Deutsches Zentrum für Luft- und Raumfahrt
German Aerospace Center

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fraunhofer
IEM

Master Thesis

Computer Science

Development of a model for analysis of census data using statistical methods and machine learning

by:

Megha Aditya
Matr.-Nr. 6882557

Supervisors:

Prof. Dr.-Ing. Roman Dumitrescu
Dr.-Ing. Sebastian von Enzberg
M.Sc. Luis Blanco
M.Sc. Ruslan Bernijazov
M.Sc. Denis Tissen

Paderborn, June 29, 2023

Master Thesis Proposal

Development of a model for analysis of Census Data using Statistical Methods and Machine Learning

Problem statement

This thesis aims to use disaggregation methods on the German Census 2011 to assign specific characteristics to individual buildings in Germany. Three specific datasets of the German Census 2011 will be processed; these are the households, buildings and population datasets.

Due to data protection laws, the individual data points (information about each specific building or person) in the census data are subject to statistical confidentiality, and therefore, the data is published in an aggregated grid - format.

This thesis proposes to develop a model that assigns special characteristics to each building based on its geographical location using machine learning techniques.

Main tasks

The solution approach of this thesis will follow the listed tasks in the specified order:

1. For the entire country, the creation of a database by merging data from the census housing, families, and population.
2. For Oldenburg city, creation of a database by spatially merging data from the previously generated census database and 3D buildings models of the city.
3. Use QGIS (Quantum Geographic Information System) to categorize the number of occupants in the building by integrating census with geographical data. QGIS is used for performing geographical and spatial data analysis.
4. Research and implement machine learning algorithms to classify the number of occupants in each building in the grid cell.

By:

Megha Aditya
Computer Science Master’s student
Universität Paderborn
Email:maditya@mail.uni-paderborn.de
Phone: +49 176 69572565

Master Thesis Nr. 187

Development of a model for analysis of census data using statistical methods and machine learning

on: June 29, 2023

HEINZ NIXDORF INSTITUT
Fachgruppe Advanced Systems Engineering
Prof. Dr.-Ing. Roman Dumitrescu
Fürstenallee 11
D-33102 Paderborn

Deutsches Zentrum für Luft- und Raumfahrt e.V (German Aerospace Center)
Institute for Solar Research
Group of Buildings and Districts
Karl-Heinz-Beckurts-Str. 13, 52428 Juelich

Non-Disclosure Notice

This work remains closed to the public due to confidential data and information.

Megha Aditya

Megha Aditya

Statutory declaration:

I hereby declare that I have produced the present work independently and without unauthorized assistance from others, that I have used no sources and aids other than those indicated, and that I have marked the passages taken from the sources used, either literally or in terms of content, as such.

Megha Aditya

Paderborn, June 29, 2023

Acknowledgement

I would like to express my sincere gratitude to Luis Blanco for his invaluable support and guidance as my thesis advisor. His unwavering patience, extensive knowledge, and constructive feedback have been instrumental in my research journey, enabling me to delve deeper into the topic and expand my technical expertise. I would like to express my gratitude to Ruslan Bernijazov and Denis Tissen for their supervisory role and the valuable feedback offered. Their guidance and inputs have played a crucial role in my growth and learning throughout this process. I am thankful for their contributions, which have greatly contributed to my development during this endeavour. I am incredibly fortunate to have been given the opportunity to undertake this thesis at DLR (German Aerospace Center). I extend my deepest appreciation to Prof. Dr.-Ing. Roman Dumitrescu for granting me this opportunity and placing trust in me to explore this subject matter. I am also grateful to Dr.-Ing. Sebastian von Enzberg for kindly accepting my request to supervise this thesis. I would like to extend my thanks to my family and friends who have provided unwavering support and encouragement during this journey. Lastly, I would like to dedicate this work to my parents, their love and support have been my greatest motivation to strive for excellence and overcome challenges. Their presence and belief in me have been invaluable.

Abstract

In the current energy crisis faced by Germany, building an energy map holds great importance for effective energy management and decision-making. This thesis contributes relevant information for constructing an energy map by classifying the number of residents per building. Building population estimation holds importance to identify the energy trends and patterns of building occupants which plays a significant role in managing energy needs and formulating appropriate energy management strategies.

The study utilizes two primary datasets, namely 3D building models for the particular study area and the 2011 German census data. Due to data privacy policies, information about individual building occupants is unavailable and this particular information is important to understand not just the final energy demand per building or household but also understand energy consumption behaviour patterns. This thesis addresses this challenge by developing a specialized and comprehensive databank that integrates the building models with census data on population, housing, and buildings. Additionally, to classify the building's number of residents, a supervised machine learning approach using the XGBoost algorithm was employed. In the absence of training datasets due to data laws, a synthetic training dataset was prepared by incorporating various building types such as single-family houses, multi-family houses, and other types. Three separate models were constructed, and their results were subsequently combined to enhance accuracy.

The evaluation of accuracy per grid cell provides insights into the effectiveness of the classification process. The developed classification model showed an average accuracy of 62%, where the models are closely related to the building form classification. The findings of this research have implications for better energy management and decision-making in the face of the country's energy crisis.

Table of Contents

Page

List of Figures

List of Tables

1 Introduction

The building sector is a key area for curbing the consumption trend and reducing energy-related emissions. In Germany, the energy consumption of the building sector accounts for about 40% of the total energy demand and contributes to almost one-third of the country's greenhouse gas emission [ED20-ol]. Building-related energy emissions include energy used for space heating, water heating, cooling, lighting and ventilation. Building energy-efficient structures can bring energy usage down as explained in [GEV+21]. Therefore, there is a pressing need to build energy-efficient buildings.

Based on the recent reports published by the International Energy Agency (IEA) [ED20-ol], [Age20], renovations of existing buildings for energy demand management can reduce energy consumption by 50-70%. Additionally, the energy management changes induced in buildings can lead to other benefits such as air quality control, reduced operating costs and indoor comforts.

The primary step to reduce building energy consumption is to measure the total energy demand by using building energy assessment methods. These methods include the development of data-driven methodologies and energy models to help estimate building energy requirements and incorporate energy-saving strategies. Data analysis and machine learning (ML) methods are great tools to evaluate the energy consumption and efficiency of buildings.

Figure 1-1 describes the trend of increased use of data-driven methodologies in building energy consumption predictions [LLL+23]. The colour scheme of Figure 1-1 has been adjusted to maintain colour consistency throughout this study. The studies highlighted in Figure 1-1 have utilised data-driven methodologies for scenarios such as the prediction of energy consumption, fault detection and optimization of energy systems in buildings. All these research studies have investigated the use of data-driven methods in reducing building energy consumption and building greenhouse gas emissions.

However, building energy models cannot be made without high-quality data about the building stock, including among others, dimensions, construction period, building types or refurbishment status. This thesis attempts to investigate how ML methods can be used to build a high-quality database of the building stock of our study area by analyzing the census 2011 database [Zen23-ol] which contains relevant information about building properties for the whole of Germany.

In Germany, residential and non-residential building data is acquired through various methods, such as dwelling registers, income statistics, housing surveys, sensor data, and, most importantly, during the national population and housing census. The national census is the procedure of systematically acquiring and recording population information about the members of a given population. However, due to data protection laws in Germany, information about individual citizens or individual buildings cannot be of open access and therefore, these datasets are altered and published in an aggregated manner. Thus, the data available for research lacks resolution and completeness. An absence of these two properties in the different datasets could lead to flawed outputs.

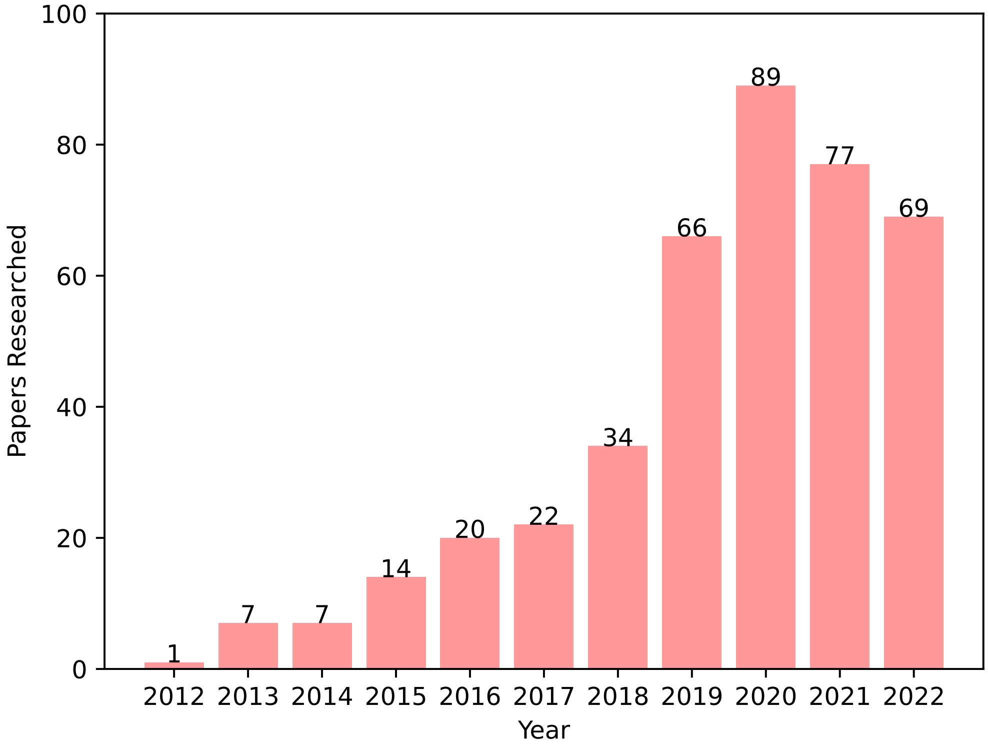


Figure 1-1: Research Statistics Highlighting the use of data-driven approaches in Buildings Energy Consumption Prediction [LLL+23]

.

Data availability differs across the country and between regions due to non-uniform administrative procedures and privacy concerns, this presents a problem since the German building stock is not complete, high-resulted, or available for the whole country. The available data differs from region to region, or it is published in an aggregate form like in the census data [Zen23-ol].

Depending on the scale of the simulation, using incomplete or aggregated data to model energy consumption estimates adds varying degrees of uncertainty, with bigger uncertainties for the simulation of single buildings. In this context, the study aims to bridge the gap between the available data and the requirements of building energy models using machine learning methods. This thesis will focus on one specific parameter of the census 2011 dataset, namely: the number of residents per building. The main objective of the thesis will be to classify individual buildings according to their number of residents. This process includes data filtering, disaggregation, modelling and evaluation of the results.

1.1 Motivation

With the energy crisis enveloping the country, the focus today is to efficiently manage and reduce the energy consumption of buildings. This thesis contributes to the bigger goal of the German Aerospace Center (DLR, in German: Deutsches Zentrum für Luft- und Raumfahrt) to develop an energy map of the country's building stock. This energy map would be used to identify building energy demand and be a source of information for the necessary renovation strategies of engineering offices and energy planers.

Building energy maps can also be used to optimize the energy efficiency and sustainability of new buildings, ensure compliance with the country's energy standards and reduce expenses during the construction and operation of buildings. The building energy demand depends on many factors. The number of residents in each building is a crucial factor to determine the energy needs of a building. If the number of residents in a given building is high, it will lead to higher consumption of heating and electricity. Additionally, having knowledge of the number of residents per building also identifies opportunities for energy conservation and efficient energy management [TD11], [SG17], [ED20-ol], [Sus12]. Therefore, there is a need to find the number of residents in each building. However, due to data protection laws in Germany, this data is restricted and made publicly available in an aggregated grid format. This thesis work attempts to find the number of residents per building by categorizing using ML algorithms for the city of Oldenburg, which is located in the state of Niedersachsen, Germany and represents the study area of this thesis work.

1.2 Objectives

The recent energy crisis engulfing the country has made it imperative to focus on energy management. The building sector holds importance in energy management as explained in section 1. In order to support the energy map development in the future (not in the scope of this thesis). The study involves the creation of a database by merging data from the census housing, families, and population of Germany. This databank will be used for energy analysis and energy map creation of buildings. The focus of this study is Oldenburg and in order to make any classifications on the buildings of the city, a database is built by merging data from the census housing, families, and population with geographical data of 3D building models. This databank will serve as a main data point for classifications done during this study. To investigate the attributes of the datasets of the census and buildings 3D models and expose hidden relationships between them we perform exploratory data analysis on the dataset. Finally, to classify the number of number of occupants in each building of the grid the study involves research and implementation of a suitable machine learning algorithm.

Summarizing the objectives:

1) For the entire country, the creation of a database by merging data from the census housing, families, and population

2) For Oldenburg city, creation of a database by spatially merging data from the previously generated census database and 3D buildings models of the city.

3) Research and implement ML algorithms to classify the number of occupants in each building in the grid cell

1.3 Approach

This study addresses the need to categorize the number of residents in each building on the grid. The data for this task is generated by creating a databank by merging the census datasets of housing, buildings and population along with geographical data of the study area. For better visualization of geospatial and census data, QGIS is used. The goal is to categorize the number of occupants in an apartment and assign unique characteristics to each building in the grid.

By following the above-mentioned tasks, our study ensures to answer the research questions and achieve the main objectives. A general database of the 3D building models of the study area with a number of residents being assigned to each building by disaggregating the census data using ML techniques.

1.4 Thesis Structure

In order to answer all objectives as discussed in section 1.2, this thesis is structured as follows. Chapter 2 discusses the problem definition tackled in this study and defines the requirements of the thesis. It also discusses the Census 2011 and Buildings 3D models datasets which are the primary datasets used in this work and highlights the inconsistencies of these datasets. Chapter 3 presents the literature overview of this study. This chapter highlights the research work done in the field of building energy consumption predictions using data-driven methodologies. Chapter 4 outlines the fundamental concepts and terminologies used in this study. The backbones of this study and the various approaches and methods used are discussed. Chapter 5 discusses the outcomes after applying the methodology are examined, discussed and analysed. Chapter 6 concludes this study with a summary and outlines the possible areas of future work.

2 Problem Analysis

This thesis aims to analyze two main datasets of the German Census 2011 database [Off23-olb], [Zen23-ol], the housing and the population datasets. The housing census provides information about the properties of buildings such as building age, construction type, and heating system and a detailed list can be seen in appendix table 7-2. The population census, on the other hand, offers data on the number of inhabitants in the country.

Due to data protection laws in Germany, the individual data points (meaning, information about each specific building or person) in the census data are subject to statistical confidentiality, and the public data is therefore published in an aggregated grid format. The census results are published in the INSPIRE-compliant $100\,\mathrm{m} \times 100\,\mathrm{m}$ grid cell format [Off23-olb] specified by the INSPIRE guideline [CG16-ol]. The INSPIRE guideline uses the Lambert-Azimuthal-Equal Area projection (ETRS89- LAEA Europe - EPSG:3035).

Bridging the gap between available data and the needs of building energy models has resulted in a variety of efforts towards the classification of different properties for residential buildings such as building age and construction type.

This thesis proposes to develop a model that assigns special characteristics to each building based on its geographical location using machine learning techniques. As part of the future use cases, the developed model will be used to forecast the energy needs of each household.

2.1 Census 2011 Data Acquisition Process

Prior to 2011, the last published census was in 1987. Every person in the German population was required to fill out a paper-based questionnaire and submit it to the enumerator. This collection of filled questionnaires is made up of the census and the attributes of the census data. It was legally required to take part in the census. Since the traditional census was not based on the survey, the only source of errors were systematic errors and no random sampling errors was observed [SK16].

United Nations recommends hosting a census every 10 years [SK16]. There was a need for the calculation of demographic indicators such as age distribution, gender distribution and many more aspects of the population distribution. Besides affecting the social and economic characteristics of a population, these indicators also impact policies and business strategies.

According to the European Union requirement, it is mandatory for every member state to host a census in 10-year intervals from 2011. To maintain uniformity and avoid any inconsistencies in the census data, all member states must submit the census results for a specified range of variables [KG19]. However, there are no restrictions for the methodology of census collection opted by the member states. The census model is explained in Figure 2-1, which highlights the data collection, data preparation and integration processes.

In Germany, unlike the past approaches used for complete enumeration, the census 2011

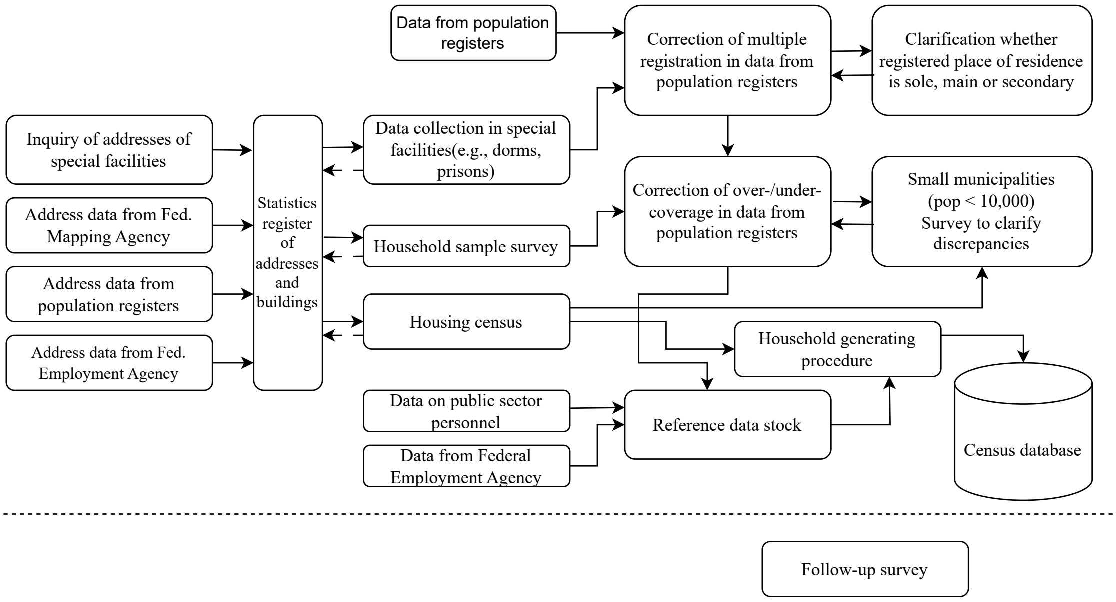

Data from population registers
Correction of multiple registration in data from population registers
Clarification whether registered place of residence is sole, main or secondary
Inquiry of addresses of special facilities
Address data from Fed. Mapping Agency
Address data from population registers
Address data from Fed. Employment Agency
Statistics register of addresses and buildings
Data collection in special facilities(e.g., dorms, prisons)
Household sample survey
Housing census
Correction of over-/under-coverage in data from population registers
Small municipalities (pop < 10,000) Survey to clarify discrepancies
Household generating procedure
Data on public sector personnel
Data from Federal Employment Agency
Reference data stock
Census database
Follow-up survey

Figure 2-1: The main data sources and process in census 2011, adapted from [Ste17].

was a register-based census where data from administrative registers was the primary data source followed by random sample surveys.

By German law, it is mandated for people to be registered when residing in the country. The population records are maintained in every municipality. These population registers are the main data sources of the census 2011 [KG19]. The population data from municipalities are then sent to the Land statistical offices and from there to the Federal Statistical Office before being integrated and submitted in a safe and secure area of statistics.

The information from population registers transmitted details of each individual's name, gender, citizenship (s), marital status, and date of birth. It also contains information about the spouse, children or any registered partner which was necessary for generating household information. Another significant feature of the population registers was indicating whether the individual is affiliated with any religious group or society under state law.

Another goal of the census 2011, was to estimate the occupation of the German population and labour force participation. The Federal Employment Agency produced data on the working class of the population contributing to social insurance. In addition to the information of name, gender, and age the economic background details are described along with the employer details and start date of the term. It was also a source for information on job seekers and the population of people recipients of unemployment benefits [Off23-olb].

The information of public officials who do not contribute to social insurance is maintained by public employers. The public employers provided data on people working in public domains such as judges, police officers, soldiers etc.

2.1.1 Understanding the Limitations of Administrative Data

The administrative data collected from various sources as discussed in the previous section is useful. However, this data is subjected to inaccuracies and inconsistencies lowering the reliability of this type of data source. For instance, the details on individuals' education attainment or job role titles and living conditions may include inconsistencies. This is primarily because every data collection process serves a specific purpose and the in-depth details gathered may not suffice other types of analysis processes. Furthermore, there may be inaccuracies due to the way the data was collected such as self-reported information or missing real-time data and old data is mentioned in the registers.

In regards to the self-employed individuals, there is no register data maintained. Similarly, there exists no register data about buildings and dwellings throughout the entire country.

2.1.2 Surveys & Census Results

Supplementary surveys were conducted to eradicate any inaccuracies induced by the data collected from administrative registers [Off23-olb]. There were three types of surveys conducted to fill in the information gaps,

1) **Buildings and housing survey**: This survey turned out to be the country's largest survey of the census 2011. The primary reason is the unavailability of any form of reliable data on housing and dwellings in the country. A complete enumeration was obtained through a postal survey to be filled by all building owners, and building managers which accounted for 17.5 million records [Off23-olc]. The result of this survey included information on the type of building (residential, industrial etc), number of rooms, number of storeys, type of heating system used and many other features.

2) **Household Survey**: the data obtained by the household survey is not available in any of the register data thereby making the survey result most important for determining the number of inhabitants in municipalities with more than ten thousand individuals [Off23-old].

   Population registers are unreliable as they are not updated, people change residences so the number of inhabitants cannot be based entirely on population registers. By conducting household surveys, outdated information can be eradicated and missing info can be filled in and real-time data can be identified.

   This survey was conducted not on complete enumeration but only for a sample population of about 10% of the total population of the country. The individuals in this sample size were selected and interviewed using a random sample method. The collected data was then extrapolated to the total population. This survey concluded with a final round of follow-up surveys where 5% of the households where the first round of household survey took place were selected randomly and questioned again. This follow-up survey only took place for municipalities with more than ten thousand occupants.

   The data obtained in this survey included the schooling certificates, vocational training, and migrant background information of individuals. This collected data was also used to fill in the gaps left by the data collected from the Federal employment agency.

3) **Survey in residential establishments**: The data collected from this survey was used to correct any statistical errors maintained in the population register data. The residential establishments and collective living quarters were divided into two groups namely, sensitive and non-sensitive facilities. Non-sensitive facilities included dormitories, boarding schools, and care homes and all individual residents were interviewed.

   However, sensitive facilities comprised of places where the exposure of individual information may create social discrimination against them. These places included prison homes, shelter homes, and psychiatric facilities. The information of occupants across sensitive facilities was provided by the manager or supervisor of the facility through face-to-face interviews or online questionnaires [Off23-ole].

   The decentralization of population register data across the country stimulates multiple registration errors where an individual is registered in more than one municipal population register data. Such errors were corrected by interviewing such individuals and thus corrected information was obtained [Off23-ola].

