A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making

Usman Ali[a], Mohammad Haris Shamsi[a], Mark Bohacek[b], Karl Purcell[b], Cathal Hoare[a], Eleni Mangina[c], James O'Donnell[a]

[a]*School of Mechanical and Materials Engineering and UCD Energy Institute, UCD, Dublin, Ireland*
[b]*Sustainable Energy Authority Of Ireland, Dublin, Ireland*
[c]*School of Computer Science and UCD Energy Institute, UCD, Dublin, Ireland*

Abstract

Urban planners, local authorities, and energy policymakers often develop strategic sustainable energy plans for the urban building stock in order to minimize overall energy consumption and emissions. Planning at such scales could be informed by building stock modeling using existing building data and Geographic Information System-based mapping. However, implementing these processes involves several issues, namely, data availability, data inconsistency, data scalability, data integration, geocoding, and data privacy. This research addresses the aforementioned information challenges by proposing a generalized integrated methodology that implements bottom-up, data-driven, and spatial modeling approaches for multi-scale Geographic Information System mapping of building energy modeling. This study uses the Irish building stock to map building energy performance at multiple scales. The generalized data-driven methodology uses approximately 650,000 Irish Energy Performance Certificates buildings data to more than 2 million buildings' energy performance. In this case, the approach delivers a prediction accuracy of 88% using deep learning algorithms. These prediction results are then used for spatial modeling at multiple scales from the individual building level to a national level. Furthermore, these maps are coupled with available spatial resources (social, economic, or environmental data) for energy planning, analysis, and support decision-making. The modeling results identify clusters of buildings that have a significant potential for energy savings within any specific region. Geographic Information System-based modeling aids stakeholders in identifying priority areas for implementing energy efficiency measures. Furthermore, the stakeholders could target local communities for retrofit campaigns, which would enhance the implementation of sustainable energy policy decisions.

Keywords: GIS modeling, building energy performance, data-driven approaches, urban building energy modeling, machine learning, urban planning

Email addresses: usman.ali@ucdconnect.ie (Usman Ali), mohammad.shamsi@ucdconnect.ie (Mohammad Haris Shamsi), mark.bohacek@seai.ie (Mark Bohacek), karl.purcell@seai.ie (Karl Purcell), cathal.hoare@ucd.ie (Cathal Hoare), eleni.mangina@ucd.ie (Eleni Mangina), james.odonnell@ucd.ie (James O'Donnell)

Nomenclature

AHP	Analytical Hierarchy Process
API	Application Programming Interface
BEH	Better Energy Homes
BEM	Building Energy Modeling
BER	Building Energy Rating
BPD	Building Performance Database
CEA	City Energy Analyst
$CityBES$	City Building Energy Saver
CRS	Coordinate Reference Systems
CSO	Central Statistics Office
$DEAP$	Dwelling Energy Assessment Procedure
$EPBD$	European Union Energy Performance of Buildings Directive
EPC	Energy Performance Certificate
GHS	GreenHouse Gas
GIS	Geographic Information System
$HVAC$	Heating Ventilation, and Air Conditioning
$LIDAR$	Light Detection and Ranging
LOF	Local Outlier Factor
$MCDA$	Multi-Criteria Decision Analysis
$MGRS$	Military Grid Reference System
$SEAI$	Sustainable Energy Authority of Ireland
SVM	Support Vector Machines
SVR	Support Vector Regression
$UBEM$	Urban Building Energy Modeling
$UEUM$	Urban Energy Use Modeling

UMI	Urban Modeling Interface
UTM	Universal Transverse Mercator
WLC	Weighted Linear Combination

1. Introduction

Building energy consumption plays a significant role in global energy supply and demand. In the building sector, energy consumption has dramatically increased over the past few years, mainly due to population growth [1]. Any further increase in energy demand will significantly increase global GreenHouse Gas (GHG) emissions that would have a significant impact on global climate change. Several opportunities exist in the building sector to reduce energy demand and emissions, thereby promoting a sustainable environment. The world has seen a major shift towards the global exchange of building energy efficiency policies, data, and performance analysis. According to IEA Efficient World Strategy report, buildings in 2040 could be nearly 40% more energy-efficient than today. In Europe, around 35% of the buildings are more than 50 years old, and 75% of the buildings exhibit inefficient energy performance [2]. One possible solution to improve building energy performance is retrofitting existing buildings to be more energy-efficient. However, based on the current trend of European energy policies, only 0.4 to 1.2% of building stock in Europe is retrofitted each year [3].

