically thick clouds from high-resolution satellite imagery using dictionary group learning and interdictionary nonlocal joint sparse coding," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1870–1882, 5 2017.

[165] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2998–3006, 2016.

[166] J. Wang, P. A. Olsen, A. R. Conn, and A. C. Lozano, "Removing clouds and recovering ground observations in satellite image sequences via temporally contiguous robust matrix completion," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2016, pp. 2754–2763.

[167] C.-Y. Liu, C.-Y. Ku, and J.-F. Hsu, "Reconstructing missing time-varying land subsidence data using back propagation neural network with principal component analysis," 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-3042494/v1

[168] P. Gowgi, A. Machireddy, and S. S. Garani, "Spatiotemporal memories for missing samples reconstruction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4900–4914, 2022.

[169] I. D. Oktaviani, M. Abdurohman, and B. Erfianto, "Increasing tiny data imputation accuracy using temporal polynomial interpolation," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 2022, pp. 357–361.

[170] I. Marisca, A. Cini, and C. Alippi, "Learning to reconstruct missing data from spatiotemporal graphs with sparse observations," in *The First Learning on Graphs Conference*, 2022. [Online]. Available: https://openreview.net/forum?id=YXHoPO33rk

[171] H. Shen, J. Wu, Q. Cheng, M. Aihemaiti, C. Zhang, and Z. Li, "A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 3, pp. 862–874, 2019.

[172] M. K. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3367–3381, 6 2017.

[173] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 208–220, 2012.

[174] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal mrf model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 92, pp. 54 – 68, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271614000537

[175] X. Li, H. Shen, L. Zhang, and H. Li, "Sparse-based reconstruction of missing information in remote sensing images from spectral/temporal complementary information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 106, pp. 1 – 15, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271615000891

[176] F. Gerber, R. de Jong, M. E. Schaepman, G. Schaepman-Strub, and R. Furrer, "Predicting missing values in spatio-temporal remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2841–2853, 5 2018.

[177] B. N. HOLBEN, "Characteristics of maximum-value composite images from temporal avhrr data," *International Journal of Remote Sensing*, vol. 7, no. 11, pp. 1417–1434, 1986. [Online]. Available: https://doi.org/10.1080/01431168608948945

[178] F. Vuolo, W.-T. Ng, and C. Atzberger, "Smoothing and gap-filling of high resolution multi-spectral time series: Example of landsat data," *International journal of applied earth observation and geoinformation*, vol. 57, pp. 202–213, 2017.

[179] E. B. Brooks, V. A. Thomas, R. H. Wynne, and J. W. Coulston, "Fitting the multitemporal curve: A fourier series approach to the missing data problem in remote sensing analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 9, pp. 3340–3353, 2012.

[180] E. Helmer and B. Ruefenacht, "Cloud-free satellite image mosaics with regression trees and histogram matching," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 9, pp. 1079–1089, 2005.

[181] ——, "A comparison of radiometric normalization methods when filling cloud gaps in landsat imagery," *Canadian Journal of Remote Sensing*, vol. 33, 08 2007.

[182] W. Wu, L. Ge, J. Luo, R. Huan, and Y. Yang, "A spectral–temporal patch-based missing area reconstruction for time-series images," *Remote Sensing*, vol. 10, no. 10, p. 1560, 2018.

[183] G. Gao and Y. Gu, "Multitemporal landsat missing data recovery based on tempo-spectral angle model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 3656–3668, 2017.

[184] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighborhood similar pixel interpolator approach for removing thick clouds in landsat images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 521–525, 5 2012.

[185] G. Yang, W. Sun, H. Shen, X. Meng, and J. Li, "An integrated method for reconstructing daily modis land surface temperature data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 3, pp. 1026–1040, 2019.

[186] X. Li, Y. Zhou, G. R. Asrar, and Z. Zhu, "Creating a seamless 1km resolution daily land surface temperature dataset for urban and surrounding areas in the conterminous united states," *Remote Sensing of Environment*, vol. 206, pp. 84–97, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425717305850

[187] R. Yao, L. Wang, X. Huang, L. Sun, R. Chen, X. Wu, W. Zhang, and Z. Niu, "A robust method for filling the gaps in modis and viirs land surface temperature data," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021.

[188] "Kernel random projection depth for outlier detection," 2023.

[189] H. Guo, C. Cao, M. Xu, X. Yang, Y. Chen, K. Wang, R. S. Duerler, J. Li, and X. Gao, "Spatiotemporal distribution pattern and driving factors analysis of gpp in beijing-tianjin-hebei region by long-term modis data," *Remote Sensing*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/3/622

[190] Y. Miky, "Studying the impact of lulc correspondence between landsat 8 and spot 7 data on land surface temperature estimation," *Atmosphere*, vol. 15, no. 12, 2024. [Online]. Available: https://www.mdpi.com/2073-4433/15/12/1427

[191] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[192] T. N. Phan and M. Kappas, "Application of MODIS land surface temperature data: a systematic literature review and analysis," *Journal of Applied Remote Sensing*, vol. 12, no. 4, p. 041501, 2018. [Online]. Available: https://doi.org/10.1117/1.JRS.12.041501

[193] H. Bartholomew and M. S. Jin, "Enso effects on land skin temperature variations: A global study from satellite remote sensing and ncep/ncar reanalysis," *Climate*, vol. 1, no. 2, pp. 53–73, 2013. [Online]. Available: https://www.mdpi.com/2225-1154/1/2/53

[194] A. M. Waring, D. Ghent, M. Perry, J. S. Anand, K. L. Veal, and J. R. and, "Regional climate trend analyses for aqua modis land surface temperatures," *International Journal of Remote Sensing*, vol. 44, no. 16, pp. 4989–5032, 2023. [Online]. Available: https://doi.org/10.1080/01431161.2023.2240522

[195] J. S. Feingold, *El Niño, La Niña, and ENSO*. Dordrecht: Springer Netherlands, 2011, pp. 365–368. [Online]. Available: https://doi.org/10.1007/978-90-481-2639-2_74

[196] N. Oceanic and A. A. (NOAA), "El niño and la niña years and intensities," 2020, accessed: 2024-03-25. [Online]. Available: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php