2.2 Census 2011 Results

The census results were published on [Zen23-ol]. Due to data privacy policies nationwide, individual data points are publicly unavailable. The census results are published in an aggregated grid format.

Table 2-1: Attributes of a grid cell in INSPIRE format

Grid Cell Attribute	Description
id	Identifier for a grid cell
x_sw	X-Coordinate of the southwest corner
y_sw	Y-Coordinate of the southwest corner
x_mp	X-coordinate of the centre of the cell
y_mp	Y-coordinate of the centre of the cell
f_staat	State area of Germany in the grid cell
f_land	Land area of Germany in grid cell
f_wasser	Water area of Germany in grid cell
p_staat	Share of Germany in the area of grid cell
p_land	Share of the German land area in the grid cell
p_wasser	Share of the German water area in the grid cell
ags	Municipality key

The use of grid cells for population representation is a widely used approach in statistics primarily because employing the use of a grid enables the distribution of the population counts to smaller units. Grids are uniformly sized squared or rectangles covering the area of interest. This leads to uniform distribution of information independent of administrative boundaries [Eur23-ol] [CG16-ol]. The uniform and equal distribution of grid sizes enable a better understanding of the population distribution over areas of interest.

The census results are published in INSPIRE grid format. INSPIRE is a European initiative for establishing a spatial data infrastructure. The main idea behind INSPIRE format was to enable Europe wide uniform geographical grid representation system. The census results use Equal Area Grids – based on the European Terrestrial Reference System (ETRS89) Lambert Azimuthal Equal Area(LAEA)(EPSG:3035) projection Coordinate Reference System (CRS)([CG16-ol]. The resolution of grid cells is 1m, 10m, 100m, 1km, 10km, and 100km. In this study, we use grid cells of (100 m × 100 m) resolution since data for grids with higher resolution(1m & 10m)is unavailable for research purposes.

A grid cell in INSPIRE format is called a cell code. The attributes of a grid cell are as shown in Table 2-1 [CG16-ol]. In this study, we utilise the grid cell id and x and y coordinate data for analysis with Buildings 3D models.

Figure 2-2 illustrates one of the results of the census 2011, Since we are focused on Buildings and their types across Germany. Other buildings with residential spaces include scenarios where a floor of the building is used for residential purposes while the other part is used for industrial purposes. It is evident from the results that residential buildings occupy the majority of buildings fraction in the country and thereby making residential buildings an important class of study for this thesis.

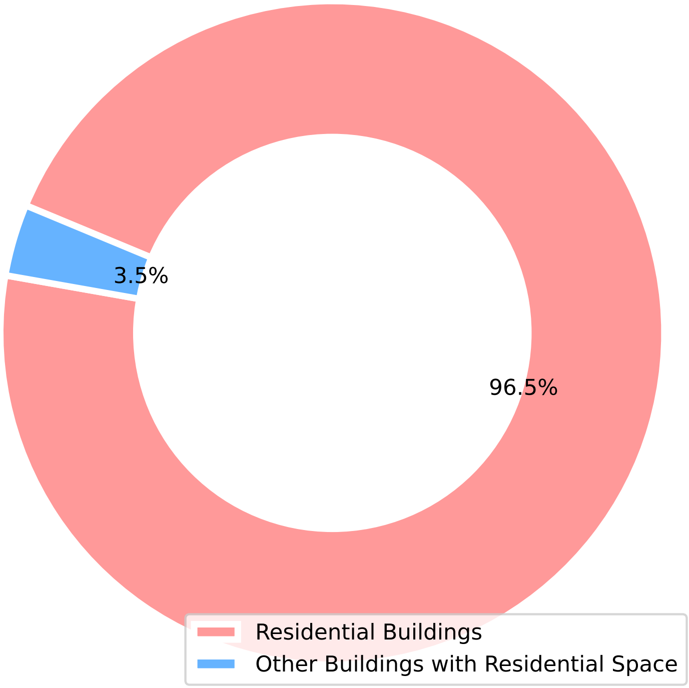


Figure 2-2: Census 2011 Results for Residential Buildings in Germany

2.3 Buildings 3D CityGML Models

Today numerous companies and cities are focused on the development of 3D city models. A wide range of advantages have been identified for building virtual 3D city models including urban planning, adopting disaster management and energy management strategies, mobile telecommunication and tourism. However, topological and semantic considerations were not taken into account when developing 3D models in the past, limiting their utility to purely visual applications. CityGML was designed to enable the use of 3D models across several sectors [Hea21-ol].

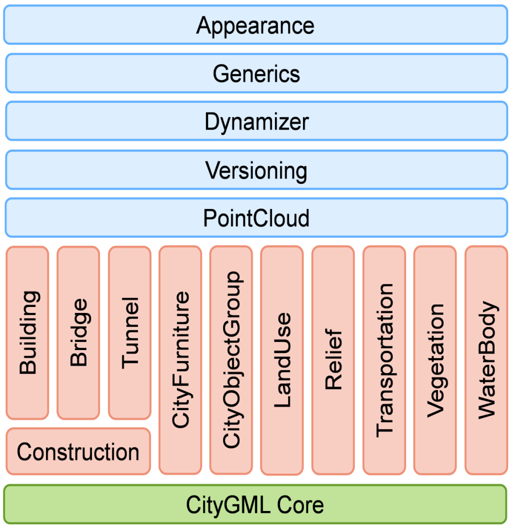


Figure 2-3: CityGML 3.0 module overview. The vertical boxes show the different thematic modules. Horizontal modules depict the concepts that are applicable to all thematic modules [Hea21-ol]

.

CityGML is a common semantic information model developed with the aim of storing and exchanging virtual 3D city models. It is defined through a Unified Modeling Language (UML)-ISO Technical Committee 211 (TC211) conceptual model standards for spatial and temporal data [Hea21-ol].

As shown in figure 2-3, a CityGML model is thematically decomposed into a Core module followed by different thematic modules. The five concept modules are utilized over all the thematic modules. These concepts define the generic characteristics and object relationships, geometry and time-series data linked to sensors for providing the provision of including real-time data in the 3D city models.

The Conceptual model of CityGML utilizes the conceptual schema of ISO 19107 which enables the usage of recursive aggregation of composite objects. For example, a building geometry(Composite Solid) can be composed of a house geometry (Composite Solid) and the garage geometry (Solid), while the house's geometry is further decomposed into the roof geometry (Solid) and the geometry of the house body (Solid) as shown in Figure 2-4 [Hea21-ol].

A CityGML conceptual model defines level-of-detail (LOD) standards to represent building information (LOD0-3). A building information model is a comprehensive digital representation of a built facility [ANM+08]. Figure 2-5 illustrates different representations of the same real-world building object in different LOD levels. LOD1 contains blocks(i.e.

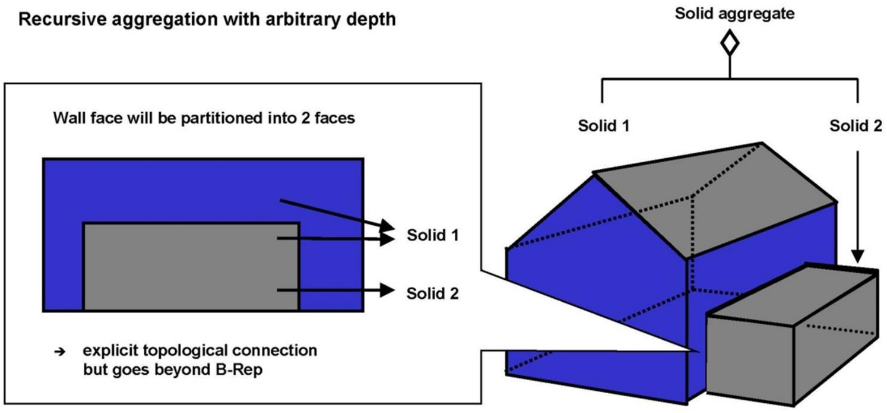

Recursive aggregation with arbitrary depth
Wall face will be partitioned into 2 faces
Solid 1
Solid 2
➔ explicit topological connection but goes beyond B-Rep
Solid aggregate
Solid 1
Solid 2

Figure 2-4: Using Conceptual schema allows recursive aggregation of geometries and objects in CityGML [Hea21-ol]
.

extruded footprints). LOD2 describes volumes with generalized roof shapes. LOD3 specifies volumetric models with greater architectural details, including windows (as well as other openings), roof overhangs, and more details. Finally, LOD4 extends LOD3 with additional indoor features like rooms or furniture.

This study utilises buildings 3D models for Niedersachsen (Oldenburg), which is open-source [GS23-ol]. We have used CityGML-LOD2 models without LOD3 and LOD4 levels, and the LOD1 level is abstract to the level of detail required for this study. The building data are available in .shp (shape file) format and consist of many geometric attributes such as the area of the building and many more. The visualization of shape files is possible on Geographic Information Systems (GIS, explained in detail in chapter 4). In this study, we use Quantum Geographic Information System (QGIS) for building data visualizations.

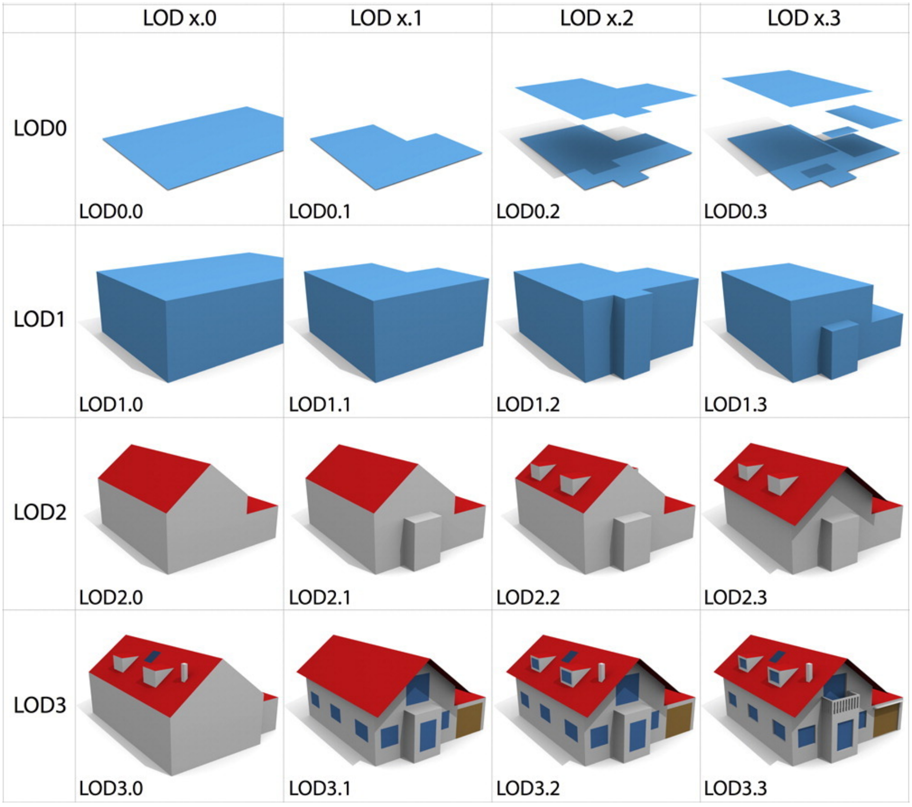


Figure 2-5: The four different LODs for building models in CityGML [BLS16] [Hea21-ol]

.

2.4 Case Study

The study site Oldenburg is located in the Federal State of Lower Saxony in Northwestern Germany as shown in Figure 2-6.

With almost 170,000 inhabitants. There has been radial urban growth in the city over the past decades. The city has a spatial structure that is predominantly centred around a single core, with a historic city centre at its centre.

The urban extensions surrounding the city core have been developed over the past few decades, resulting in a diverse mix of building types within the city's morphology such as single-family houses, multi-family houses and other types. For instance, the central area of the city is characterized by block development, while the suburbs consist primarily of free-standing or residential houses. Additionally, large industrial sites or non-residential buildings are situated at the periphery of the city, occupying substantial areas of land. In total, the study area covers 102.96 square meters [WDS+21].

The primary reason for choosing Oldenburg is the availability of resources and datasets, some of which were provided by local project partners at DLR. The source information about the building models of Oldenburg is available on the data portal of Lower Saxony [GL23-ol].

A total of 56,749 building models were exported for the city of Oldenburg. About 75% (42 875) of them are residential buildings. The other 25% are distributed among industrial, commercial, agricultural and educational buildings. Nevertheless, the database is incom-

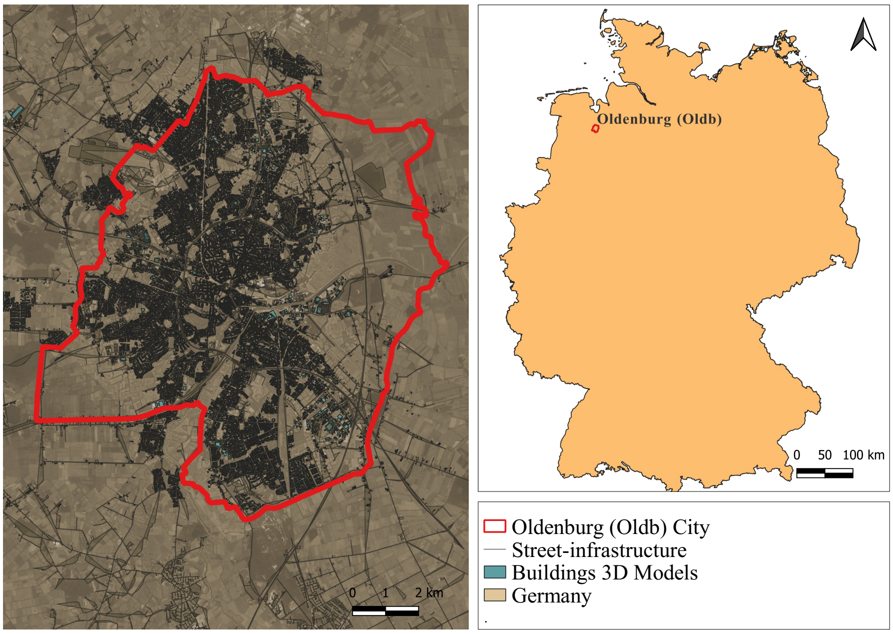


Figure 2-6: Illustration of Oldenburg City, Germany, the study area for this thesis

.

plete, and errors in such classifications are expected, which will be explained in detail in the next section 4.1.

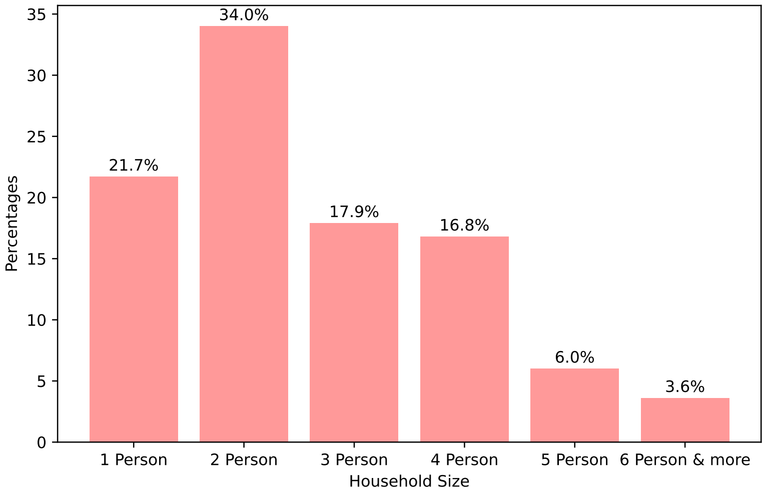


Figure 2-7: Census 2011 Results for the percentages of buildings in Oldenburg, Germany, based on Household Size

Bar plots in figure 2-8 and figure 2-7 are some of the published census results. The results are available at [Zen23-ol] and users can customize based on parameters and regions across Germany.

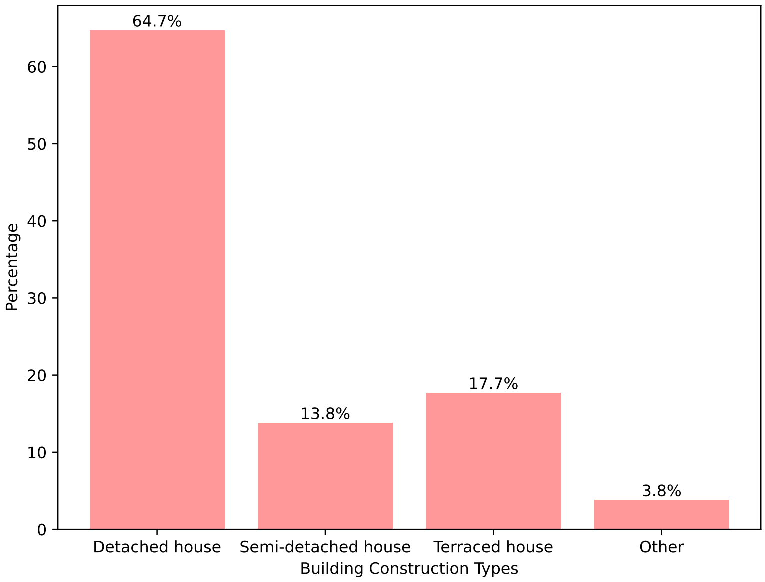


Figure 2-8: Census 2011 Results for the percentages of buildings in Oldenburg, Germany, based on Construction Type

2.5 Derivation of Requirements

This section explores the requirements for this thesis study and provides an overview of how they will be satisfied. The requirements for this study are as follows:

1) **R1 – Unified census data management to access different census datasets at once**: The thesis aims to build a databank integrating three different datasets of the census namely, population, buildings and household census across Germany and define the structure of the database such that each grid cell id is uniquely identifiable.

2) **R2 – Data integration for analysis**: This involves gathering data from multiple sources and ensuring that collected data is accurate and reliable. For any classification problem, in order for it to function properly and give results with good accuracy. A low-quality dataset will produce a low-quality machine learning model [AP22]. Data cleaning and preparation are also included in this step. In this thesis, census data and buildings 3D models for Oldenburg are the primary datasets.

3) **R3 – Grid-based building analysis**: This requirement involves the integration of Buildings 3D dataset and Census datasets using a spatial join for analysis. It also highlights the inconsistencies present in both datasets and are explained in detail in section 4.1. This combined dataset will be as a basis for further analysis and requirements.

4) **R4 – Population estimation for individual buildings using ML methods**: This involves studying different machine learning algorithms and analysing their performances. This study utilises unsupervised clustering K-Means algorithm and a supervised XGBoost algorithm.

5) **R5 – Supervised ML analysis & prediction**: In this thesis, a supervised machine learning model called XGBoost is utilized. In the absence of training data for the machine learning model due to data privacy regulations, the thesis attempts to create a synthetic training dataset for the model, which will later be explained in detail in Chapter 4.

6) **R6 – The model should be optimized for better performance and accuracy**: Once a machine learning model is built, it is necessary to perform hyper-parameter optimization such that the model is fit to the best possible parameter combinations as explained in detail in chapter 4. This type of optimization technique improves the accuracy and overall performance of the model. In this study, it is necessary to employ such model-tuning techniques to select the best parameter combinations in XGBoost algorithm and thereby improve overall model classification results.

7) **R7 – The model should be evaluated for accuracy and performance measurements**: The developed machine learning model must be evaluated on some defined metrics of evaluations and analysis. In this thesis, we are measuring the accuracy of the model using the grid-based population census to the predicted individual occupants in each building of a grid cell.

3 State of the Art

This chapter elaborates on the related work utilizing the importance of Census data and how various heterogeneous Census collection methods can be assessed using a common framework. The significance of estimating the population distribution of individual buildings and the evaluation based on the derived requirements from the problem the analysis section is also discussed.

3.1 Introduction to Data Analysis

This section studies the core concepts used in this master thesis. The first section highlights the importance of data analysis. The second section gives a short introduction to machine learning and the classification of data. The last section lists the tools and software used in this study.

In the information era, data is available in abundance, but it is scattered in different formats and across various sources. This gives rise to the necessity to find ways to handle vast amounts of unstructured data and transform it into information. Data analysis provides a framework for the complete exploitation of data, making it available in a consistent form and this data can then be used to generate information. Once the data is cleaned and transformed, it can then be studied to uncover hidden patterns and details which would be otherwise ignored or which were not expected when initiating the analysis [Tuk62].

Data analysis finds its place in the field of research and statistics to present accurate and reliable data [Bla22]. It provides efficient solutions to problems which occur when handling huge amounts of data such as statistical errors in data, outliers and missing data. Data analysis also provides researchers with different tools such as statistical analysis and quantitative analysis. It is because of all these methodologies associated with data analysis; it has now become an integral part of every research project [Bla22].

Data Analysis Foundations

The word data is derived from the Latin word *datum* meaning plain facts or which is used as a basis of any calculation [Tha23-ol]. Fundamentally, information and knowledge are represented by symbols and hence no difference occurs between them. However, Figure 3-1 explains the difference between different levels of data. Data levels and complexity are arranged in the pyramid. With the widespread use of the internet generating huge amounts of data for one's disposal [Aga13], extracting useful information from huge datasets has silently emerged as a new challenge for researchers. Without processing, this data serves no purpose and therefore the need for data analysis strives.

The main idea of performing data analysis is to have a better understanding of the data present to us. This data before undergoing analysis is called raw data. Raw data is a collection of facts that rarely provides meaningful insights into the data [Bla22]. This type of data could possess the potential to transform into information after it is cleaned,

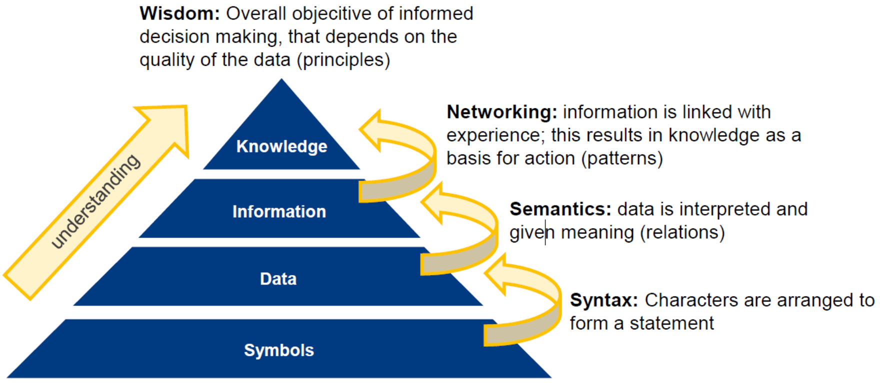

Wisdom: Overall objecitive of informed decision making, that depends on the quality of the data (principles)
Networking: information is linked with experience; this results in knowledge as a basis for action (patterns)
Semantics: data is interpreted and given meaning (relations)
Syntax: Characters are arranged to form a statement
understanding
Knowledge
Information
Data
Symbols

Figure 3-1: The Knowledge pyramid [KB09]

analyzed and transformed which could be as simple as taking an average of the data points or as complex as applying complex algorithms to the data such as regression algorithms.