Planning and implementation of large scale sustainable energy systems pose significant challenges for stakeholders due to the complexities. Due to rapid growth in building data availability, there are opportunities to analyze existing building data and develop strategic and efficient energy planning. However, systematic approaches are required for integrating available energy and planning data. One possible solution for large scale building energy analysis is through a spatial analysis of energy data by using Geographic Information System (GIS) modeling [4]. This approach has been extensively used for regional, urban, and national planning [5] and is one of the primary tools to present large geographical scale data in a visual format. GIS provides a framework for gathering, managing, and analyzing large scale data in a geographic context. Visual representation of data in a GIS system can help the stakeholders to perform qualitative and quantitative analysis for support decision making [6].

GIS-based energy planning requires extensive data to make an energy policy decision [4]. Individual building analysis is often difficult on a large scale due to the limited availability of data and users' privacy issues [7]. One of the most promising solutions for building energy analysis with limited information can be accomplished through building stock modeling [8]. However, majority of studies focus on developing building stock models without considering aspects that integrate spatial information for decision-making processes [9].

Generally, building stock modeling at a large scale takes two approaches, namely, engineering based and data-driven modeling [10]. Engineering-based approaches use building archetypes that represent various dwelling types of building stock to calculate the energy use using numerical simulation models [11]. However, existing urban energy modeling studies

often rely on aggregated building data and henceforth, do not account for a fine grained analysis of building characteristics [8]. The data-driven models use historical building stock data to build relationships between input and output data using statistical or machine learning techniques [12]. This approach is beneficial when limited historical data is available. However, existing studies focus on traditional statistical techniques and only a limited number of studies implement machine learning techniques at large scale using spatial features.

Urban planners, local authorities, and energy policymakers are often required to conduct energy planning and analysis at the district or neighborhood-scale. While national level authorities often find it difficult to coordinate large disparate sources of individualized information, local authorities do not have access to building stock data outside the concerned area of authority. As previous building energy modeling studies mostly focus on national or city-scale analysis for energy policy planning [13], these strategies are therefore not adequately addressed within local or regional level detailed analysis. As such, the local authorities are not wholly informed when making energy policy decisions in their locality as energy planning is often not adequately addressed within local or regional level planning structures. [6]. Furthermore, existing energy modeling approaches lack spatial information for detailed GIS-based analysis at multiple scale.

There are several challenges associated with the implementation of multiple scale GIS-based building energy modeling that include: (1) data availability, (2) data inconsistencies, (3) data scalability (4) data integration (5) geocoding and, (6) data privacy issues. Building energy performance data is typically unavailable for the entire spatial area. Moreover, due to inconsistencies in available large scale energy data and lack of scalable building energy mapping approaches, a gap persists between building energy modeling and traditional planning practices [14]. Stakeholders face scalability issues because of the requirement that energy planning be implemented at a national level. Similarly, integration issues exist in large scale GIS mapping for planning and analysis because the available data is sparse, inconsistent, diverse, and heterogeneous [15]. The available data does not provide complete coverage and is of unknown quality. Unfortunately, most of the building stock survey data are not geocoded for GIS mapping. Furthermore, data privacy is also a significant challenge for granular level GIS mapping of results [6]. Therefore, a robust GIS-based modeling approach is required that would help in predicting the energy performance of the entire building stock data using limited resources for complex decision analysis.

This study introduces a generalizable bottom-up data-driven approach for multi-scale GIS-based mapping of residential building energy performance. Previous studies often devise non-scalable frameworks suited for a particular application. The methodology described in this research is generalizable and scalable and henceforth, could be applied to existing available building stock data. Furthermore, the devised methodology is integrated with a novel data-driven solution to support geocoding of building stock data for GIS mapping. The bottom-up data-driven approach predicts building energy performance using available limited building stock data. The methodology, further, compares different supervised machine learning algorithms to determine the optimal data-driven building energy model for large scale implementation.