In the next section, the different processes and methods for data analysis tasks are discussed. For applying those approaches, numerous tools and libraries exist to perform data analysis tasks, which will also be covered in this section.

Data Analysis Process

Data analysis defines the process to bring meaning to the gathered data. It is important to prepare the data before it can be used in the data analysis process [Tah22]. Figure 3-2, illustrates the life cycle of data from data acquisition, transforming it to meaningful information and extracting knowledge using a model.

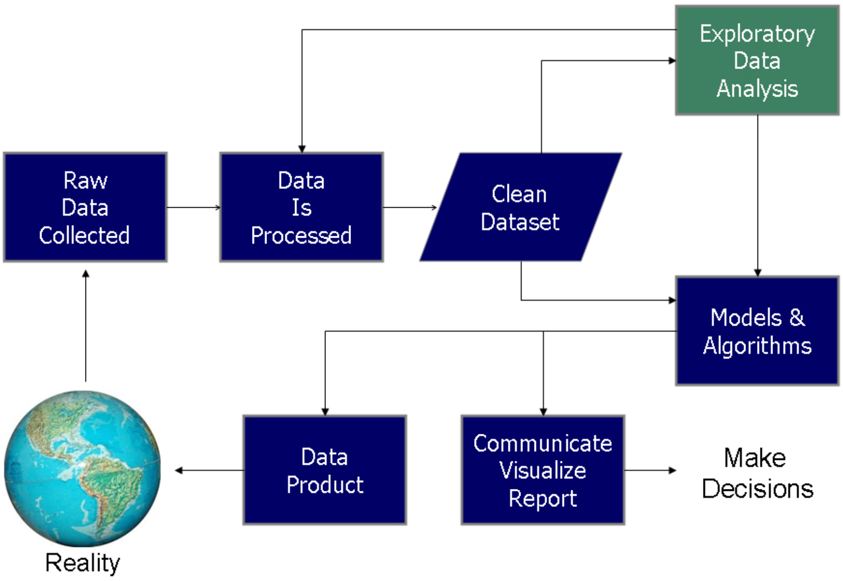


Figure 3-2: Data Analysis steps [OS13]

Data preparation involves four stages as mentioned below. All these stages are necessary to prepare the data and later use it as a feed for machine learning algorithms.

Initial Data Analysis Stage

Data analysis defines the process to bring meaning to the gathered data. However, it is important to prepare the data before it can be used in the data analysis process [Tah22]. Data preparation involves four steps as mentioned below. All these steps are necessary to prepare the data and later use it as a feed for machine learning algorithms.

1) **Data Preparation**: It is the initial step for data analysis and involves cleaning and transforming data before it can be processed and analysed. Raw data gathered from different sources might exist in different formats. Data Preparation ensures combining data in different formats into a single format ensuring its consistency [Bla22].

2) **Data Gathering**: This step involves finding relevant data sources based on the problem scenario. Acquisition of data is possible from numerous sources such as the web, articles, project work etc [Gau21]. In this thesis work, there are two main data sources namely, Census 2011 and Buildings 3D models datasets. The Census of households, buildings and population datasets along with Buildings 3D models which describe the geometry of buildings such as height, perimeter, roof type and many more characteristics which are necessary to categorize the number of inhabitants in each building were utilized in this study.

3) **Discover and Assess Data**: This step involves developing an understanding of the data and how can the available data be transformed and analysed based on the problem scenario [Bla22]. This thesis work focuses on classifying the number of residents in each building of the grid cell, therefore buildings, households and population census 2011 were used along with Buildings 3D model.

4) **Data Cleaning**: The discovered data must be cleaned before it can be used for further processing. Data cleaning focuses on removing noise, and inconsistent and incomplete datasets. This step involves removing faulty data and filling gaps. For example, removing outliers from the dataset, and filling in null values or missing values [Bla22].

After the above-described phases of data analysis, it is important to verify the data quality. Data dimensions are the concept established to measure the overall quality of data [CR19]. It is imperative that a prepared dataset adheres to the below-mentioned data standards to qualify as a high-quality dataset. Figure 3-3 illustrates the five main data dimensions which contribute to the data quality metrics.

The data dimensions are as follows:

1) **Completeness**: It means whether or not we have all the data available. Data completeness reflects if the available data is adequate for the task [CR19].

2) **Accuracy**: Data accuracy highlights whether the available data is reliable and correct in real-time [BBK+14].

3) **Consistency**: Data consistency identifies whether or not the data kept at different databases, systems and applications same and any changes made to the data at one source would reflect in all other sources[Sol23-ol].

4) **Timeliness**: Data timeliness defines whether or not the data is up-to-date. If there exists any new information related to the data making previously available data useless must be updated.

5) **Accessibility**: This metric defines the ease of data availability and data retrievability [CR19].

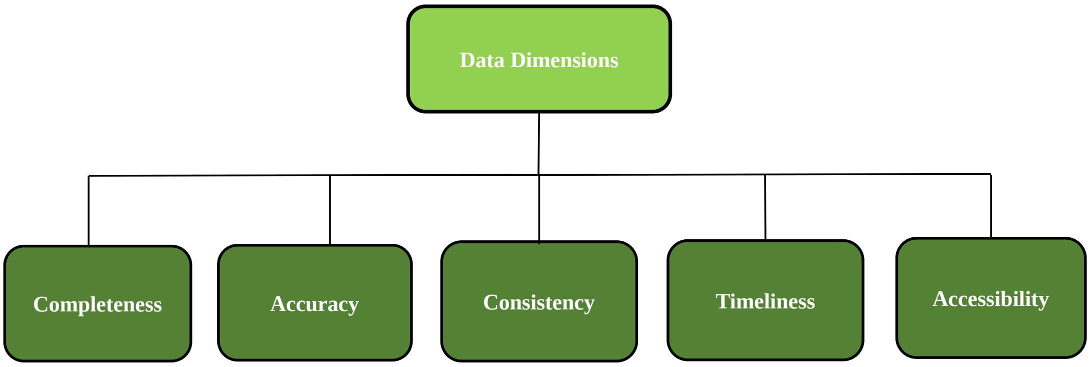


Figure 3-3: The different data dimensions contribute to the data quality metrics.

Data Analysis Stage

After the data is cleaned and adheres to the quality standards, it can be used for analysis based on the problem scenarios. Analysing the data can be categorized broadly into four main categories as explained below and as seen in figure 3-4,

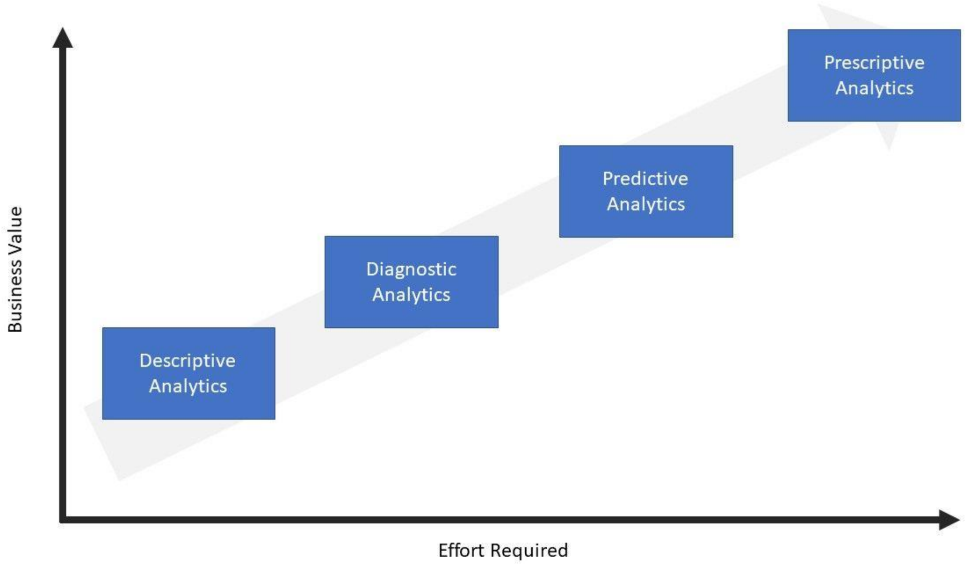


Figure 3-4: The different Data Analysis techniques based on their value and effort estimations [Sar20]

1) **Descriptive Analysis**: It defines data distributions and characteristics of the data. This type of analysis deals with interpreting the available data using statistical approaches. For example, the Mean, median and mode of the data measures central [HAK+17].

2) **Diagnostic Analysis**: This type of analysis answers the question of why an occurrence or anomaly occurred in the data. For example, exploratory analysis is a type of diagnostic analysis used to investigate and examine datasets [Tea23-ol].

3) **Predictive Analysis**: It identifies what is likely to happen in the near future based on the historical dataset. For example, predicting future energy trends based on current energy consumption data. Methods such as clustering, regression and many more ML algorithms can be used for prediction analysis [AS19].

4) **Prescriptive Analysis**: This type of analysis defines the steps to be taken to achieve the desired result or mitigate risk in the near future. It utilizes optimization algorithms to identify the best case among different available cases. For example, simulation optimization [CCK+19].

Data Analysis Techniques

Once data is pre-processed, it is free from inconsistencies and noise distributions. A dataset adhering to the high data quality standards as discussed above can then be used for further processing and analysis tasks. There are a number of different data analysis techniques to extract meaningful information from pre-processed data. The usage and selection of data analysis technique depends on the use-case and desired form of result such as [Ste23-ol],

1) **Regression Analysis**: In this analysis technique the aim is to discover how a number of variables impact a dependent variable. This approach highlights the relationship between a set of different variables. Regression can be applied to study a dependent variable and any number of independent variables as shown in Figure 3-5. While identifying variables dependencies this technique does not provide information and details on the identified variable relationships. This technique finds its use scenarios involving in making predictions and identifying trends, and patterns.

   There are numerous types of regression techniques available and their use depends on the type of data being experimented. For example, regression techniques for continuous dependent variables such as linear regression and non-linear regression. Another type of regression is for categorical values, these are values which can be grouped into a finite number of classes or categories. Logistic regression and binary logistic regression are a few examples which belong to categorical regression techniques.

   The paper [SSR16] compares linear regression and support vector regression techniques on time-series data for prediction and modelling. Extending the use of regression models in model checking, authors in the paper [YDR09] have presented a regression model checking(RMC) technique that can be applied to models after regression analysis. This RMC ensures performance optimizations such as model tuning and therefore contributes to improving the regression results.

2) **Monte Carlo Simulation**: The Monte Carlo method helps study and analysis all different possible outcome sets and their probabilities as shown in Figure 3-6. This technique helps the user to identify the effect of unpredictable variables on a specific output variable and thereby making it an ideal choice for requirements involving risk analysis. Thus, enabling the making of better decisions and risk management

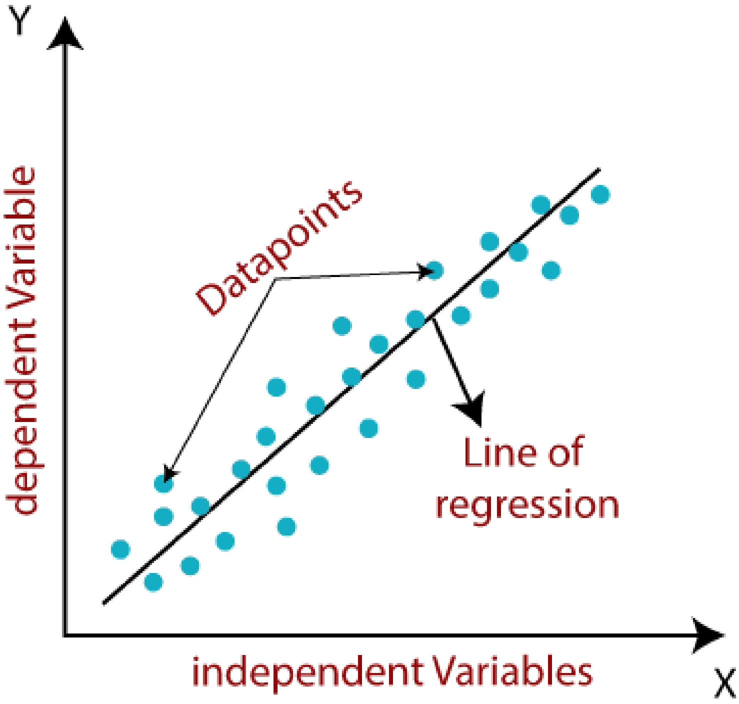


Figure 3-5: Linear Regression used for modelling relationship between a dependent variable with a given set of independent variables. [Sar20]

strategies for the future.

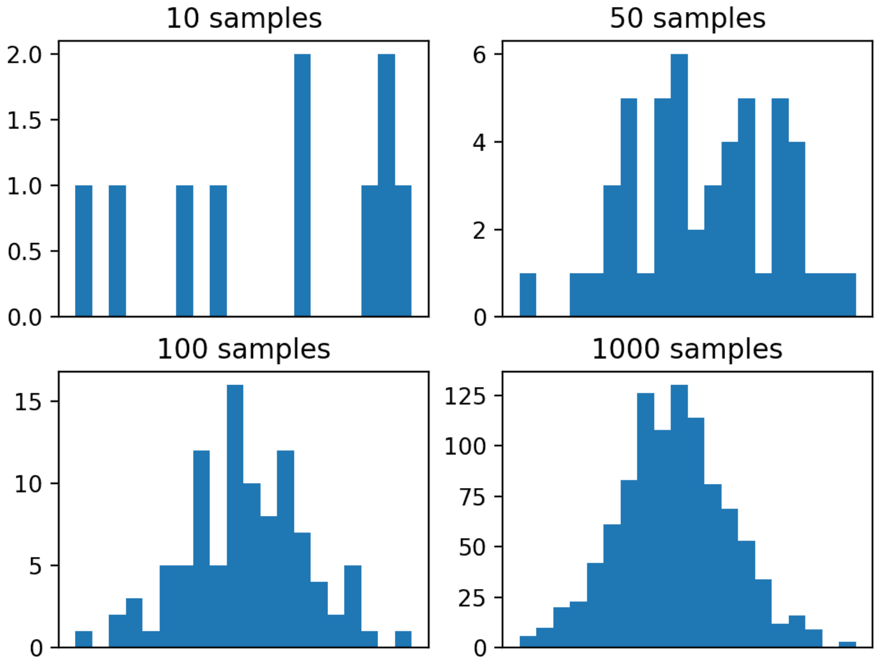


Figure 3-6: An example employing the Monte Carlo technique and highlighting the benefit of increased accuracy with the number of samples drawn. [Bro16]

In the paper [ASH+21], authors propose a new algorithm for energy management algorithm and study its performance on varying energy systems. Using the Monte

Carlo method, different characteristics of the energy system, and the different scenarios are defined according to a morphological analysis are analysed.

In an effort to utilize this method in the buildings energy domain, authors in the paper [Sim19] have used a Sequential Monte Carlo method to train a lumped building energy model (RC model), and estimate a heat loss coefficient.

3) **Cluster Analysis**: Cluster Analysis is used to identify structures or patterns within the dataset of our interest. The individual data points are sorted into groups. However, clustering does not provide insights into how and why the data points are grouped together or do not exist in the same group.

In the paper [SPP+21], to minimize the operational costs the authors study and analyse a cluster of educational buildings, equipped with photovoltaic installations and hybrid systems. Another such study in the field of building energy management was done by Chicco et al. [CBP+15]. The authors designed a framework for performance assessment employing clustering for thermal energy management in buildings. The clustering technique is discussed in detail in section 3.2.

4) **Time Series Analysis**: statistical technique to uncover patterns and cycles over a time period. This technique uses data points to study the same variable over a time frame thereby helping to identify and forecast how the variable of interest may change in the future.

In the paper, [ZPH+13] a data-driven time-series-based model is proposed for building energy consumption prediction and applies it to two actual commercial buildings. This model is used to predict energy consumption based on monthly actual energy consumption data. Another research study proposed a framework for utilizing learning-based modelling. This model will be used for the prediction of relevant time series to support comfort satisfaction and resource efficiency in building energy management [SPK18].

3.2 Machine Learning

Using machine learning and data analysis techniques possesses the potential to transform raw data into very valuable assets which is one of the key reasons why machine learning techniques are widely used around the globe [CPC19].

The idea of teaching a computer to acquire human-like intelligence roots back to the 1950s when Alan Turning's experiment demonstrated that a computer can trick a human into thinking it is a human [Mar16-ol]. In this century, with the availability of an abundance of heterogeneous data and computational power, machine learning systems' predictive performances have shown significant improvement over the years [NAM+17]. These systems have made their home in a diverse range of domains, products and services. For example, personal voice assistants, and movie recommendation systems. Machine learning has established its way into our day-to-day lives. There has also been an increasing trend in using ML across healthcare systems, criminal justice, environment monitoring, etc [PRM21].

Machine learning is a branch of Artificial Intelligence (AI) which is used to predict

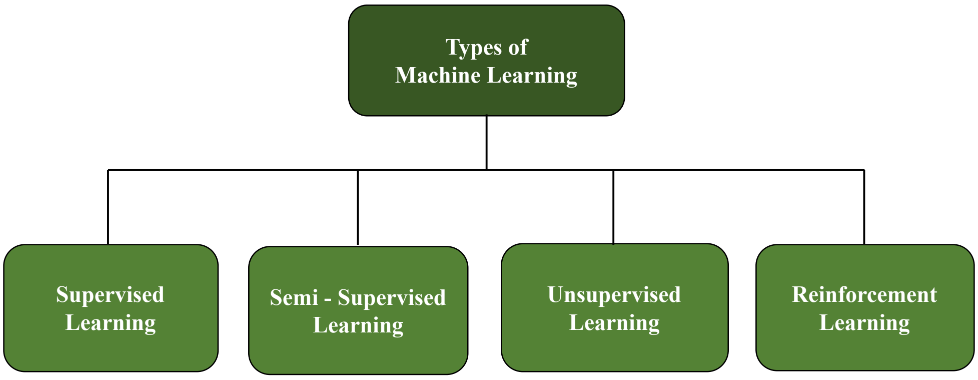


Figure 3-7: Classification of Machine Learning Techniques Based on Learning Problems

outcomes of applications without being explicitly programmed to do so. Machine learning methods learn from historical data for prediction and detection tasks [Bur-ol].

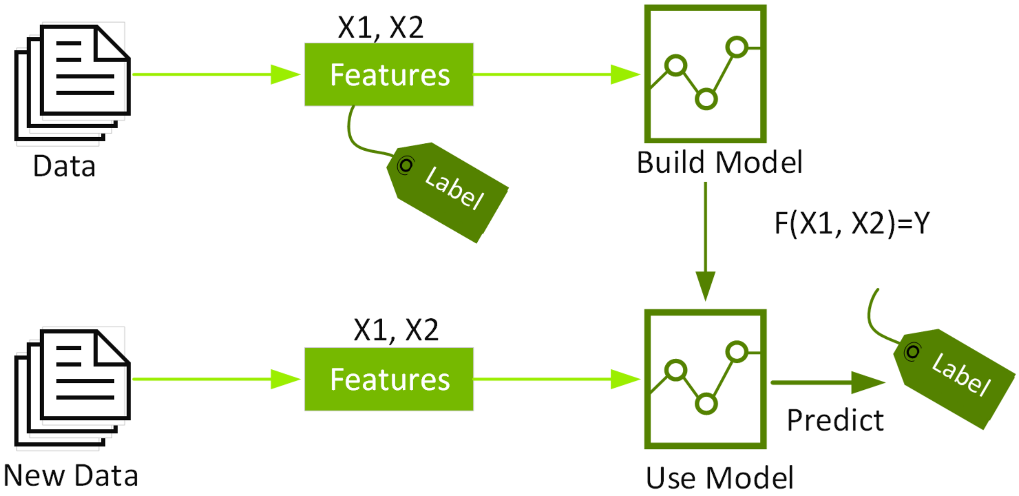


Figure 3-8: Supervised machine learning algorithmic approach [GFL18]

.

When presented with new data, these machine learning applications learn, grow and adapt to develop by themselves to uncover meaningful information and perform the predefined task without being told where to look from. This becomes possible due to the availability of numerous ML algorithms that learn from data in iterative processes. Data Scientists and Data Analysts can utilise these algorithms based on their problem scenario and expected results [Pri23-ol]. One such example of using machine learning in research is in the paper, [SDA+22] the authors have proposed a machine learning model for statistical analysis of datasets. The new model GPT3 can be used for the prediction of data insights from the calculated statistics of the large dataset. Utilizing deep neural networks GPT3 ensures performance and reliability with huge datasets. This framework will efficiently remove the dependencies on traditional data analysis methods such as correlation matrices, p-values and many more such methods.

A typical ML algorithm consists of three components [IBM-ol]:

1) **A Decision Process**: Generally, ML is used for classification and prediction tasks. After performing a sequence of calculations on the input data which may be labelled or unlabelled, the ML algorithm provides results based on the knowledge it has acquired from the pattern present in the input data.

2) **An Error Function**: An error function defines the model's accuracy. In scenarios where the original output is available, the error function can be used to make a comparison and calculate the accuracy of the model.

3) **An Optimization Process**: Based on the calculated error function, the model tries to fit itself better and tries to avoid making the same mistakes as done in previous iterations. This can be done by updating the necessary parameters of the algorithm. The optimization process can run iteratively, updating the parameters autonomously until the desired accuracy is reached by the model.

Figure 3-7, depicts the classification of ML algorithms based on the nature of the information made available during the learning process [Bur-ol], [Pri23-ol], [Raj22-ol].

1) **Supervised Learning**: This learning algorithm uses labelled datasets to train the model. The input and output of the algorithm are specified in this case.

2) **Reinforcement Learning**: This type of learning consists of three main components namely, the agent, environment and actions. The agent is responsible for making action by interacting with the environment. The algorithm works when the agent chooses actions that maximise the reward for a given time period as shown in figure 3-10.

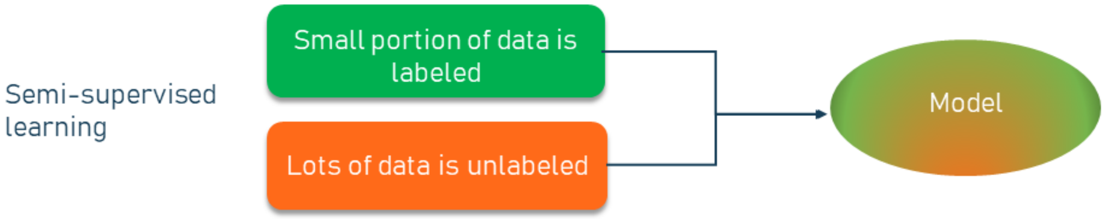


Figure 3-9: Semi-Supervised machine learning algorithmic approach [Alt22-ol]

.

3) **Semi-supervised Learning**: This set of learning algorithms is based on supervised and unsupervised machine learning algorithms. This learning helps to solve the problem where there is an unavailability of sufficient training datasets as shown in figure 3-9.

4) **Unsupervised Learning**: In an unsupervised learning algorithm, the training data is not dependent on labelled training data. The algorithm learns from the structure of the input data pattern and gives the output.

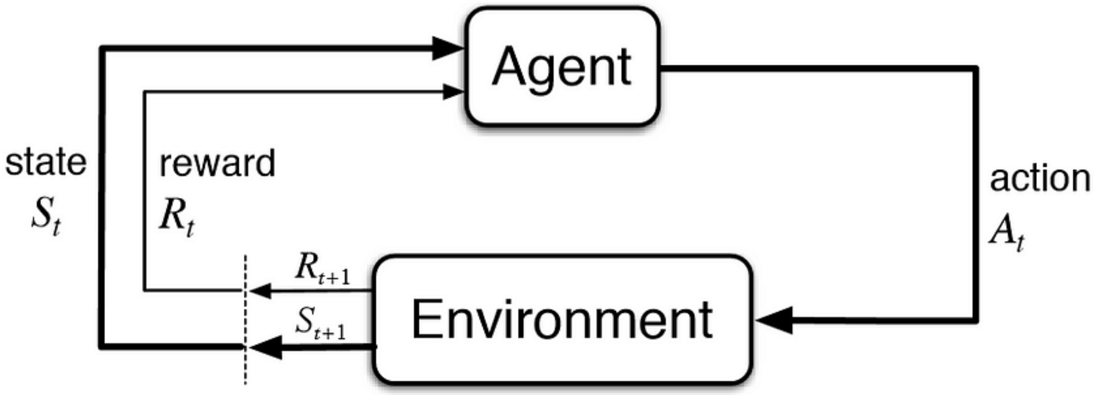


Figure 3-10:Reinforcement machine learning approach [Bha18-ol]

.

Clustering

The word clusters come from an old English word *clysters*, meaning a bunch. Clustering is the method of grouping similar things together. Data points residing within one cluster have similar properties. Machine learning makes use of this characteristic for data segregation [Mar21-ol], [Tea22-ol].

Clustering belongs to a class of unsupervised machine learning algorithms which uses unlabelled datasets to train the model. The aim of developing this technique was to, try to build human-like cognitive abilities in machines.

Machines, unlike humans, find it challenging to group data together when the input is unlabelled data sets, therefore techniques like clustering offered by ML are of great importance and find their applications in many domains for example, anomaly detection, spam detection, social-network analysis etc. [Mar21-ol], [Tea22-ol].

This technique exploits the idea that data points in a group have similar properties and data points belonging to different groups exhibit different characteristics. In the absence of labelled data, this technique interprets the data by finding patterns and observing different attributes of the data set. In Clustering, new data points are mapped to already identified clusters based on their characteristics.

Figure 3-11 depicts the output of a clustering algorithm. Here, different colours represent different clusters and data points belonging to the same cluster exhibit similar properties while data points in different clusters possess dis-similar characteristics.

Every analysis of clusters follows a series of steps,

1) **Extracting and selecting features**: this is the first step in any cluster analysis. Here, features are extracted from the input to interpret the unlabelled data set.

2) **Selecting clustering algorithm**: ML offers a range of clustering algorithms to choose from based on the problem scenario.

3) **Evaluation and interpretation of result**: This step involves the evaluation of the obtained clusters and an explanation of how the clusters were formed by the algorithm, and which features were considered.

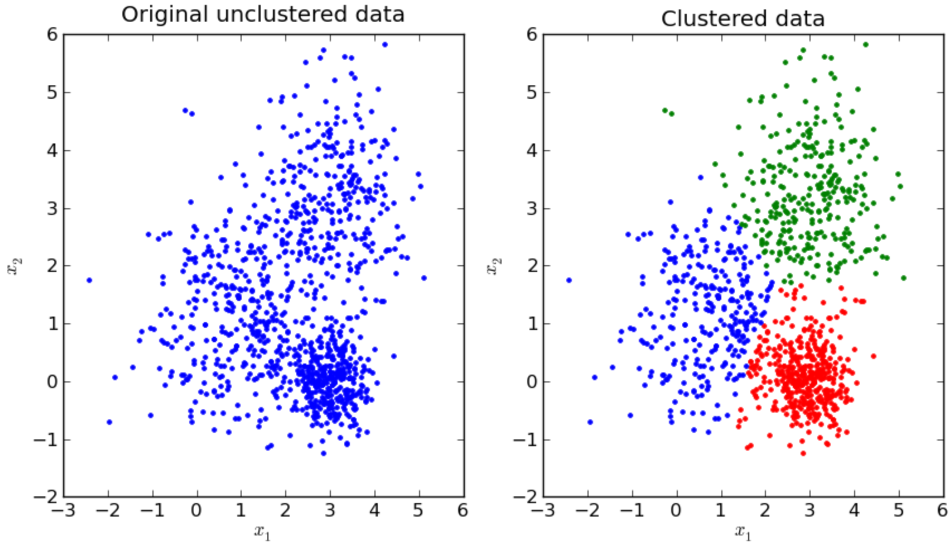


Figure 3-11:The application of clustering algorithm on an unlabelled dataset [GFL18]

.

K-Means Clustering Algorithm

K-means belongs to the class of partition-based clustering algorithms. As the name suggests, the data is divided into *k* number of clusters and any new data point is fed to already identified clusters based on their features and properties.

This algorithm forms clusters between data points based on their spatial distances and hence provides a range of distance calculation algorithms to choose from, for example, the Manhattan distance, and Euclidean distance. It is the most popular clustering algorithm and easy to understand. Its applications are spread across many fields image detection, sensor measurements etc.,

In this algorithm, the number of clusters *k*, can be chosen randomly or defined at input using cluster optimization techniques like the elbow method. Once the number of clusters is defined, respective cluster centroids are chosen and data points are grouped into subsequent clusters based on spatial distances to the centroids of the clusters. This process of choosing the cluster centroid and assigning data points to a cluster is an iterative process until the metric of the data points to their nearest cluster centre is minimized [Mar21-ol].

Cluster Optimization

In clustering algorithms, the number of clusters *k* can be defined in input or one can use optimal clustering techniques like elbow method [CMM21].

The elbow method works in iterations as, for a range of *k* for example, from *k* =1 to 10. Each iteration, Within- Sum of Clusters (WCSS) is calculated for each cluster. WCSS defines the difference between the sum of squares of data points and the centroids. As the number of cluster increases, the WCSS values begins to drop, this point where there is a sharp change in WCSS value is identified as the elbow point and marks the optimal number of clusters for clustering algorithms.

This thesis utilises the K-Means clustering algorithm for the classification of the number of residents in each building of a grid cell. The study determines the optimal number of clusters using the elbow method. The output and details are explained in the results section.

XgBoost Algorithm

Gradient Boosting algorithms have received a lot of attention from researchers around the globe in recent years. This is mainly due to the improved performances of decision tree algorithms as compared to other machine learning approaches.

In the Gradient Boosting approach, the model's performance in terms of accuracy, efficiency, and interpretability is improved using an ensemble of weak learners. These models are an extension of decision tree algorithms and the output is a combined result of all decision trees. The main idea is to use the error residuals of the previous model to fit the next model as shown in figure 3-12 thereby improving the performance of a single weak model by combining it with multiple weak models and generating a collective strong model [Nvi23-ol] as shown in figure 3-12.

One of the most popular Gradient boosting approaches is the Xgboost algorithm which focuses on computational speed and performance [Sah23-ol]. This algorithm can be utilized for regression and classification problems and possesses the capability to handle structured and tabular data efficiently. It is integrated with machine learning tools available in Scikit learn package in Python and caret in R [Nvi23-ol]. Other highlighting features are parallel processing support is enabled, cache optimization is included and efficient memory management for large datasets exceeding RAM. The main advantages of employing Xgboost algorithm are as follows [Jai23-ol],

1) **Provision for Regularization**: Standard GBM methods lack regularization and hence suffer from overfitting issues. However, Xgboost prevents overfitting by adding additional constraint penalties. It includes regularization terms in the objective function to be optimized during the training process and hence is also known as the 'regularized boosting' technique.

2) **Parallel Processing**: Traditional gradient boosting methods build sequential decision trees and each tree tries to correct the mistakes of the previous one in a sequential manner thereby limiting performance adversely. Xgboost is empowered to support parallel processing and hence it utilizes available computational resources such as multiple threads, organized cache access and even support Hadoop and distributed computing frameworks. This significantly reduces training time and the overall performance of the model improves.

3) **High Flexibility**: Xgboost provides users with the ability to customize to define custom objective functions and evaluation metrics. The objective function is the loss function which is to be minimized by the algorithm. This helps the user to define evaluation metrics for requirements which are not dealt with by the standard Xgboost algorithm. Additionally, users can define custom evaluation metrics to study the model's performance if the already available ones do not suit their requirement definitions. Hence, Xgboost allows custom tuning of the model parameters. This

helps to cater to specific problem scenarios which are not addressed by the algorithm.

4) **Tree Pruning**: Tree pruning refers to the process of collapsing decision trees to reduce the complexity and overfitting of the model. Xgboost uses post-pruning using maximum depth and minimum loss reduction parameters. The maximum depth parameter establishes a value for the maximum permissible depth of a tree. A tree exceeding this parameter is then stopped splitting further and converted to a leaf node which helps combat overfitting and complexity. Similarly, a leaf node is formed when a value below the minimum loss reduction parameter threshold is reached. These parameters ensure only meaningful splits are made.

5) **Built-in Cross Validation check**: In traditional gradient boosting (GBM) a grid search over a predefined set of values for the number of iterations is required. However, Xgboost algorithm has an in-built early stopping functionality which efficiently determines the optimal number of boosting iterations during the training process thereby enables to obtain the best model performance while saving time and computational effort.

Xgboost algorithm has been employed successfully in the field of building energy management.

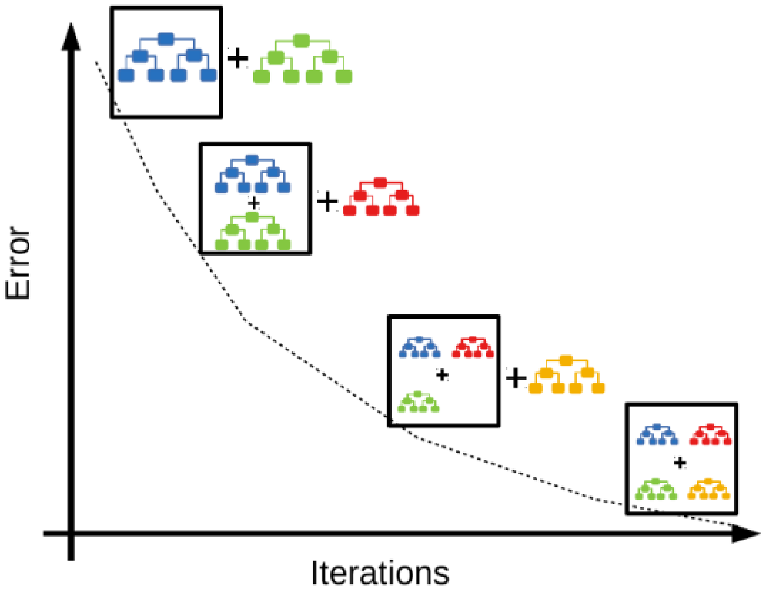


Figure 3-12: Boosting approach using the error residuals of the previous model to fit the next model [Sah23-ol]

In the paper, [AGA+19] authors propose an ensemble learning approach for reducing load while maintaining occupants' comfort in residential buildings. Xgboost has been employed in their work to avoid overfitting problems and build an efficient and robust prediction model.

Another research involving Xgboost is discussed in the paper [CWA+22]. The authors propose a building energy prediction algorithm (AM-GBDT) that combines the gradient-boosted decision tree (GBDT) to improve the accuracy of the building energy prediction model. Simulation results highlight that the proposed building energy prediction algorithm improved the predicted building energy performance metrics and also reduced the training time of the model.

3.3 Tools and Software

Python

There is mounting evidence of Python's growing popularity from various surveys [NG19],[Pyt21-ol] which suggests that Python is the most studied and used language for data engineering and machine learning. This section covers fundamental elements of the language which were the main reasons for choosing Python for the majority of the implementation in this thesis work.

Python came into existence in the late 1990s [NG19], [PNS19] is an object-oriented, high-level language and is built on top of abstract data types which enables faster application development. Being an interpreter-based language eliminates the need to compile the code before executing it. This makes the debugging easy and in turn, reduces the development time effectively.

The availability of a wide range of libraries leads to fewer lines of code, helping the programmer to focus more on the logic and testing and spend less time on the syntax. To implement the same logic, Python uses 1/5th of code as compared to other OOPs languages [NG19].

Data analysis and data science involve working with a huge amount of data. Python's ability to support large and robust standard libraries makes it easier and a popular choice for data analysis and visualization [BRS19]. The popularity of Python in the ML domain is also due to the enormous online community which provides help in various stages of the data analysis process. While ML involves complex algorithms, python offers simple and readable code enabling the developers to put all their efforts into solving the ML problem instead of focusing on the technical issues of the language. Besides supporting a rich technology stack with several libraries for the machine learning domains, namely NumPy, Pandas and many more. Python's platform-independent feature has led to Python's wide usage.

Numerous operating systems are supported such as Linux, Windows, and macOS. Python, which is widely used for data analysis tasks, also appears to provide efficient interfaces to all major commercial and open-source databases like MySQL. Python's efficient interface support and open-source availability are the key reasons for using it in this thesis study.

GIS

A geographical information system (GIS) is a system that maps and analyzes spatial data. There are numerous features and applications of GIS namely:

1) GIS supports computer cartography which is map making.

2) GIS allows associating information (non-geographical data) with places (geographical).

3) Spatial analysis: GIS is used for spatial and statistical analysis. One of the methods includes adding layers to a map which creates a separate view of the original map using some attributes.

4) The result of the analysis via GIS can be derivative information, interpolated information, or prioritized information.

In this thesis, the software QGIS is used to visualize GIS data in both vector and raster formats. The key reasons for choosing QGIS were that it's open-source software and provides support to various operating systems Windows, MacOS and Unix platforms [KSS+18] making it platform independent. The well-documented tutorials along with well-explained examples and popular online community support of QGIS developers also contribute to choosing this software for the thesis work. This study uses QGIS for visualization of the 3D Buildings models along with census databanks.

QGIS helped to identify the scenarios where a building is present in two grid cells concurrently and therefore centroid algorithms (built-in QGIS) were used such that each building exists only in one grid cell. In this thesis, QGIS has also been used for result verifications using its attribute tables feature which made it possible to verify the output of spatial join produced from Python code to the output of spatial join done on QGIS.

SQL

The popularity of SQL is highlighted due to its high-level syntax which requires very less programming for most of the queries. Its platform-independent nature makes it possible to be implemented in many kinds of DBMS(Database Management Systems), from desktops to open sources such as MySQL to commercial types such as Oracle, Microsoft SQL Servers and more [FS15]. The First SQL standard was published in the late 1980s and since then it has gone through various updates.

ML and data analytics developers acquire data from heterogeneous data sources and formats (.gpkg, .csv, etc.). To analyse and extract meaningful information from the data, it must be imported and transformed. SQL is capable of extracting data from huge database systems based on relational algebra with very low run-time complexity.

The core of SQL resides in its SELECT which can be manipulated for filtering of records, columns, grouping and data processing tasks (GROUP BY and HAVING clauses). Moreover, the result of these queries can not only be stored inside a database or a view but also be exported in many different formats such as text files, CSV files, pickle files, etc.

In this thesis work, SQLite has been used for data management [BP15]. SQLite is an open-source advanced database system. Besides storing vast amounts of data, it also enables to perform data processing using SQL. SQLite's serverless interface (inbuilt server in Python), zero-configuration, and freely available and easy-to-be-established connection with Python were the main reasons behind choosing it for this thesis work.

GitHub

Git serves as the repository on GitHub. Git is a distributed version control system for software development that was first made available in 2005 [RA12]. It has made it simpler to contribute to open-source projects by providing support for version control, issue tracking (bugs and feature requests), documentation, notifications, diffs, and status dashboards.

Before GitHub, a developer who wanted to contribute to a project had to download the source code, make changes, and then email the patch for review. Using GitHub, you can "fork" the source code, make your modifications, and then submit a "pull" request for review. In GitHub, the project's collaborators can review and contribute to documentation files. The thesis work chooses to make the implementation (code) open-source thereby making GitHub a perfect choice

3.4 Related Work for Census and Census Quality

Census collection is a crucial process for any country. It estimates the country's population by age, sex, and region. It serves as of great importance for municipal infrastructure project planning. Additionally, it is the census data which helps the government to make public funds distribution decisions among states and regions.

A study by Michaela and Rembrandt [SK16] highlight the importance of Census collection and also explains in detail the history of the census in Germany. They also discuss the census collection method used i.e., the register-based approach for the first time in the country in 2011.

In the study [Pre13], authors Prewitt and Kenneth explain the importance of the United States Census and how this collected data significantly contribute to research. Besides being a demographic measure, the census data serves to understand the social and economic characteristics of the population and thereby impacting the policies in the country. The authors also discuss the challenges involved in maintaining the privacy and accuracy of the collected and published census data.

Census affects policy-making at the local and international levels. Countries like the United States and European Union host census after every ten years, but the data collection method varies. The census results, besides being used by the statistical department as a benchmark are also used for facilitating international comparisons such as Millennium Development Goals thereby making census quality measurement an important attribute. Census quality measure is important to maintain the integrity and utility of the information produced. However, errors are inevitable in the census data collection process which makes it a more challenging task to measure its quality. Another reason measuring census data quality is difficult is the absence of a standard method that applies to all the different census collection strategies and processes used by different countries. In the paper [BV12], the authors introduce a common census quality assessment framework and assess the quality of different census methodologies. According to the paper, the census is defined as a multi-dimensional representation of information with six characteristics namely,

1) **Relevance**: The relevance of the collected census data measures whether the needs of the stakeholder, users and population are met. The degree of the 'value for money'

obtained from the census results defines its relevance measure.

2) **Accuracy**: Accuracy measure in terms of the census is related to the accuracy and the errors afflicted during various stages of census design, collection and processing procedures. This metric is closely linked to coherence and timeliness as the decisions to improve accuracy impact the delivery time span of the census results.

3) **Timeliness**: This metric measures the time taken for the entire census process from census design till the availability of results. If this time is very long then it affects the validity and relevance of the results thereby highlighting a significant trade-off between the accuracy metric.

4) **Accessibility**: The metric includes not only the availability of census results but also includes the confidentiality of sensitive data sets without comprising the data quality. The results of the census are utilized by government, private agencies, researchers and various businesses and hence this data must be accessible, interpretable and coherent.

5) **Interpretability**: This metric defines the degree to which the census results are easy to understand and easy to find. It also includes the availability of supplementary data or metadata which brings meaning to the census results for users and how well-defined the metadata is in terms of definitions, terminologies, classifications and references.

6) **Coherence**: Coherence measures how well the census results can be used together with other statistical results. It also focuses on the integrity of data which can be checked using administrative data, older survey results and validation checks on the census data explaining any deviations of the data trends.

3.5 Designing Energy-Efficient Buildings

Energy-efficient buildings can help reduce the energy demands of cities [ÜEG+12]. This has inspired intensive research in the field of estimating the energy consumption of buildings and infrastructure for planning and evaluating energy-saving strategies but the availability of high-quality and complete data for this type of research has proven to be challenging. Additionally, the availability of building stock data including information on building height, and floor area is limited. To address these research limitations, WURM et al. [WDS+21]; leveraged crowd-generated databases, such as Open Street Map Projects (OSM), and employed automated image analysis of remotely sensed images to obtain superior quality building stock information.

The authors have used a convolutional neural network (CNN) to build a stock model from aerial images and used the attributes of construction type and building age of the census dataset to develop a random forest model. Figure 3-13 provides an overview of the different methods used to model a building stock.

The importance of employing CNN for not only high-quality images but also increased performance when compared to traditional methods of land surveying and manually digitized building footprint generations making CNN a valuable tool for large-scale building stock generation is also discussed. However, their work was limited to a few cities in the country

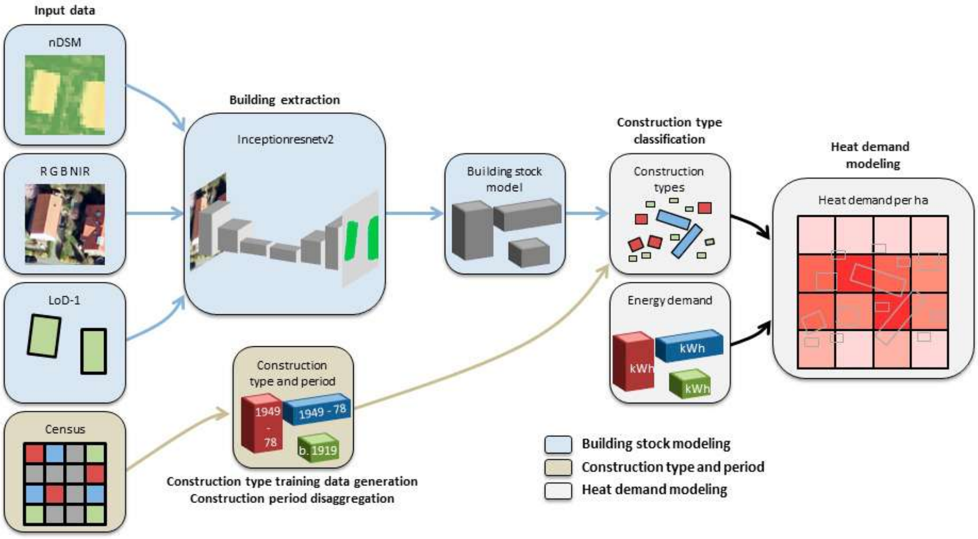


Figure 3-13:Workflow of methods and datasets utilized by WURM et al. [WDS+21] to build a building stock using CNN

and focused on building age predominantly. This thesis aims to develop a general building stock of the country using the census 2011 dataset, integrating not just the building age but characteristics like heating type, ownership, and number of inhabitants per building.

Another eminent contribution to the field of building energy-efficient buildings is highlighted in the work of Garbasevschi et al. [ASV+21]. Their work develops a machine learning model for the classification and prediction of building age in many cities of the North-Rhine Westphalia state of Germany. The authors propose a random forest model which incorporates street and block metrics along with building attributes as classification features. They also discuss the adverse impacts of building age misclassification on building heat estimation. The thesis focuses on the classification of the number of residents per building in a city of Lower Saxony State i.e., Oldenburg. However, due to the unavailability of training data for the number of residents per building in Oldenburg. This thesis study attempts to develop a training data set and then utilize it for building classifications which will be explained in detail in the next chapter.