The novelty of this study includes the implementation of feature engineering and de-

termines the optimum features for data-driven model development, thus, significantly enhancing model accuracy. Moreover, the proposed approach implements a spatial aggregation approach to determine the energy performance at the neighborhood, district, city, and county levels. The methodology further couples the predicted results with available spatial resources (social, economic, or environmental data) for planning and decision making using the Multi-Criteria Decision Analysis (MCDA) approach. Overall, this study derives a novel integrated approach to help local authorities analyze residential sector energy consumption and CO_2 emissions at different geographical scales ranging from local to national levels. This research demonstrates the implementation of the methodology for the residential building stock of Ireland.

The study introduces a novel integrated scalable approach to implement building stock modeling that includes a combination of bottom-up, data-driven, and spatial modeling approaches. As the approach is generalizable, further studies could be conducted to ensure the applicability to different national databases. The potential impact of this study on the international academic community is vast as it provides a guideline to ensure the implementation of data-driven approaches is as per the data analytics standards. For instance, extraction of optimum features would aid the model development process of energy rating prediction.

This paper is structured as follows: Section 2 describes an overview of existing work done in GIS mapping and building energy performance prediction; Section 3 describes the devised methodology, including an explanation of the different steps followed in the GIS-based mapping of multi-scale model development; Section 4 states the results of Irish case study followed by Section 5 that includes discussions about possible implications and improvements in case study. Section 6 includes conclusions and potential challenges and future work.

2. Literature Review

GIS-based building stock models can be effectively used to develop and optimize urban scale sustainable energy planning. GIS-based modeling involves the use of data from varied sources. The associated GIS modeling approaches differ on the based of available and required data, as described in the following sections.

2.1. GIS-based Data Modeling

Building data required for energy modeling comprise three main categories, namely, simulated, benchmark, and measured data. Simulated data are generated from engineering-based building energy modeling tools such as EnergyPlus [16], Modelica [17] and TRNSYS (a transient system simulation program) [18]. Benchmark data can be acquired from publicly-available datasets available for researchers to compare modeling results, and validate the models performance. Real data are gathered through census, survey, billings, energy meters, and environmental sensors [8]. Data-driven modeling often makes use of real data. Within the real data category, census data includes statistical building stock data at various scales (local, national, and international), while survey data involves additional sampling studies

of individual buildings within a defined population area. Building electricity and meter data can be available in the different granularity of the time-series (measurement frequencies), such as per minute, hourly monthly, and yearly. The use of these data depends upon applications such as forecasting, prediction, and energy use intensity (EUI) estimation [14].

The methodologies used to collect real building stock data vary by country. For instance, the United States Department of Energy maintains one of the largest building stock databases, the Building Performance Database (BPD), which includes information about residential and commercial building stock [19]. Similarly, each member state in the EU maintains its own EPC database containing essential building energy performance information about its building stock [20]. However, using available data for decision making is often challenging for stockholders (urban planners, local authorities, and energy policymakers) as the data is inconsistent, diverse, sparse, and heterogeneous [15].

The available data for energy modeling are typically of incomplete coverage and inadequate quality. For instance, any Energy Performance Certificate (EPC) dataset only represents a proportion of the entire building stock. Unfortunately, most of the survey data are not geocoded where geocoding is the process of transforming data into a location-based format. The users often do not follow a standardized format while collecting the building addresses. This unstructured address format introduces inconsistencies in GIS mapping [21].

There are two different ways for geocoding existing datasets, namely, geocoding Application Programming Interface (API) and a data-driven approach. Geocoding API is a commercial service provided by some of the leading mapping companies like Google, Economic and Social Research Institute (ESRI), and Bing. However, these services do not perform well if the data is unstructured and unformatted. Such services match the address based on predefined descriptive data. Furthermore, these services can be costly for geocoding large scale datasets. On the other hand, the data-driven approach implements fuzzy string matching algorithms for geocoding. This process is useful for survey-based and inconsistent data. This approach works effectively with complicated addresses and case-specific priorities. For instance, even when the address does not match correctly, this approach formulates results based on the user definition of minimum address matching criteria [21].