The designing of energy-efficient buildings is affected by many factors and conditions. One such parameter is the number of occupants residing in individual buildings. For designing energy-efficient buildings, the estimation of the distribution of population for individual buildings is a significant factor due to the below factors,

1) **Estimating Load Profiles of Occupants**: energy needs of individual buildings depend on resident activities and patterns. Estimating the population distribution in individual buildings help researchers and designers to design accurate and consistent load profiles which reflect occupant energy requirements and thereby help in making energy management strategies and decisions such as renovation plan, and heating and cooling systems [CLA+19].

2) **Effective Peak Load Management**: Peak load predictions and their efficient management can be accomplished by estimating the population distribution in

individual buildings. Peak Loads occur when there is a high demand an ideal scenario would be a building with the maximum number of residents. Studying the residents per building can help in predicting such peak load patterns and reduce the need for oversized and inefficient building energy systems [DDG+22].

3) **Enhanced Building Occupants' Comfort**: By understanding building resident distributions can help designers to ensure the indoor air quality, thermal comfort and lighting conditions are maintained in areas with high occupant density. This leads to occupant comfort and satisfaction, productivity and overall physical and mental well-being of the residents [DDG+22].

Understanding the significance of population estimation for building energy management, SCHUG et.al [SFL+21] propose a population mapping across Germany using weighting layer from building characteristics using Earth Observation Data and census disaggregation. In their study, they found that using building density, type and volume along with the information on living floor area is suitable to produce accurate large-area bottom-up population estimates.

Highlighting the importance of using high-resolution aerial images to map urban complexities and map features at a fine scale is explained in the paper [UHS11]. The obtained images along with surface models (GIS models) and digital terrains were utilized to determine building footprints. Additionally, city zoning maps were used for further classification of buildings to identify residential and non-residential buildings. To further classify residential buildings and identify houses and apartments, ancillary geographical data such as topographical data, climate data and infrastructural details were used as shown in Figure 3-14.

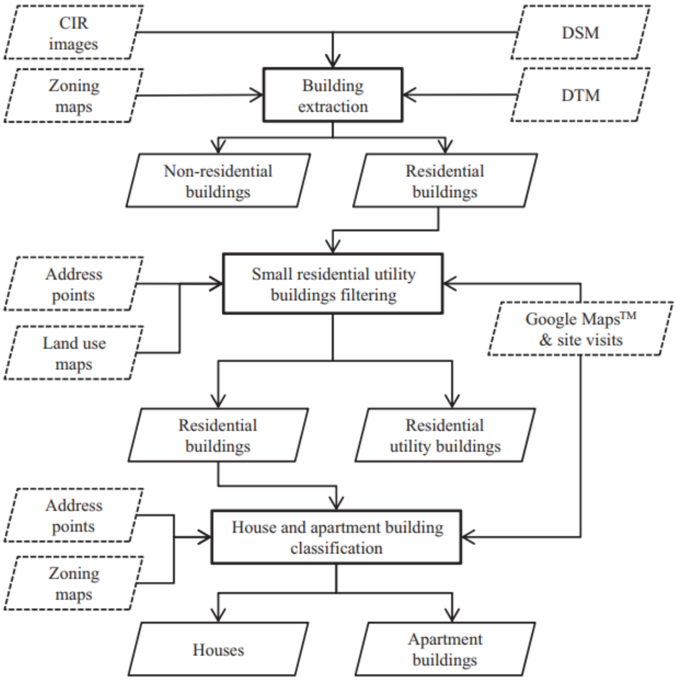

CIR images
DSM
Zoning maps
Building extraction
DTM
Non-residential buildings
Residential buildings
Address points
Small residential utility buildings filtering
Land use maps
Google Maps™ & site visits
Residential buildings
Residential utility buildings
Address points
House and apartment building classification
Zoning maps
Houses
Apartment buildings

Figure 3-14: Workflow of methods and datasets utilized by URAL et al. [UHS11] to estimate population mapping in areas of West Lafayette, Lafayette, and Wea Township, all in the state of Indiana, USA.

Results of the individual buildings were evaluated per census block with reference to the known U.S census records as shown in Figure 3-15.

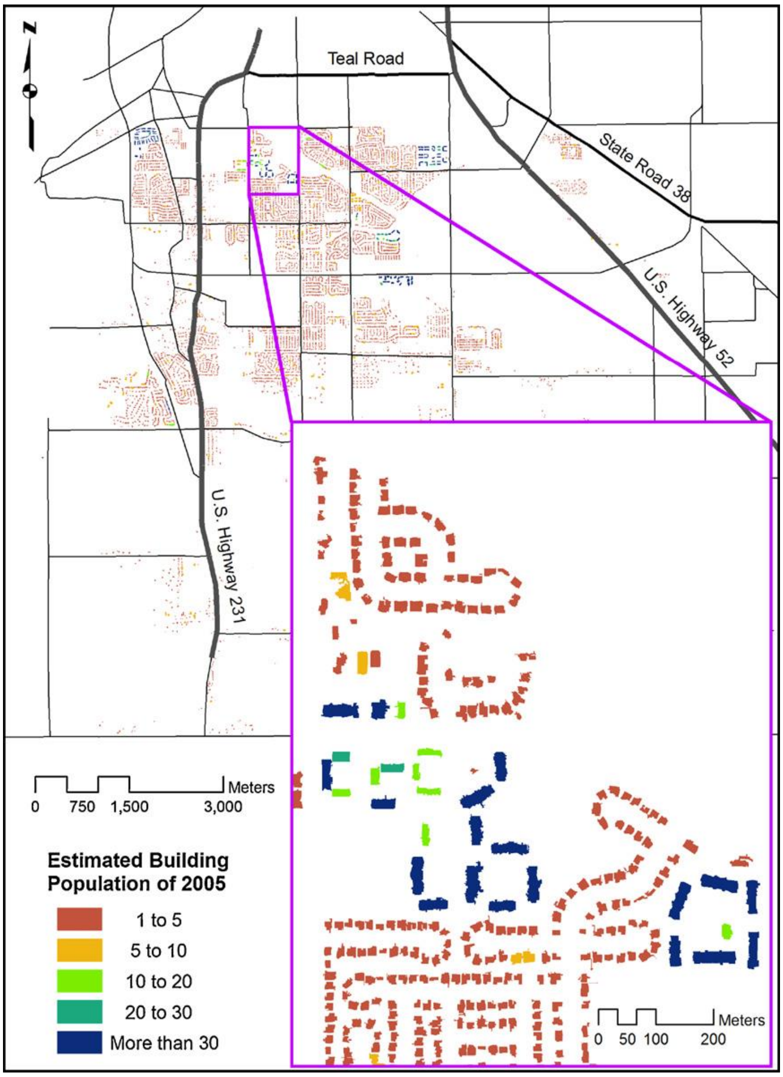
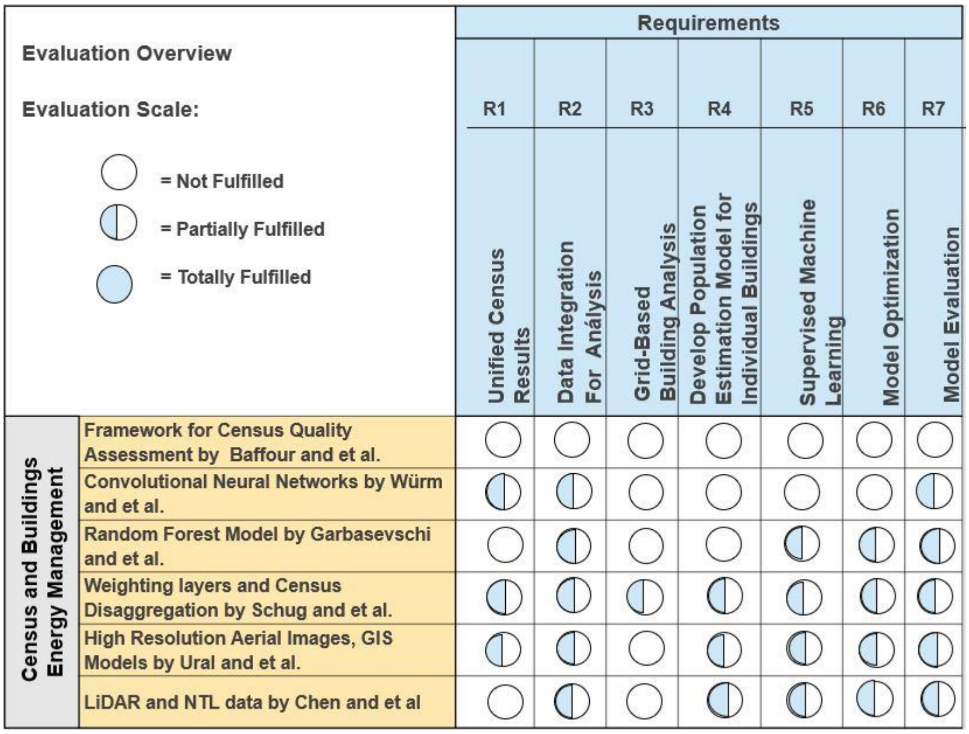


Figure 3-15: Estimated population distribution by URAL et al. [UHS11] using weighting layers.

Another study in the domain of estimating population distributions at a fine scale was done by CHEN et al. The authors found that integrating LiDAR data along with nighttime light data (NTL) for studying building infrastructures improves the quality of population estimation results. A random forest model was built on the integrated training data to estimate the population for individual buildings, followed by a validation test performed using the data from Huangpu District in Shanghai, China, [CWY+21].

3.6 Evaluation Based on the Requirements Specified

Figure 3-16 illustrates the evaluation overview of related work of Census and building energy management methodologies according to the previously defined requirements in the problem analysis chapter 2.



Evaluation Overview

Evaluation Scale:

○ = Not Fulfilled
◐ = Partially Fulfilled
● = Totally Fulfilled

		Requirements						
		R1	R2	R3	R4	R5	R6	R7
		Unified Census Results	Data Integration For Análysis	Grid-Based Building Analysis	Develop Population Estimation Model for Individual Buildings	Supervised Machine Learning	Model Optimization	Model Evaluation
Census and Buildings Energy Management	Framework for Census Quality Assessment by Baffour and et al.	○	○	○	○	○	○	○
	Convolutional Neural Networks by Würm and et al.	◐	◐	○	○	○	○	◐
	Random Forest Model by Garbasevschi and et al.	○	◐	○	○	◐	◐	◐
	Weighting layers and Census Disaggregation by Schug and et al.	◐	◐	◐	◐	◐	◐	◐
	High Resolution Aerial Images, GIS Models by Ural and et al.	◐	◐	○	◐	◐	◐	◐
	LiDAR and NTL data by Chen and et al	○	◐	○	◐	◐	◐	◐

Figure 3-16: Evaluation overview of the State of the Art approaches according to the previously defined requirements.

1) **R1 - Unified census data management to access different census datasets at once**: In the paper [BV12], the authors focus on building a census quality assessment framework and the idea of accessing multiple census datasets is not explored. The significance of combining different census datasets is highlighted in the work of [WDS+21], [SFL+21] and [UHS11]. However, the authors explore different data sources such as high-resolution images or CNN with the census of buildings and housing. The population census population is not explored in any of these studies and hence requirements are fulfilled partially.

2) **R2 - Data integration for analysis**: As seen in Figure 3-16, most of the requirements are satisfied by the research studies discussed above. All the studies except for [BV12], discuss data preparation and data analysis concepts. However, data integration processes are not explored and hence requirements are fulfilled partially.

3) **R3 - Grid-based building analysis**: In the previous studies examined, there is a lack of research addressing the use of grid-based building analysis. However, a few authors mention the significance of data distribution from a developer's perspective. In the paper[SFL+21] authors have used a census dataset (published in a grid format) and combined it with high-resolution aerial images for their results. Another study [CWY+21] utilises the grid-based building population for their model validations.

4) **R4 - Population estimation for individual buildings using ML methods**: Out of

all the research discussed, only three studies [WDS+21], [SFL+21] and [UHS11] discuss the estimation of the population to individual buildings to ensure energy management strategies and decisions. However, the discussions are brief and no results for occupant prediction are discussed and hence requirements are fulfilled partially.

5) **R5 - Supervised ML analysis & prediction**: More than fifty per cent of the studies discussed here have employed supervised learning techniques. It is mainly due to the reason that employing supervised machine learning results in better classifications than unsupervised techniques. Hence, this study also utilizes supervised ML techniques. However, supervised ML is the studies are used in combination with other techniques such as CNN and hence requirements are fulfilled partially.

6) **R6 - The model should be optimized for better performance and accuracy**: Once a machine learning model is built, its optimization becomes a crucial step to obtain best-case results. This involves utilizing methods such as hyperparameter tuning to make the model efficient and have a better prediction accuracy. Most of the papers use such model optimization methods to have better results however the optimization techniques vary and hence requirements are fulfilled partially.

7) **R7 - The model should be evaluated for accuracy and performance measurements**: This requirement is satisfied by all the research papers mentioned relating to model development. Model evaluation is a crucial and mandatory requirement for any such study project and helps to determine the quality of the trained model. Most of the papers use such model evaluation methods to have better results however the techniques vary and hence requirements are fulfilled partially.

None of the listed approaches completely meet all of the requirements. However, all of the mentioned methods contribute to at least two to three requirements, which aids in achieving success to some extent; therefore, a new solution needs to be developed, which will be further described in the upcoming chapter 4.

4 Solution Concept

This chapter is the core of the present work. It describes the solution approach for the development of a classification model for the building population estimation using Census data and Building 3D models of the city of Oldenburg. The proposed solution must meet the requirements derived from the problem analysis in section 2 and satisfy the call for the actions defined in chapter 3.

Section 4.1 describes the inconsistencies in the utilized datasets and section 4.2, section 4.3, section 4.4 and section 4.3 describe the methodology of the solution approach used in this study. These sections highlight databank preparation, preparation of training datasets, development of a classification model and optimization of the developed model.

4.1 Inconsistencies between Census 2011 and Building 3D Models

The primary datasets used in this study are not complete and are inconsistent. In this section, the different inconsistencies among the datasets are explored and a solution concept is defined.

The Buildings 3D Models and the Census present several inconsistencies therefore the need for a careful treatment of the data to ensure reliability and accuracy of the developed model. It is imperative to handle such data inconsistencies. Failing to address these issues can significantly affect the study's outcome.

It is important, in particular, to resolve the data inconsistencies such that the integrity and reliability of the findings are preserved. Inaccurate and inconsistent data can adversely affect the conclusions of the study, potentially leading to erroneous interpretations and misleading insights. By addressing these discrepancies, this thesis seeks to enhance the overall quality and reliability of its results.

Furthermore, to achieve a comprehensive and holistic understanding of the problem under investigation, it becomes necessary to address data inconsistencies. Data inconsistencies can lead to the creation of biases and gaps in the information available for analysis thereby hindering important findings. By addressing these discrepancies, this thesis aims to fill these gaps and mitigate any potential biases, thus ensuring a robust and comprehensive analysis.

For accurate classification and characterization of buildings, the resolution of the discrepancies existing among the datasets of the Buildings model and Census is important. Without addressing these issues, could lead to misclassification errors and lead to incorrect inferences about the building type and characteristics. By maintaining consistent data, the study can improve the accuracy of building classifications thereby enhancing the reliability of subsequent analyses.

The subsequent section delves into the inconsistencies identified within the datasets and attempts to propose suitable resolutions for each,

1) Addressing the unavailability of training data constitutes a significant challenge addressed in this study. Owing to data protection and privacy regulations, individual

building population data is solely accessible in an aggregated INSPIRE 100 m grid format.

To overcome this challenge, synthetic training data was specifically tailored for this study. This involved utilizing heterogeneous grid cells representing different building types. Subsequently, this training data was used to train our model, enabling us to predict the classification of buildings' occupant population. This approach allows us to work around the limitations imposed by the unavailability of test and training data, ensuring the feasibility of our analysis and the accuracy of our predictions.

2) The unavailability of combined census data of buildings, households and population poses a challenge for this study. Inconsistencies arising from the absence of combined census data hinder the study's ability to comprehensively understand and analyze the complex dynamics between the housing, buildings and population census.

To accurately capture and comprehend the inter-dependencies of the datasets, a comprehensive database is proposed. Thereby, multiple census results can now be investigated simultaneously. This study can then overcome these inconsistencies and enable insightful analyses.

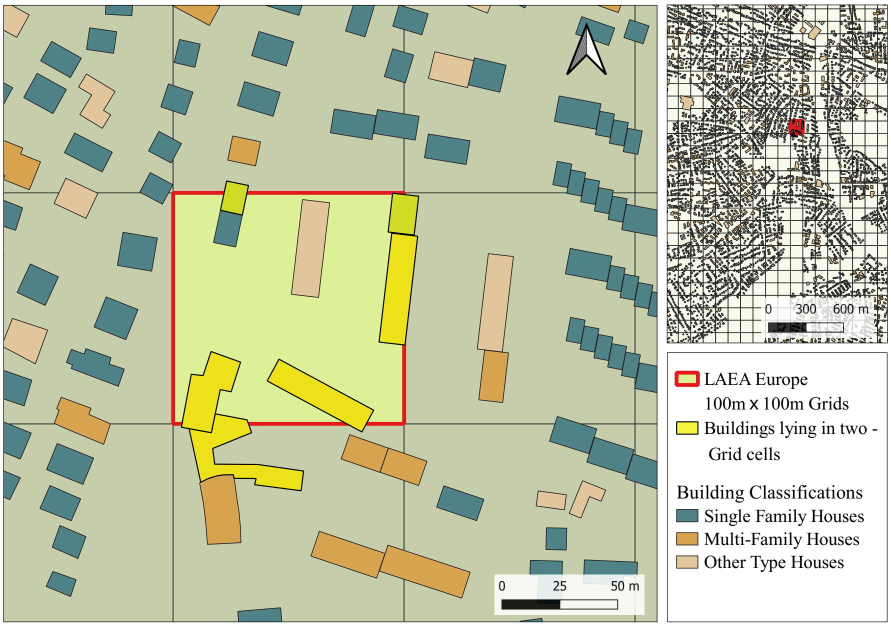


Figure 4-1: Buildings overlapping on multiple grid cells simultaneously

3) Addressing inconsistencies such as building-to-grid assignment is crucial for accurate spatial analysis. Scenarios where a building is positioned in between two grid cells, uncertainties arise regarding the appropriate grid cell assignment according to the census records as can be seen in figure 4-1. This highlights the ambiguity and lack of resolution in the published census and is one of the causes of misclassifications in this study. Failure to resolve this issue can result in inaccurate spatial and analytical outputs.

An attempt to resolve such scenarios by using the centroid-based assignments as shown in figure 4-2. This method reduces the polygon of a building to a point and this point can then be assigned to a single census grid cell. Utilizing a centroid-based assignment, resolution of ambiguity and enhancing the accuracy of building-to-grid assignments, leading to more reliable spatial analyses is our focus. However, the problem still persists as the centroid method employed is our proposed approach but the results may differ from building-to-grid assignments as published by the census results dataset.

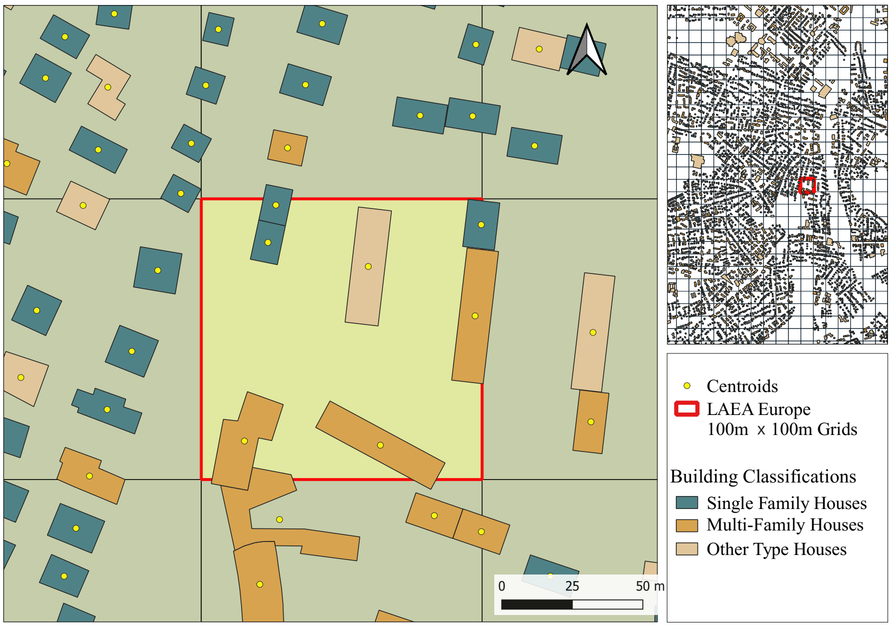


Figure 4-2: Centroid-based assignment of buildings to grid cells

4) Another source of inconsistency in the classification stems from the mismatch between the rate of building construction and model development which leads to discrepancies in the representation of the built environment. This has led to a major inconsistency in the data of the census and buildings model. Figure 4-3, shows one of the many scenarios where for a grid, the building models are not present.

Our study addresses the population estimation of each grid by utilizing census data and assigning residents to buildings for which models are available. Consequently, in the future, with the availability of consistent building models, the census results utilized in our study can be employed to accurately assign residents to each individual building.

5) Misclassifications are anticipated because the 3D CityGML models include all of the buildings within the study area (including the non-residential ones like administrative buildings, commercial, trade and other types). In contrast, the census database specifically focuses only on residential buildings. Even after filtering the buildings, substantial misclassification still persists. This can be attributed to the fact that the 3D building models comprise only geometry information while potentially lacking other relevant parameters beyond just the building's geometry.

By resolving these inconsistencies, the thesis aims to enhance classification accuracy, enable precise spatial analyses, capture inter-dependencies between variables, reflect

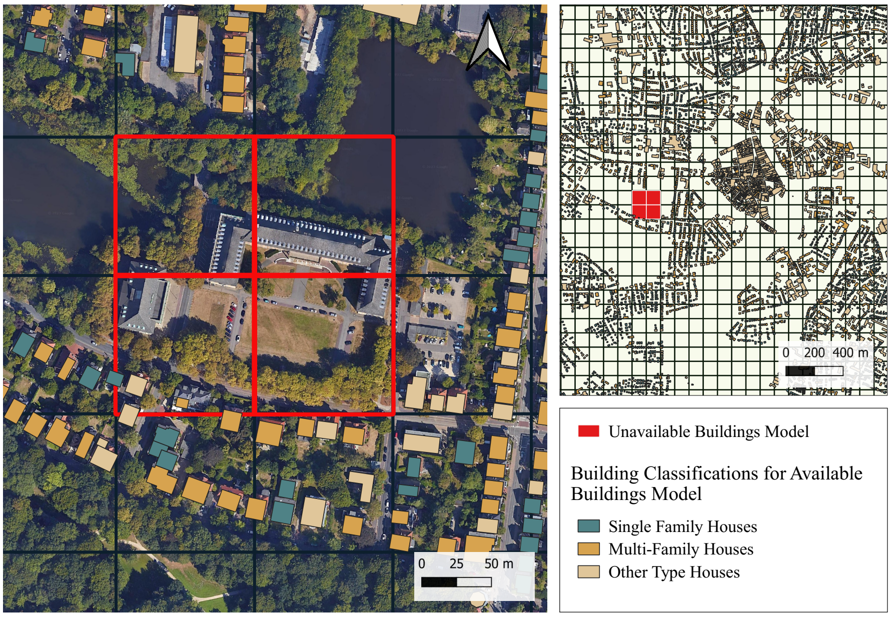

0 25 50 m
0 200 400 m
Unavailable Buildings Model
Building Classifications for Available Buildings Model
Single Family Houses
Multi-Family Houses
Other Type Houses

Figure 4-3: Unavailability of Buildings 3D models

current scenarios accurately, and prevent misleading or biased conclusions thereby improving the overall performance of the study.

4.2 Databank Preparation

Databank integrating Census datasets for Germany. This step involves the development of a comprehensive databank consolidating various census datasets across Germany. It aims to organize data and have unified access to census information of households, buildings and population datasets as shown in figure 4-4.

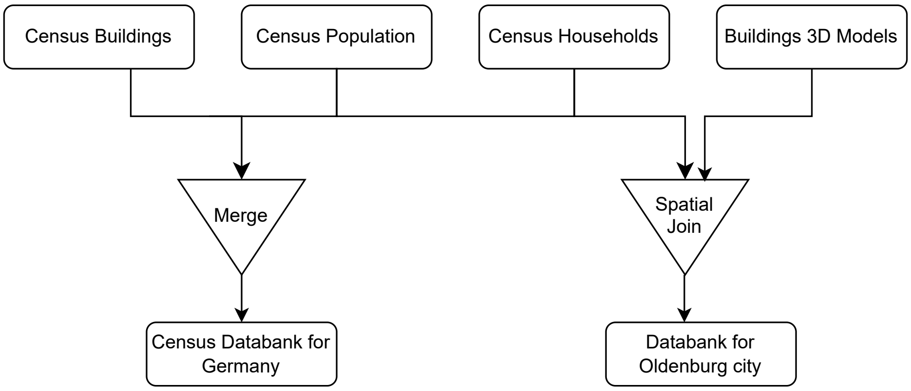


Figure 4-4: Overview of Databank Preparation step

Databank integrating Buildings 3D models and Census datasets for Oldenburg city, Germany. This step involves the development of a specialized databank comprising of Building's 3D models and previously mentioned Census datasets thereby facilitating a holistic analysis by integrating the spatial representation of buildings with demographic information as shown in figure 4-4.

To build a databank integrating various census datasets for Germany, census buildings, households and population datasets were used. The structures of these datasets are defined in table 4-2, table 4-1 and table 4-3.

Table 4-1: Data structure of published census household dataset

Grid Id	**Features**	**Specification Text**	**Quantity**
100mN26912E43412	Size of private household	1 Person household	2
100mN26912E43412	Size of private household	2 Person household	11

The databank is designed with a specific structure to ensure efficient data traversal and retrieval. Each grid ID within the databank is assigned a unique identifier, acting as a primary key for accessing the associated data. This unique identifier allows for easy navigation within the databank.

Table 4-2: Data structure of published census buildings dataset

Grid Id	**Features**	**Specification Text**	**Quantity**
100mN26912E43412	Heating Type	Central Heating	2
100mN26912E43412	Heating Type	No Heating	11

The structure and organization of the databank are discussed in chapter 5.

Table 4-3: Data structure of published census population dataset

Grid Id	**Residents**
100mN26912E43412	-1
100mN26912E41388	12

To build a specialized databank integrating various census datasets and buildings 3D models for Oldenburg, census buildings, households, population datasets and buildings 3D models using a spatial join. To extract coordinates of census datasets, shape files for Oldenburg city were used. The census coordinates were then joined with buildings models coordinates such that buildings in grids can be identified for spatial and demographic analysis.

4.3 Creation of Training Datasets

Initially, due to the unavailability of training and test datasets, experimentation with unsupervised learning methods was performed. Specifically, the K-means clustering algorithm to cluster the buildings into different types, such as single-family houses (SFH), multi-family houses (MFH), trade buildings, commercial buildings, service buildings, and others was used. The optimal number of clusters was determined using the elbow method, and scatter plots were used to study the distribution range of residents within each building type, as depicted in Figure 4-5

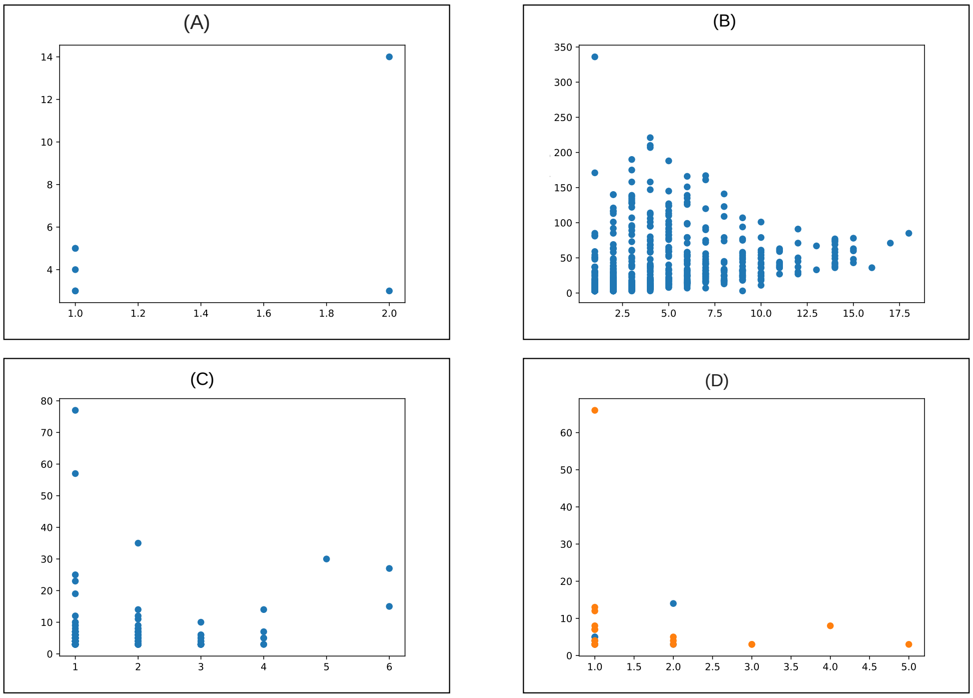


Figure 4-5: Scatter plots illustrating the distribution of residents per grid in Oldenburg across different building types: A - Trade, B - Residential, C - Industrial, and D - Service.

To assign residents to each cluster, a brute force approach within the identified ranges of residents for each building type was used.

The study aimed to find the best solution that minimized the error between the assigned cluster and the actual number of residents. However, this method resulted in significant errors, rendering it unsuitable for further building classifications. Therefore, our focus shifted towards supervised learning techniques.

Training datasets were generated for each building type class, specifically Single-Family Houses (SFH), Multi-Family House (MFH), and other building types.

Figure 4-7 and the geographical representation as shown in figure 4-6 shows the distribution of SFH, MFH and other building types across Oldenburg.

Homogeneous grid cells of the census data with the same building types were selected in order to equally distribute the number of residents to each one of the buildings within that grid cell according to the relative volume proportion of the building (footprint area and number of storeys).

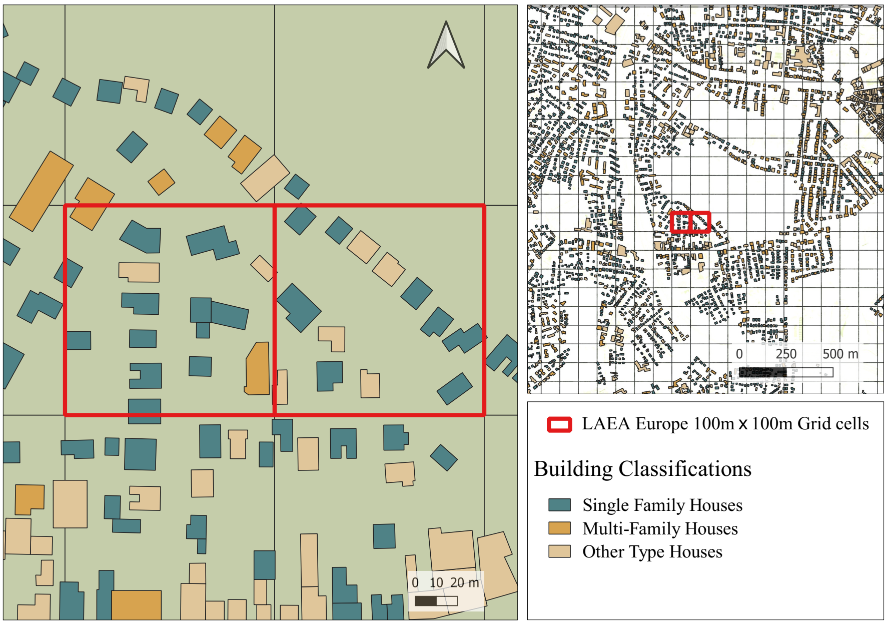


Figure 4-6: Different building types in Oldenburg City

It is important to note that census data contains a lot of private or secured data labelled as -1, these were not included in the models. These grid cells are considered sensitive due to potential discrimination and confidentiality concerns. Examples of sensitive areas include prisons and rehabilitation centres. The exclusion of these grid cells ensures the protection of an individual's privacy and prevents any discriminatory impact.

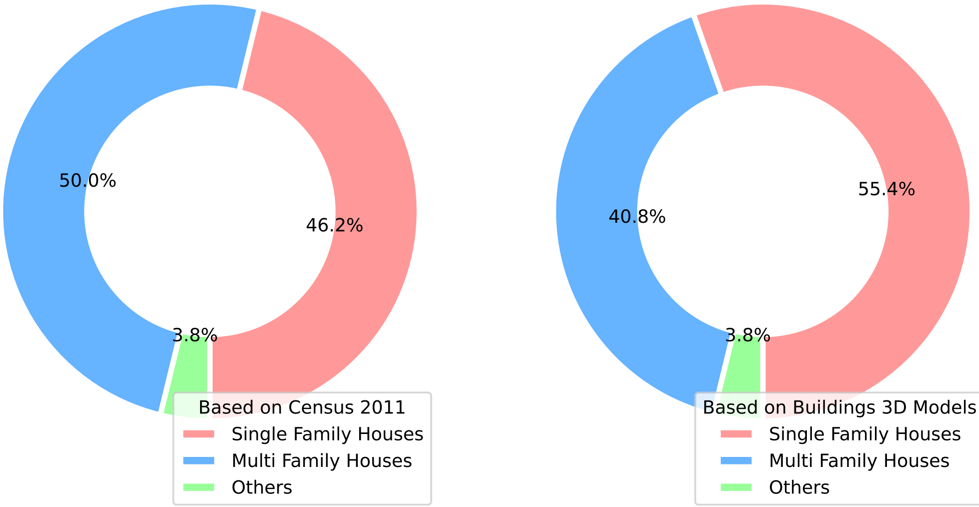


Figure 4-7: Pie plots illustrating building type classification based on Census 2011 and Buildings 3D Models dataset

4.4 Performance Optimization

To enhance the model performance and accuracy, model tuning is performed on the prepared classification model. Model tuning involves adjusting various parameters and fine-tuning to enhance the model's predictive capabilities.

Model tuning was carried out by using hyperparameter optimization techniques offered in Xgboost algorithm. The XGBRegressor class offers several hyperparameters that can be tuned to optimize the model's performance. GridCV search was implemented for the model tuning. The hyperparameters tuned in this study are discussed in detail in the Appendix chapter 7 section 7.2.

4.5 Classification Model

In search of an effective approach, supervised learning techniques were explored. Initially, logistic regression was used to classify individual building occupants. However, the results obtained from logistic regression did not meet our expectations in terms of accuracy and performance.

As a result, other supervised machine learning algorithms, specifically logistic regression and Xgboost were explored. For the evaluation of both the algorithms, mean-squared error metrics were used and Xgboost showed better performance as shown in figure 4-8 and hence it was utilized.

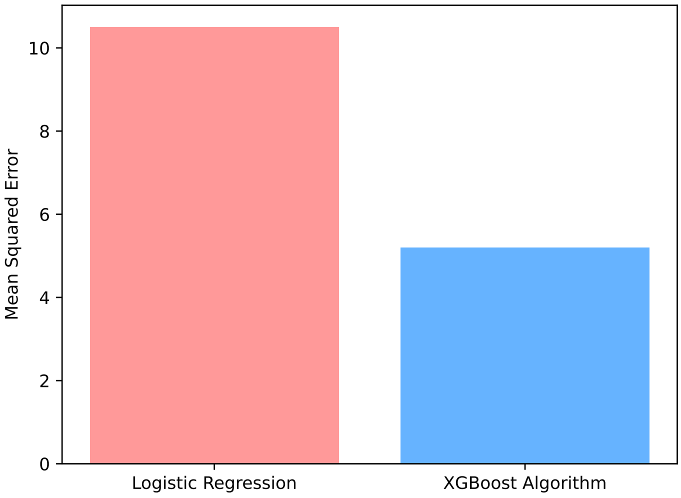


Figure 4-8: MSE on Logistic regression And XGboost algorithms

Xgboost is explained in detail in the section 3.2. Xgboost is an ensemble learning method that combines the power of multiple decision trees to make more accurate predictions. Three separate models were developed and trained each pertaining to each building type (SFH, MFH and Other building types) and their training datasets as shown in figure 4-9, figure 4-10 and figure 4-11.

The model's objective was to allocate residents per grid cell to the buildings within the heterogeneous grid cells, considering the proportional building area and a number of storeys.

The model takes into consideration the number of buildings within each grid cell and their

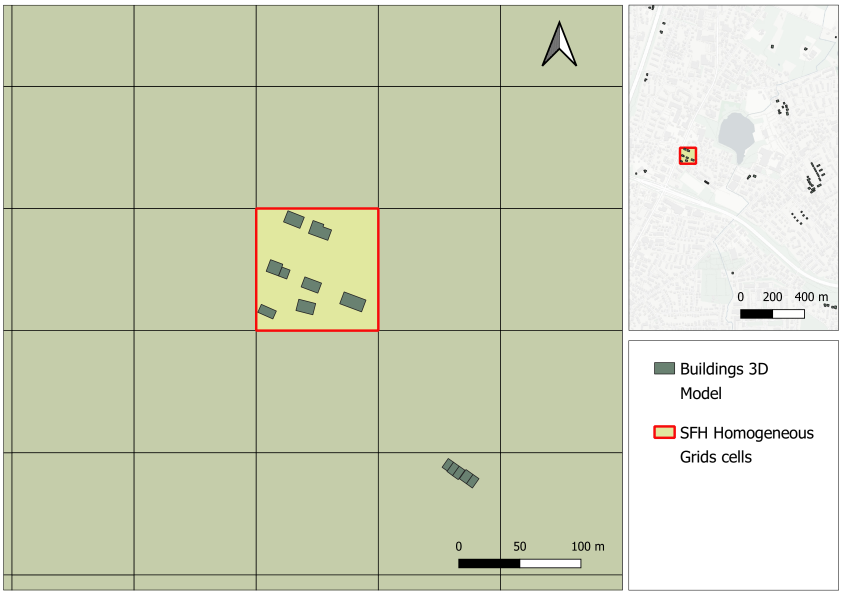


Figure 4-9: SFH Homogeneous grids used as training dataset

geometric characteristics (3D CityGML LoD2 models), the number of stories of each building within each grid cell (assuming that each floor is approximately 3.0 m and the number of residents at the grid cell level.

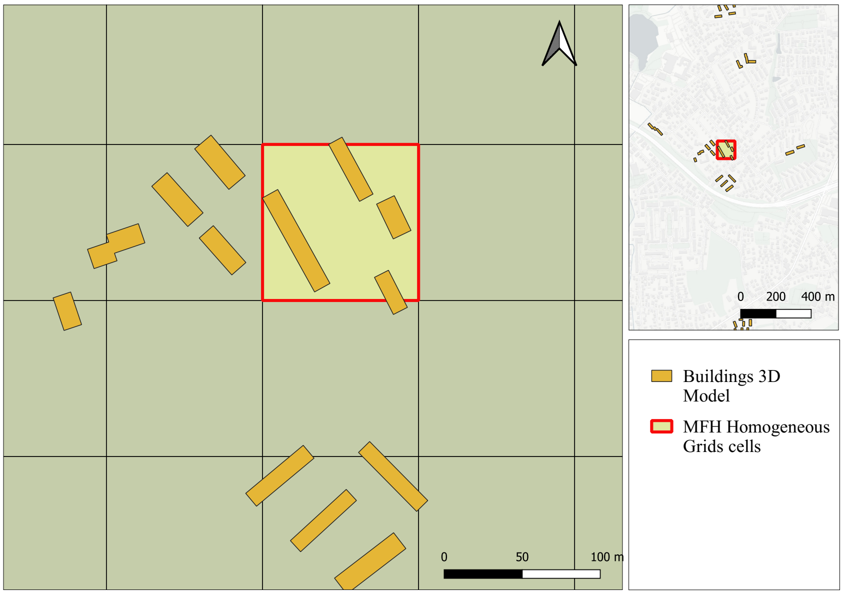


Figure 4-10:MFH Homogeneous grids used as training dataset

The selection of Xgboost as our preferred algorithm was based on the comparison of its results and metrics with those of other models.

Model's accuracy can be evaluated at the grid cell level, providing a measure of the precision in predicting the number of residents for each building within the respective grid cell. Model evaluation using various metrics is discussed in detail in the results chapter 5.

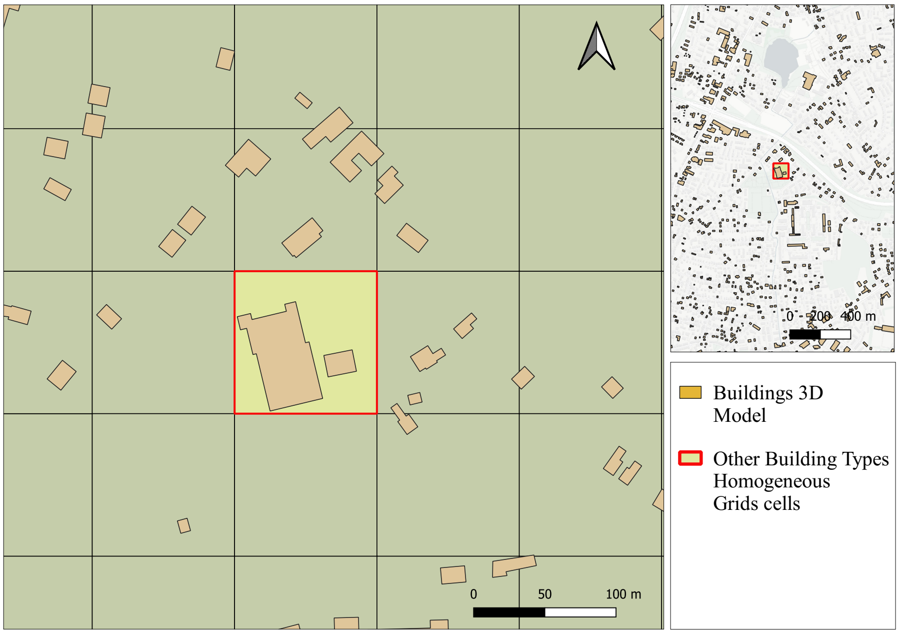


Figure 4-11:Other Building Type Homogeneous grids

4.5.1 Summary

The proposed approach is further explained through a flowchart, which provides a visual representation of the detailed steps involved in the implementation of the approach. This flowchart as shown in figure 4-12 serves as a roadmap, guiding through the sequential and interconnected stages of data integration, preparation of training dataset, model development, and tuning.

Overall, the combination of these key components and the outlined steps in the flowchart form a comprehensive and systematic approach to address the objectives of the study, enabling the accurate classification of building population estimation based on integrated Census and Buildings 3D models data.

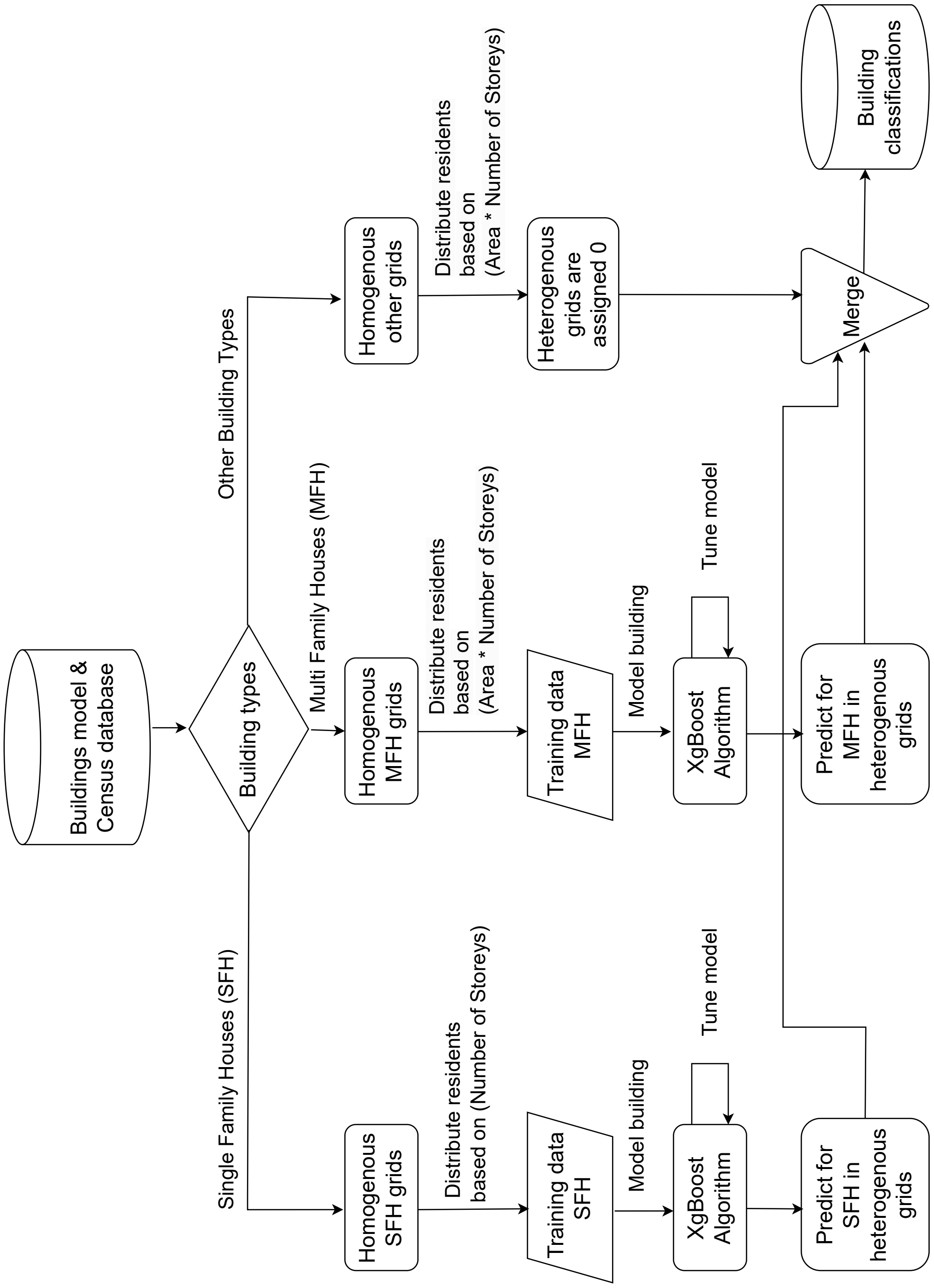


Figure 4-12: Workflow of methods and datasets used in this study

5 Results

In this section, the results obtained for each section are discussed in the methodology chapter 4. Section 5.1 discusses the final databank structures. Section 5.2 defines the main results for the synthetic training datasets by illustrating comparisons with the primary dataset. Lastly, section 5.3 and section 5.4 discuss the results obtained from the developed and tuned model.