There are limited studies that implement geocoding using a data-driven approach. Among these, the majority of the research deals with enhancing the efficiency of address matching and, thereby, does not provide a generalized, scalable solution for different scenarios [22]. Several opportunities exist to extend the existing work for address pre-processing along with address matching [21].

2.2. GIS-based Building Energy Modeling

Building stock modeling at a large scale usually implements two approaches, namely, engineering and data-driven approaches [10]. The engineering approach uses detailed building physics to identify energy performance. These tools often require detailed inputs about geometric and non-geometric properties of buildings; failure to provide accurate inputs can produce incorrect results. Henceforth, a massive amount of data would be needed to simulate an entire district. The use of building archetypes simplifies this approach by classifying the building stock using representative buildings. Several recent projects base on urban energy

modeling used the engineering approach, including City Building Energy Saver (CityBES) [23], Urban Modeling Interface (UMI) [24] and City Energy Analyst (CEA) [25].

These studies mostly use engineering methods with synthetic experimental data (Table 1). As engineering methods using archetypes implement a limited number of typologies, there are numerous assumptions and uncertainties embedded in energy simulations. These assumptions directly affect the accuracy of results and hence, limit the reliability of decision-making at large scale [8].

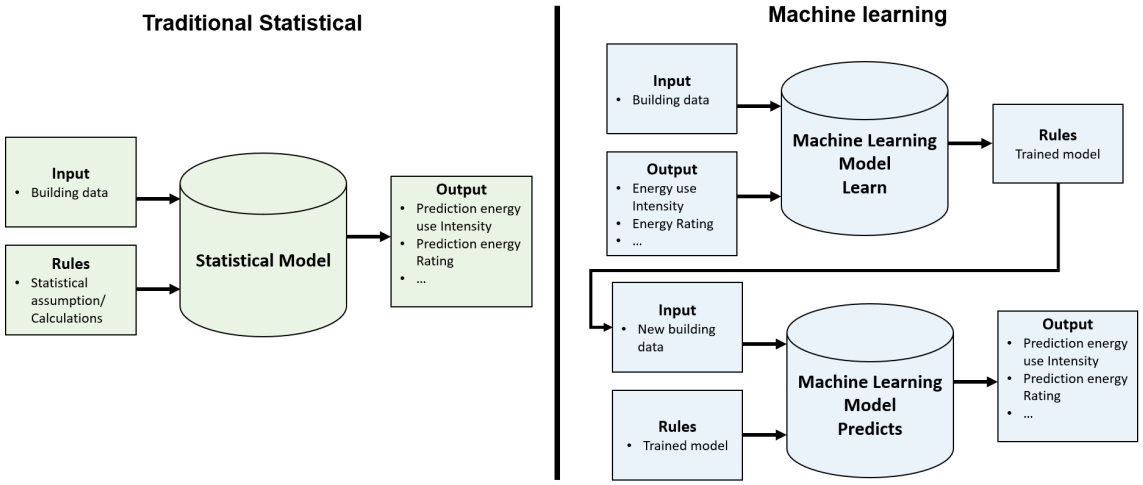

Figure 1: The methodological differences between traditional statistical and machine learning modelling techniques.

Data-driven approaches, on the other hand, do not require detailed knowledge about the building as these approaches estimate building energy performance based on historical data either using statistical or machine learning models [26]. While statistical models use sample data about buildings to build a mathematical relationship between the building's energy consumption and characteristics [11], machine learning models implement algorithms that learn from data to predict building energy performance with minimal assumptions [27]. The traditional statistics model takes input (building data) and pre-defined rules (statistical assumption/ calculations) to predict outputs such as energy use intensity and energy rating. On the other hand, the machine learning model comprises two steps. The first step uses inputs (building data), and outputs (energy use intensity and energy rating features) to train the learning model (trained model). The second step uses these rules (trained model) and model inputs (new building data) to predict the output (Fig. 1) [28]. As machine learning models can predict energy performance with limited information, these approaches have gained a lot of attention in the energy sector during the past few years [12]. Furthermore, these approaches often provide highest levels of accuracy using the available building energy usage data [14]. However, only a limited number of studies implement data-driven approaches at multiple scales using machine learning models (Table 1).