5.1 Databank Preparation

As discussed in section 4.2 in this study, two databanks have been prepared, databank 1 integrating census population, housing and buildings and the other specialized databank 2 integrating the previously mentioned census datasets and Buildings 3D models.

The databank 1 has 1783987 records and 74 columns with Grid id serving as a unique identifier. A detailed overview of the structure of the final databanks is given in Appendix 7 table 7-1 and table 7-2. It consists of characteristics of all three census datasets thereby providing details information about the buildings in the grid, heating types (e.g., central heating, floor heating) used across the grid, type of households per grid (e.g., 1-Person Household, 2-Person Household), the number of different building types (e.g., SFH, MFH, Other) across the grid and lastly the number of residents per grid.

The databank 2 has 56750 records and 46 columns containing grid cell level building geometric information such as height, area, surface area, number of storeys along with the number of people living on a grid.

These databanks were created using Sqlite in Python which provides serverless storage and faster retrieval of large datasets as discussed in section 3.3 and section 3.3.

The creation of a databank helps faster extract, load and transformation of information and can also be used for further analysis tasks related to census datasets. Additionally, it facilitates data management and enhances data processing efficiency thereby opening possibilities for deeper analysis.

5.2 Creation of Training Datasets

As discussed in section 4.3 in the absence of training data due to data privacy laws. Synthetic training datasets for the classes SFH, MFH and other building types were prepared. The training data for SFH was prepared by distributing the residents per grid considering the proportion of the building area.

Consequently, the training data for MFH and other building types were prepared by distributing the residents per grid considering the proportion of building area and the number of storeys as shown in section 4.5.1 and figure 4-12.

Figure 5-1 illustrates the comparison of the training dataset and the original dataset for SFH. The histogram in blue shows the distribution of residents per grid based on the

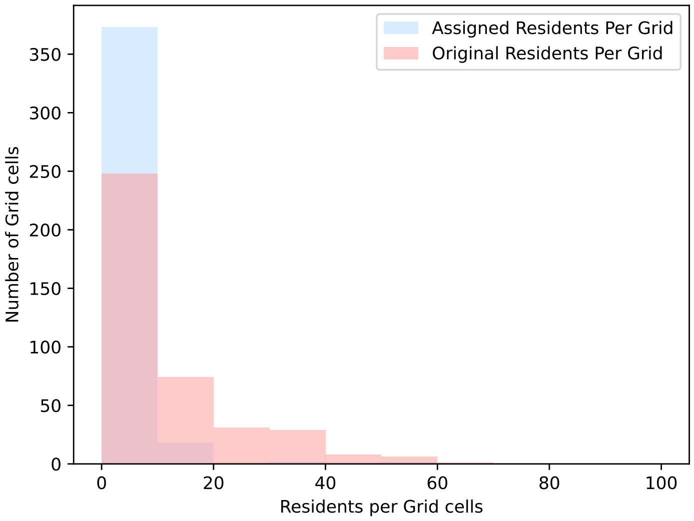


Figure 5-1: Resident distribution in original and assigned grids for SFH training data

assigned residents(as of building area proportion). The histogram in salmon pink shows the distribution of residents per grid based on the original residents (as of census data). It can be seen that the distribution of assigned residents is concentrated between 0 to 20 residents per grid. On the other hand, the distribution of assigned residents per grid cell is concentrated between 0 to 40 residents per grid. This highlights that for the training dataset most of the grids have between 0 to 20 residents although the distribution based on the census data, is still concentrated towards the lower end but there are grids that have a larger number of residents as compared to the training dataset.

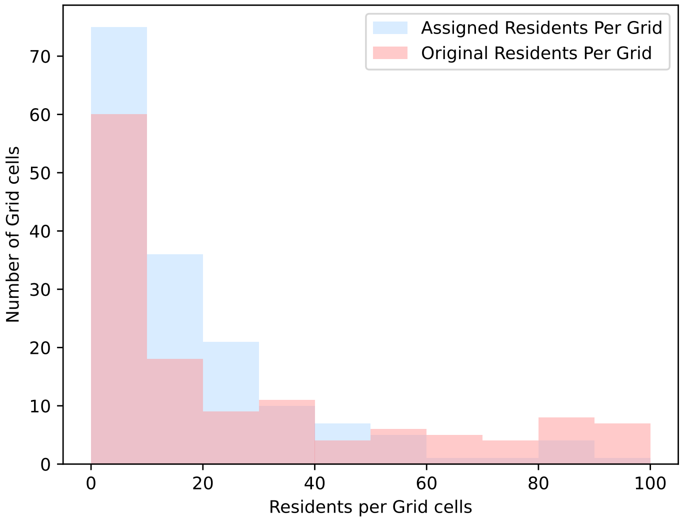


Figure 5-2: Resident distribution in original and assigned grids for MFH training data

Figure 5-2 illustrates the comparison of the training dataset and the original dataset for MFH. The histogram in blue shows the distribution of residents per grid based on the assigned residents(as of building area and the number of storeys proportion). The histogram in salmon pink shows the distribution of residents per grid based on the original residents (as of census data). It can be seen that the distribution of residents in both cases is concentrated between 0 to 100 residents per grid. This indicates that the maximum number of residents per grid is consistent in both cases.

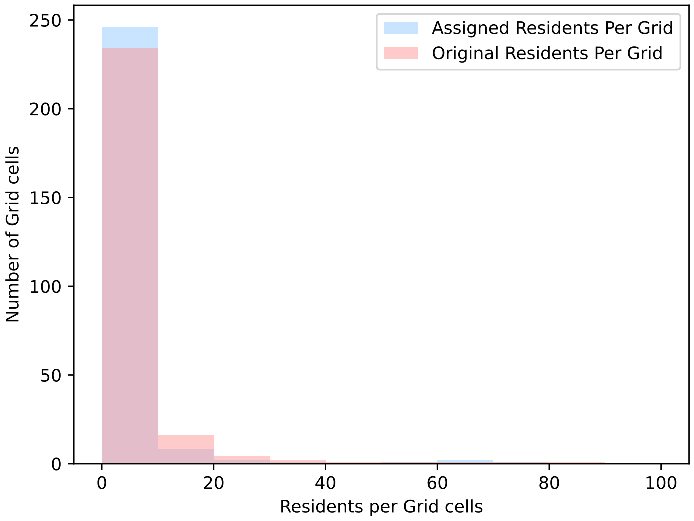


Figure 5-3: Resident distribution in original and assigned grids for other building types training data

Figure 5-3 illustrates the comparison of the training dataset and the original dataset for other building types including trade, commercial and service buildings. The histogram in blue shows the distribution of residents per grid based on the assigned residents(as of building area and the number of storeys proportion). The histogram in salmon pink shows the distribution of residents per grid based on the original residents (as of census data). It can be seen that the distribution of assigned residents closely follows the original resident's distribution. However, the maximum number of residents per grid varies i.e., for training dataset and original dataset are 20 and 40.

Overall, it is evident from the three histograms that the majority of grids in the training dataset have a relatively low number of residents. However, for grids between 0 to 20 residents, the training dataset is able to capture the distribution of the original dataset. This is important because according to the original dataset the average number of residents per grid is 30. Hence, the trained models follow the average distribution for the original residents' dataset.

5.3 Performance Optimization

In this study, the model is fine-tuned with a grid search method. The details of each hyperparameter used are discussed in section 4.4. The tuned values of these hyperparameters are discussed in Appendix 7 Table 7-3.

To compare the optimization results, the R2 score and Mean Squared Error(MSE) metrics were used. The R2 score defines how well the developed model fits the dataset. It ranges between 0 to 1 and a R2 score closer to 1 indicates a better fit. Figure 5-4 highlights the values of the R2 score before and after performing the model tuning.

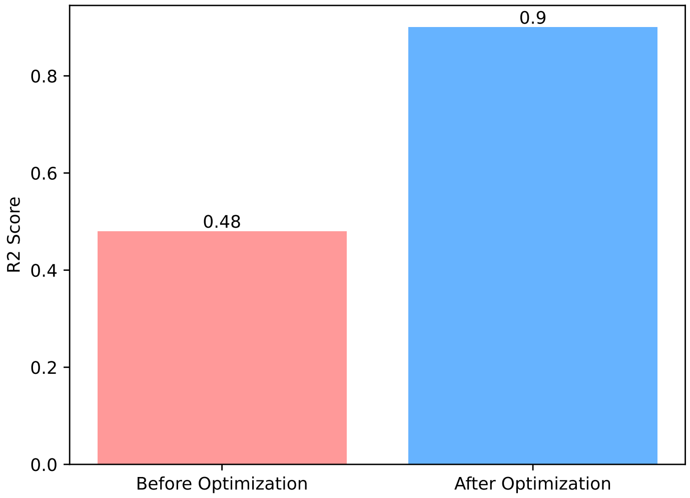


Figure 5-4: XGBoost R2 score before and after hyperparameter optimization

Additionally, MSE measures the average squared difference between the predicted values and the actual values. A lower MSE value indicates a better fit. Figure 5-5 shows a drastic drop in the MSE value after tuning the model. Therefore, it is evident that the model is better fit after tuning.

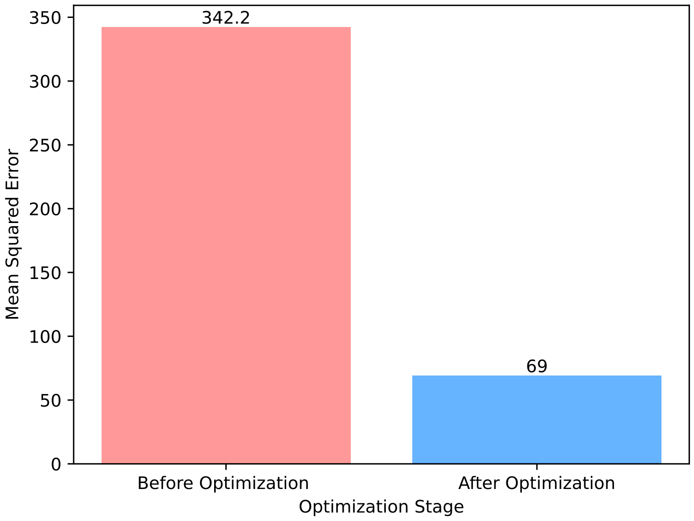


Figure 5-5: Bar plot illustrating the XGBoost MSE before and after hyperparameter optimization

5.4 Classification Model

The model's accuracy can be evaluated at the grid cell level as defined in equation 5-1, providing a measure of the precision in predicting the number of residents for each building within the respective grid cell.

$$\text{Accuracy per grid} = \left(100 - \left|\frac{\Sigma R_{\text{predicted}} - \Sigma R_{\text{original}}}{\Sigma R_{\text{original}}}\right|\right)$$

where,
$R_{\text{predicted}}$ represents the predicted building residents per grid
R_{original} represents the original residents per grid

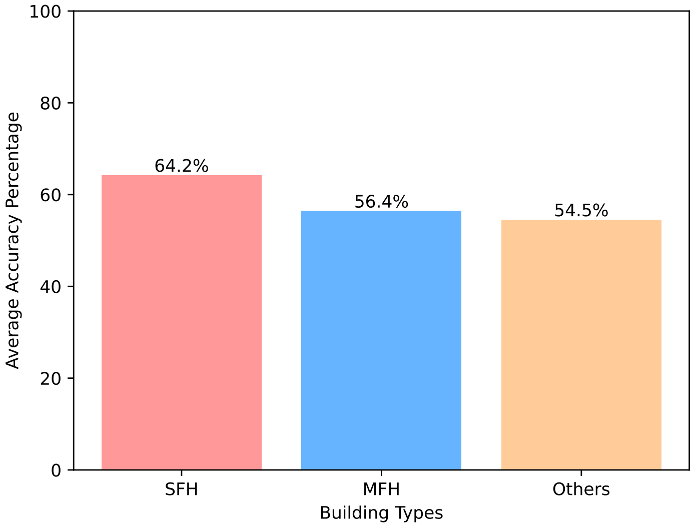


Figure 5-6: Bar plot illustrating the average prediction accuracy over different building types

Figure 5-6 highlights the average accuracy percentage for SFH, MFH and other building types. The overall accuracy is observed to be higher than 50% in all the building types with the highest accuracy observed in the case of SFH. This is important as SFH occupy the highest share of building types in Oldenburg which is evident from figure 4-7.

Figure 5-7 illustrates the comparison of predicted and the original residents per grid across Oldenburg. The histogram in blue shows the distribution of predicted residents while the histogram in salmon pink shows the distribution of original residents per grid cell.

It can be seen that the distribution of predicted residents is concentrated between 0 to 25 residents. However, the original dataset covers larger distribution, with the maximum number of residents per grid at 100. The inconsistencies in the datasets as discussed in section 4.1 and irregular distribution of building types in the Census and Buildings 3D models mainly due to building type misclassifications as can be seen from 4-7 are the major causes of such variation in distributions.

It was observed that the performance of the synthetic training dataset closely followed

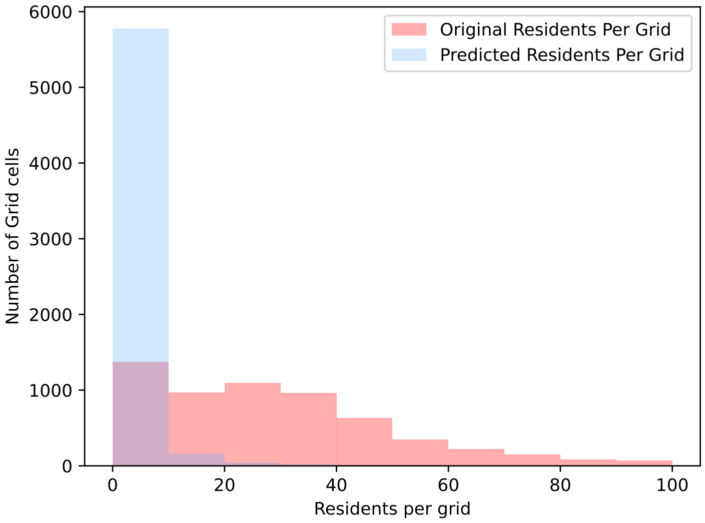


Figure 5-7: Plot illustrating the distribution of original and predicted residents per grid cell across Oldenburg

the distribution of primary datasets with minor deviations(due to already discussed inconsistencies) from the distribution curve. The main reason for the good performance of the synthetic training sets is as they were based on the calculating proportion of resident grids and distributing them among the buildings of the grid as can be seen in figure 4-12. However, when the final model was built integrating the training sets which comprised minor deviations, it resulted in high deviations from the original distribution.

Additionally, it was observed that the performance of the model is tied to the accuracy of the building type classification. A better building type classification will lead to a better prediction of the number of residents. Overall, the accuracy of grid cells spans from 10% to 100%, with an average of 62.4%. Approximately, 25% of all buildings were classified with an accuracy of 80% or higher.

In conclusion, census data provides useful information that can be used to parameterize building energy models. This study shows the possibilities of using statistical, machine learning and GIS-based models in order to classify buildings and generate a detailed national building stock while still complying with data privacy laws. The results obtained can be enhanced by developing a larger machine-learning model that incorporates and learns from more parameters.

6 Summary and Outlook

This section summarizes the primary findings and conclusions of the study and outlines the potential directions for future research and development.

6.1 Conclusion

This study plays a significant role in providing relevant information for the development of the energy map for Germany, which is crucial considering the ongoing energy crisis in the country. Focusing on the understanding of residents per building provides valuable insights into the trends and patterns of individual energy needs. This knowledge is essential for making informed energy management decisions and strategies, thereby leading to effective and efficient energy usage.

Two primary datasets, Census 2011 and Buildings 3D models, were used. The Census 2011 databank consists of three main datasets: Population, Households, and Buildings. The Census 2011 collection process and quality checks for the gathered data were also discussed. This data is published in the INSPIRE-compliant 100m grid format with the highest resolution of $100\,\mathrm{m} \times 100\,\mathrm{m}$ grid cells. The aim of the study is to utilise this $100\,\mathrm{m} \times 100\,\mathrm{m}$ information of grids and build a classification model for Oldenburg.

In the beginning to obtain a classification model, unsupervised learning techniques such as clustering using K-means were explored. In this method, an optimal number of clusters were identified for the dataset under consideration using elbow methods, and scatter plots of different building types helped to analyse the distribution of residents per grid. However, the results were not satisfactory, and therefore, a new approach was chosen for further analysis.

This study used a supervised machine learning algorithm and a GIS-based model for individual building population estimations. In the absence of training data due to data regulation policies, synthetic training datasets were prepared for different building types (SFH, MFH and other building types). For the number of residents, a gradient-boosted decision tree model was developed and tuned. The model shows an average accuracy of 62%, where the models are closely related to the building form classification. It is important to note that census data containing private or secured data labelled as -1, were not included in the models. These grids include areas such as prisons, rehabilitation centres, asylums and many others which are considered sensitive due to potential discrimination and confidentiality concerns.

A general visualization of the classification results for all of the models is found in figure 6-1, where for a specific area of Oldenburg, all buildings are classified into the respective parameters. In conclusion, this study shows that census data offers valuable insights that can be used to parameterize building energy models.

This study highlights the potential of utilizing statistical, machine learning, and GIS-based models to classify buildings and create a comprehensive national building stock, all while adhering to data privacy regulations. The findings of the study can be further improved by

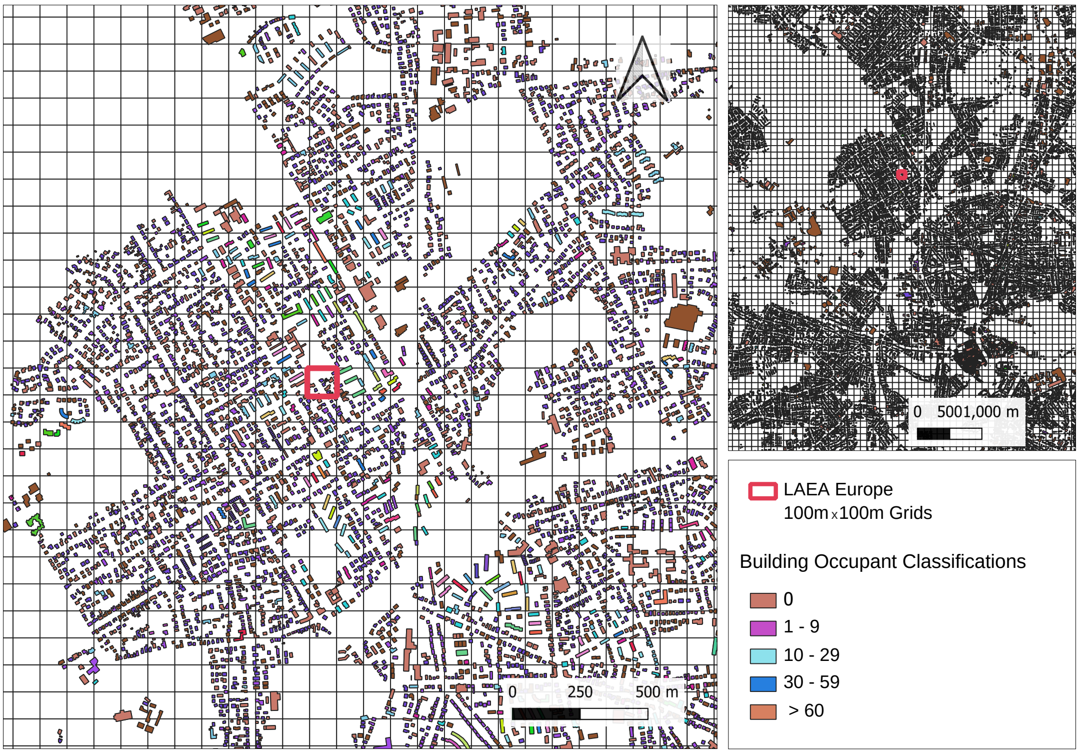


Figure 6-1: Building occupant classification results

developing a more extensive machine learning model that incorporates and learns from a wider range of parameters. By expanding the model's scope and incorporating additional relevant features, a more accurate and detailed classification of buildings can be achieved, leading to enhanced insights and a deeper understanding of the building stock.

6.2 Future Work

This study makes a significant contribution to the development of the energy map for Germany, which is of utmost importance given the current energy crisis in the country. A key aspect of this study involves understanding the number of residents per building, as it provides valuable insights into individual trends and patterns. This understanding, in turn, enables better decision-making and strategies for energy management.

The availability of comprehensive training datasets plays a crucial role in training the model and improving its performance. In this study, the importance of obtaining high-quality training data to enhance the accuracy of the model's predictions is recognized. By incorporating a wide range of building samples with different characteristics, the model becomes more robust and capable of accurately classifying buildings in various scenarios.

The inclusion of a training dataset ensures that the model learns from a wide spectrum of building types, architectural styles, and geographical variations and thereby making the model generalize better and make accurate predictions on unseen data.

By emphasizing the significance of training datasets and their impact on the model's performance, this study highlights the need for ongoing efforts to collect and curate high-quality data. A continuous focus on improving the training datasets will ultimately lead to more reliable and accurate building classifications, thereby facilitating better energy management decisions and strategies.

In this study, the performance of the developed model relies heavily on the accurate classification of building types. A more precise classification leads to improved model performance. Utilizing Census 2022, datasets instead of Census 2011 have been estimated to produce better results than the former census. The main reason for this is while both censuses use register-based data to estimate the population of the country. However, the difference comes in the way the inaccuracies in the register data are dealt with. In Census 2022, it was found that the need for adjustments in smaller municipalities was greater than in larger municipalities therefore random sampling in municipalities of all sizes, both large and small is to be used to correct any inaccuracies introduced during the upcoming census [Off20-ol]. This new approach could lead to better and more precise classifications than Census 2011.

Another approach to obtain a more precise classification can be of using high-resolution aerial images introduced by Würm [WDS+21] and this methodology can be extended by integrating these images with the primary datasets. Thereby one of the main inconsistencies in the classification of buildings can be addressed. This integration will result in better performance of the model and improved classification accuracy.

In summary, this study not only contributes to the energy map development for Germany but also emphasizes the significance of accurate building classification. By leveraging high-resolution images and integrating them with primary datasets, the model's performance is enhanced, leading to improved building classification and consequently, more reliable energy management strategies.

7 Appendix

7.1 Final Databank Structures

This section elaborates on the data integration results of this study. The different parameters are shown, followed by their descriptions and labels. Table 7-1 describes the databank structure for Germany comprising different census datasets joined based on grid id acting as a unique identifier. In the table 7-2, Building coordinates in table 7-2, (N, S, E, W) refer to the North, South, East and West coordinates. The final databank for Oldenburg consists of Table 7-1 and Table 7-2 as its structure. The results presented here are to give an intuition of how the databanks and their components are structured.

Table 7-1: Databank structure for Germany comprising census buildings, households and population

Census type	Parameter	Description	Pred. Label
Buildings census	Grid Id	Unique identifier	-
	Building type	Buildings with living space Residential buildings Residential buildings (excluding dormitories) Dormitory Other buildings with living space	1 2 3 4 5
	Building age	Before 1919 1919–1948 1949–1978 1979–1986 1987–1990 1991–1995 1996–2000 2001–2004 2005–2008 2009 and later	1 2 3 4 5 6 7 8 9 10
	Type of Ownership	Community of apartment owners Private person Housing cooperative Municipality or municipal housing company	1 2 3 4

Continued on next page

Table 7-1 – Continued

Census type	**Parameter**	**Description**	**Pred. Label**
		Private housing company	5
		Other private company	6
		Federal or state government	7
		NGO	8
	Building form	Detached Semi-detached Terraced house	SFH
		Detached Semi-detached Terraced house 3–6 apt. 7–12 apt. > 13 apt. Other building types	MFH
	Number of rooms	1 Room	1
		2 Rooms	2
		3 Rooms	3
		4 Rooms	4
		5 Rooms	5
		6 Rooms	6
		7 Rooms or more	7
	Total buildings	Number of buildings	> 0
Households census	Type of private household (by family)	Single person households	1
		Couples without child	2
		Couple with child	3
		Single parent households	4
	Type of private household (living arrangement)	Single person households	1
		Married couples	2
		Registered partnerships	3
		Single mothers	4
		Single fathers	5
	Size of private household	1 Person	1

Continued on next page

Table 7-1 – Continued

Census type	Parameter	Description	Pred. Label
		2 Person	2
		3 Person	3
		4 Person	4
		5 Person	5
		6 Person or more	6
	Total households	Number of households	> 0
Population census	Number of residents	Number of people	> 0
		Uninhabited or secret	-1

Table 7-2: Databank Structure for Oldenburg comprising different census datasets and buildings 3D models

Parameter	Description	Pred. Label
Building Id	Unique identifier	-
Area	Area of the building	> 0
Height	Height of the building	> 0
Category	Residential buildings	1
	Service buildings	2
	Industrial buildings	3
	Trade buildings	4
	Other buildings	5
Type	SFH	1
	MFH	2
	Others	3
Storeys	Number of storeys	$(0-5)$
Volume	Volume of the building	> 0
Perimeter	Perimeter of the building	> 0
Surface Area	Surface area of the building	> 0
Geometry	Building coordinates	(N, S, E, W)
Centroid	Building centroid coordinates	(x, y)

7.2 Model Optimization Parameters

In this section, the hyperparameters tuned during model tuning are discussed. Table 7-3 illustrates the parameter values used in this study. The results presented here are to give an intuition of the configuration setting. The following are the hyperparameters used in the study:

1) **Maximum depth**: This parameter is used to define the maximum depth of the tree in the algorithm. The higher the maximum depth value, the higher the probability of capturing complex relationships, but it also increases the risk of model overfitting.

2) **Minimum loss reduction**: This parameter defines the minimum loss reduction (gamma) required to do a further split in the tree. This parameter defines the minimum threshold for creating new tree nodes. A higher value results in fewer splits, leading to a more conservative model. At the same time, lower values make the model more complex by allowing more splits.

3) **Training instances**: This parameter refers to the fraction of training instances (subsample) that are randomly sampled to train each tree in the XGBoost ensemble. It helps prevent overfitting by controlling the sampling of the training data.

4) **Number of estimators**: This parameter determines the number of trees (or estimators) to be created in the model. Each tree contributes to the final prediction, and increasing the number of trees can improve model performance. However, a higher number of estimators also increases the computational complexity and the training time.

5) **Evaluation metrics**: This metric is used for data validation and consists of a variety of different parameters such as root mean squared error (RMSE), mean absolute error (MAE) and many more. In this study, we have used RMSE with an objective function of squared error for validation.

Table 7-3: Hyperparameters of XgbRegressors used in this study

Hyperparameters of XgbRegressor	**Values**
Maximum Depth	7
Learning Rate	0.3
Number of Estimators	300
Gamma	1
Subsample	0.8
Objective Function	squared error
Evaluation Metrics	RMSE

Acronyms

AI Artificial Intelligence

CNN Convolutional Neural Network

CRS Coordinate Reference System

ETRS European Terrestrial Reference System

GBDT Gradient-Boosted Decision Tree

GBM Gradient Boosting Method

GML Geography Markup Language

IEA International Energy Agency

ISO International Organization for Standardization

LAEA Lambert Azimuthal Equal Area

LOD Level-Of-Detail

MFH Multi-Family Houses

ML Machine Learning

OSM Open Street Map Projects

QGIS Quantum Geographical Information System

RMC Regression Model Checking

SFH Single Family Houses

SQL Structured Query Language

UML Unified Modeling Language

WCSS Within-Sum of Clusters

Bibliography

References

[AGA+19] Al-Rakhami, M.; Gumaie, A.; Alsanad, A.; Alamri, A.: An Ensemble Learning Approach for Accurate Energy Load Prediction in Residential Buildings. IEEE Transactions on Smart Grid 1, 2019

[Aga13] Agarwal, S.: Data Mining: Data Mining Concepts and Techniques. In: 2013 International Conference on Machine Intelligence and Research Advancement, 2013, pp. 203–207

[Age20] Agency, I. E.: Energy Efficiency 2020, 2020, Unter: https://www.iea.org/reports/energy-efficiency-2020

[Alt22-ol] AltexSoft: Introduction to Semi-Supervised Learning. Unter: https://www.altexsoft.com/blog/semi-supervised-learning/, Mar. 18, 2022

[ANM+08] Azhar, S.; Nadeem, A.; Mok, j.; Leung, B.: Building Information Modeling (BIM): A New Paradigm for Visual Interactive Modeling and Simulation for Construction Projects. In: Aug. 2008

[AP22] Azimi, S.; Pahl, C.: The Impact of Data Completeness and Correctness on Explainable Machine Learning Models, Feb. 2022

[AS19] Asniar; Surendro, K.: Predictive Analytics for Predicting Customer Behavior. In: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 230–233

[ASH+21] Arens, S.; Schlüters, S.; Hanke, B.; Maydell, K. von; Agert, C.: Monte-Carlo Evaluation of Residential Energy System Morphologies Applying Device Agnostic Energy Management. IEEE Access PP, Dec. 2021, pp. 1–1

[ASV+21] Ana, M. G.; Schmiedt, J.; Verma, T.; Lefter, I.; Korthals Altes, W.; Droin, A.; Schiricke, B.; Wurm, M.: Spatial factors influencing building age prediction and implications for urban residential energy modelling. Computers, Environment and Urban Systems 88, 2021, p. 101637

[BBK+14] Behkamal, B.; Bagheri, E.; Kahani, M.; Sazvar, M.: Data accuracy: What does it mean to LOD. In: 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE), 2014, pp. 80–85

[Bha18-ol] Bhatt, S.: Reinforcement Learning 101. Unter: https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292, May 19, 2018

[Bla22] Blanco, L.: Analysis of Phase Shift Data of Atomic Force Microscopy on Solar Glass with Different Functional Surfaces and Solar Collectors after Desert Exposure. Dissertation, University of Freiburg, 2022

[BLS16] Biljecki, F.; Ledoux, H.; Stoter, J.: An improved LOD specification for 3D building models. Computers, Environment and Urban Systems 59, 2016, pp. 25–37

[BP15] Bhosale, S.; Patil, P.: International Journal of Computer Science and Mobile Computing SQLite: Light Database System. International Journal of Computer Science and Mobile Computing 44, Apr. 2015, pp. 882–885

[Bro16] Brownlee, J.: Monte Carlo Sampling for Probability, Mar. 2016

[BRS19] Butwall, M.; Ranka, P.; Shah, S.: Python in Field of Data Science: A Review. In: 2019, pp. 20–24

[Bur-ol] Burns, E.: Machine Learning. Unter: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

[BV12] Baffour, B.; Valente, P.: An evaluation of census quality. Statistical Journal of the IAOS 28, Mar. 2012, pp. 121–135

[CBP+15] CHICCO, G.; BOSIO, F. de; PASTORELLI, M.; FANTINO, M.: Clustering-based performance assessment of thermal energy management in buildings. In: 2015 IEEE International Telecommunications Energy Conference (INTELEC), 2015, pp. 1–6

[CCK+19] COCHRAN, J. K.; COX, L. A.; KESKINOCAK, P.; KHAROUFEH, J. P.: Prescriptive analytics: Optimizing decisions in real-time. INFORMS Journal on Applied Analytics 49, INFORMS, 2019, pp. 321–337

[CG16-ol] CARTOGRAPHY, F. A. for; GEODESY: Geographical Grid System of Germany. Unter: https://gdz.bkg.bund.de/index.php/default/inspire/sonstige-inspire-themen.html, Feb. 19, 2016

[CLA+19] CSOKNYAI, T.; LEGARDEUR, J.; AKLE, A. A.; HORVÁTH, M.: Analysis of energy consumption profiles in residential buildings and impact assessment of a serious game on occupants' behavior. Energy and Buildings 196, 2019, pp. 1–20

[CMM21] CHHABRA, A.; MASALKOVAITĖ, K.; MOHAPATRA, P.: An Overview of Fairness in Clustering. IEEE Access 9, 2021, pp. 130698–130720

[CPC19] CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S.: Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 8, 2019

[CR19] CICHY, C.; RASS, S.: An Overview of Data Quality Frameworks. IEEE Access 7, 2019, pp. 24634–24648

[CWA+22] CHEN, D.; WANG, B.; ADEEL, M.; YANG, Y.; KE, J.: Integrated attention mechanism for GBDT building energy consumption prediction algorithm. In: 2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AHPCAI), 2022, pp. 222–227

[CWY+21] CHEN, H.; WU, B.; YU, B.; CHEN, Z.; WU, Q.; LIAN, T.; WANG, C.; LI, Q.; WU, J.: A New Method for Building-Level Population Estimation by Integrating LiDAR, Nighttime Light, and POI Data. Journal of Remote Sensing 2021, 2021

[DDG+22] DARWAZEH, D.; DUQUETTE, J.; GUNAY, B.; WILTON, I.; SHILLINGLAW, S.: Review of peak load management strategies in commercial buildings. Sustainable Cities and Society 77, 2022, p. 103493

[ED20-ol] ECONOMIC CO-OPERATION, T. O. for; DEVELOPMENT: Germany's energy policy landscape. Unter: https://www.iea.org/reports/germany-2020, Feb. 1, 2020

[Eur23-ol] EUROSTAT: POpulation Grids. Unter: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_grids, May 15, 2023

[FS15] FOTACHE, M.; STRIMBEI, C.: SQL and Data Analysis. Some Implications for Data Analysits and Higher Education. Procedia Economics and Finance 20, Dec. 2015

[Gau21] GAUTAM, S.: Role of Data Analysis in Higher Education. In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021, pp. 741–744

[GEV+21] GARBASEVSCHI, O. M.; ESTEVAM SCHMIEDT, J.; VERMA, T.; LEFTER, I.; KORTHALS ALTES, W. K.; DROIN, A.; SCHIRICKE, B.; WURM, M.: Spatial factors influencing building age prediction and implications for urban residential energy modelling. Computers, Environment and Urban Systems 88, 2021, p. 101637

[GFL18] GATTAL, A.; FAYCEL, A.; LAOUAR, M.: Automatic Parameter Tuning of K-Means Algorithm for Document Binarization. In: Dec. 2018, pp. 1–4

[GL23-ol] GEOINFORMATION, S. O. for; (LGLN), S. S. L. S.: 3D BUILDING MODEL (LOD2). Unter: https://opengeodata.lgln.niedersachsen.de/#lod2, May 7, 2023

[GS23-ol] GEOINFORMATION, S. O. for; SAXONY, S. S. L.: 3D Building Models LOD2. Unter: https://opengeodata.lgln.niedersachsen.de/, May 15, 2023

[HAK+17] HANEEM, F.; ALI, R.; KAMA, N.; BASRI, S.: Descriptive analysis and text analysis in Systematic Literature Review: A review of Master Data Management. In: 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), 2017, pp. 1–6

[Hea21-ol] HEAZEL, C.: OGC City Geography Markup Language (CityGML) 3.0 Conceptual Model Users Guide. Unter: https://docs.ogc.org/guides/20-066.html, Sept. 13, 2021

[IBM-ol] IBM: Machine Learning. Unter: https://www.ibm.com/topics/machine-learning

[Jai23-ol] Jain, A.: Mastering XGBoost Parameter Tuning. Unter: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/, May 19, 2023

[KB09] Klimešová, D.; Brozova, H.: DSS and GIS in Knowledge Transformation Process. Journal on Efficiency and Responsibility in Education and Science 2, June 2009

[KG19] Körner, T.; Grimm, E.: Towards a register-based census post 2021 in Germany. Conference of European Statisticians 1, 2019, pp. 1–11

[KSS+18] Kranjac, M.; Sikimić, U.; Salom, J.; Tomic, S.; Bulajić, S.: Visualization of smart specialisation process using QGIS tools. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1444–1448

[LLL+23] Liu, H.; Liang, J.; Liu, Y.; Wu, H.: A Review of Data-Driven Building Energy Prediction. Buildings 13, 2023

[Mar16-ol] Marr, B.: A Short History of Machine Learning. Unter: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=43c7a72415e7, Feb. 19, 2016

[Mar21-ol] Marzell, T.: Types of Machine Learning Algorithms. Unter: https://rocketloop.de/en/blog/clustering-machine-learning-comprehensive-guide/s, May 27, 2021

[NAM+17] Nemirovsky, D.; Arkose, T.; Markovic, N.; Nemirovsky, M.; Unsal, O.; Cristal, A.: A Machine Learning Approach for Performance Prediction and Scheduling on Heterogeneous CPUs. In: 2017 29th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), 2017, pp. 121–128

[NG19] Nagpal, A.; Gabrani, G.: Python for Data Analytics, Scientific and Technical Applications. In: 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 140–145

[Nvi23-ol] Nvidia: XGBoost. Unter: https://www.nvidia.com/en-us/glossary/data-science/xgboost/, June 16, 2023

[Off20-ol] Office, F. S.: Census 2022 and Census 2011. Unter: https://www.destatis.de/EN/Press/2020/12/PE20_Z01_122.html, Dec. 15, 2020

[Off23-ola] Office, F. S.: Quality assurance 2: check for multiple registrations in the population registers. Unter: https://www.zensus2011.de/EN/2011Census/Methodology/Methodology_Survey_to_clarify_discrepancies_node.html, May 4, 2023

[Off23-olb] Office, F. S.: Register data as the basis of the 2011 Census. Unter: https://www.zensus2011.de/EN/2011Census/Methodology/Methodology_Register_data_node.html, May 3, 2023

[Off23-olc] Office, F. S.: The census of buildings and housing: an important basis for planning the future. Unter: https://www.zensus2011.de/EN/2011Census/Methodology/Methodology_Census_of_buildings_and_housing_node.html, May 3, 2023

[Off23-old] Office, F. S.: The household survey. Unter: https://www.zensus2011.de/EN/2011Census/Methodology/Methodology_The_household_survey_node.html, May 4, 2023

[Off23-ole] Office, F. S.: The survey on residential establishments. Unter: https://www.zensus2011.de/EN/2011Census/Methodology/Methodology_Survey_in_residental_establishments_and_collective_living_quarters_node.html, May 4, 2023

[OS13] O'Neil, C.; Schutt, R.: Doing Data Science. O'Reilly Media, Inc., 2013 – isbn: 9781449358655

[PNS19] P.N.Siva jyothi, R. Y.: A Review on Python for Data Science, Machine Learning and IOT. In: 2019

[Pre13] Prewitt, K.: The Importance of the Census and the Future of the American Community Survey. Public Opinion Quarterly 77, 2013, pp. 377–387

[Pri23-ol] Priyadarshini: Machine Learning. Unter: https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning, Mar. 10, 2023

[PRM21] PUGLIESE, R.; REGONDI, S.; MARINI, R.: Machine learning-based approach: global trends, research directions, and regulatory standpoints. Data Science and Management 4, 2021, pp. 19–29

[Pyt21-ol] PYTHON SOFTWARE FOUNDATION, J.: Python Developers Survey 2021 Results. Unter: https://lp.jetbrains.com/python-developers-survey-2021, Sept. 1, 2021

[RA12] RODRÍGUEZ-BUSTOS, C.; APONTE, J.: How Distributed Version Control Systems impact open source software projects. In: 2012 9th IEEE Working Conference on Mining Software Repositories (MSR), 2012, pp. 36–39

[Raj22-ol] RAJ, R.: Types of Machine Learning Algorithms. Unter: https://www.enjoyalgorithms.com/blog/classification-of-machine-learning-models, Apr. 19, 2022

[Sah23-ol] SAHA, S.: XGBoost vs LightGBM: How Are They Different. Unter: https://neptune.ai/blog/xgboost-vs-lightgbm, Apr. 25, 2023

[Sar20] SARKAR, S.: Regression in Machine Learning, Mar. 14, 2020

[SDA+22] SHARMA, A.; DEVALIA, D.; ALMEIDA, W.; PATIL, H.; MISHRA, A.: Statistical Data Analysis using GPT3: An Overview. In: 2022 IEEE Bombay Section Signature Conference (IBSSC), 2022, pp. 1–6

[SFL+21] SCHUG, F.; FRANTZ, D.; LINDEN, S. van der; HOSTERT, P.: Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates. PLoS ONE 16, 2021, e0249044

[SG17] SMITH JR., C. B.; GREENBERG, F. S.: Energy Management Principles and Practice. Springer International Publishing, 2017

[Sim19] SIMON ROUCHIERA Maria Jose Jimenezb, S. C.: Sequential Monte Carlo for on-line parameter estimation of a lumped building energy model. Energy and Buildings 187, 2019, pp. 86–97

[SK16] SCHOLZ, R.; KREYENFELD, M.: The Register-based Census in Germany: Historical Context and Relevance for Population Research. Comparative Population Studies 41, Aug. 2016

[Sol23-ol] SOLANKI, J.: Data Consistency, Definition, examples and best practice. Unter: https://www.decube.io/post/what-is-data-consistency-definition-examples-and-best-practice, Apr. 1, 2023

[SPK18] SCHACHINGER, D.; PANNOSCH, J.; KASTNER, W.: Adaptive learning-based time series prediction framework for building energy management. In: 2018 IEEE International Conference on Industrial Electronics for Sustainable Energy Systems (IESES), 2018, pp. 453–458

[SPP+21] SIMA, C. A.; POPESCU, M. O.; POPESCU, C. L.; ALEXANDRU, M.; POPA, L. B.; DUMBRAVA, V.; PANAIT, C.: Energy Management of a Cluster of Buildings in a University Campus. In: 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2021, pp. 1–6

[SSR16] S, K.; S, V.; R, R.: A comparative analysis on linear regression and support vector regression. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1–5

[Ste17] STEFAN DITTRICH, T. K.: The combined census model in Germany – origins, lessons learned and future perspectives. Conference of European Statisticians 49, United Nations Economic Commission for Europe, 2017, pp. 321–337

[Ste23-ol] STEVENS, E.: The 7 Most Useful Data Analysis Methods and Techniques. Unter: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/#data-analysis-techniques, May 10, 2023

[Sus12] SUSTAINABLE DEVELOPMENT, W. B. C. for: Energy Efficiency in Buildings: Business Realities and Opportunities, 2012

[Tah22] TAHERDOOST, H.: Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects. International Journal of Academic Research in Management, 2022, New York, SSRN, 2022

[TD11] TURNER, W. C.; DOTY JR., S.: Energy Management Handbook. The Fairmont Press, Inc., 2011

[Tea22-ol] TEAM, G. L.: Types of Clustering Algorithms. Unter: https://www.mygreatlearning.com/blog/clustering-algorithms-in-machine-learning/, Oct. 21, 2022

[Tea23-ol] TEAM, G. L.: Types of Analytics. Unter: https://www.analytics8.com/blog/what-are-the-four-types-of-analytics-and-how-do-you-use-them, Mar. 21, 2023

[Tha23-ol] THAKUR, D.: Difference between Data and Information. Unter: https://ecomputernotes.com/fundamental/information-technology/what-do-you-mean-by-data-and-information, Apr. 9, 2023

[Tuk62] TUKEY, J. W.: The Future of Data Analysis. The Annals of Mathematical Statistics 33, Institute of Mathematical Statistics, 1962, pp. 1–67

[ÜEG+12] ÜRGE-VORSATZ, D.; EYRE, N.; GRAHAM, P.; HARVEY, D.; HERTWICH, E.; JIANG, Y.; KORNEVALL, C.; MAJUMDAR, M.; MCMAHON, J. E.; MIRASGEDIS, S.; AL., et: Energy End-Use: Buildings. In: Global Energy Assessment: Toward a Sustainable Future. Cambridge University Press, 2012, pp. 649–760, DOI: 10.1017/CBO9780511793677.016

[UHS11] URAL, S.; HUSSAIN, E.; SHAN, J.: Building population mapping with aerial imagery and GIS data. International Journal of Applied Earth Observation and Geoinformation 13, 2011, pp. 841–852

[WDS+21] WURM, M.; DROIN, A.; STARK, T.; GEISS, C.; SULZER, W.; TAUBENBÖCK, H.: Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling. ISPRS International Journal of Geo-Information 10, 2021, p. 2

[YDR09] YANG, G.; DWYER, M. B.; ROTHERMEL, G.: Regression model checking. In: 2009 IEEE International Conference on Software Maintenance, 2009, pp. 115–124

[Zen23-ol] ZENSUS: Zensus. Unter: https://www.zensus2011.de/EN/Home/home_node.html, Mar. 21, 2023

[ZPH+13] ZHOU, R.; PAN, Y.; HUANG, Z.; WANG, Q.: Building Energy Use Prediction Using Time Series Analysis. In: 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications, 2013, pp. 309–313