Generally, machine learning models implement either regression or classification algorithms [12]. Regression algorithms estimate real value (numerical or continuous) output variables, such as energy consumption. The most common regression algorithms include linear regression, decision trees, random forest, deep learning, generalized linear models, gradient boosted trees, and Support Vector Regression (SVR) [29]. Classification algorithms are effective when the output variable represents a designated label (discrete or categorical), such as energy rating or building type. Commonly used classification algorithms include the nearest neighbor, naive bayes, generalized linear model, logistic regression, deep learning, decision trees, random forest, gradient boosted trees, and Support Vector Machine (SVM), rule induction,and neural networks [30].

This study implements the data-driven approach using machine learning models for building energy modeling at multiple scales. The data-driven approach delivers robust energy modeling results when building stock data is available. Mostly, building energy data-driven studies focus on either single building energy use prediction or building clusters of limited typologies [31]. These studies implement traditional statistical models, namely, linear regression , multiple linear regression, non-linear regression, and conditional demand analysis [32]. The majority of these models rely on the nature of the data; model assumptions are far too strict and not representative of reality. To counter these limitations, machine learning models use techniques such as data pre-processing, feature selection, and cross-validation to improve the quality of data before generating system models.

Few studies implement GIS-based building energy modeling using machine learning models at a large scale (Table 1). For instance, Ma and Cheng devised a framework to estimate the building energy use intensity at the urban scale by integrating GIS and big-data technology [33]. Similarly, another study by Kontokosta and Tull formulated a data-driven predictive model to estimate the city-scale energy use in buildings [34]. It is worthwhile to mention that existing studies mostly focus on formulating an urban scale framework that uses synthetic data to generate models with a limited focus on GIS modeling. For instance, Nutkiewicz, Yang and Jain developed a framework for integrating engineering simulations (synthetic data) and machine learning methods in a multi-scale urban energy modeling workflow [13]. Similarly, Abbasabadi and Azari proposed an Urban Energy Use Modeling (UEUM) framework to model urban building and transportation energy using machine learning [31]. Several opportunities exist to extend the previous literature by introducing a generalized methodology for multi-scale modeling.

Table 1: Existing GIS-based building energy modeling studies to compare the scale, approaches, application and scope.

Paper	GIS	Scale	Approaches	Application	Scope	Context
[33]	2D	City	Machine Learning	Energy use intensity	C	United States
[35]	2D, 3D	Building, City	Engineering-based	Energy modeling	B	United States
[36]	2D	Building, City	Statistical	Energy consumption	R	Italy
[37]	3D	City	Engineering-based	Energy performance	C	Japan
[38]	2D, 3D	Building, City	Engineering-based	Retrofit analysis	C	United States
[39]	2D	Building, City	Statistical	Built environment energy use	R	Italy
[40]	2D	Building, City	Statistical	Energy consumption and solar potential	R	Italy
[41]	2D	Counties	Engineering-based	Energy demand	C	United States
[42]	2D	Neighbourhood, City	Statistical	Energy use	R	Korea
[5]	2D	City	Engineering-based	Optimal integration of flexibilisation technologies	B	Germany
[43]	2D	City	Engineering-based	Estimate wind and solar potential	C	UK
[44]	2D	Building, District	Engineering-based	Energy performance	C	China
[45]	3D	Building, City	Engineering-based	Community energy system design and operation	C	Global
[46]	3D	Building, City	Engineering-based	Energy modeling	R	Netherland
[47]	3D	Building, District	Engineering-based	Building simulation	B	Germany
[48]	2D, 3D	Building, City	Engineering-based	Energy modeling	R	UK
[49]	2D	Building, City	Engineering-based	Thermal building simulation	R	Germany
[50]	2D	Building, Neighbourhood	Engineering-based	Assessing energy profiles	R	Austria
[13]	2D, 3D	Building, City	Engineering-based, Machine Learning	Energy modeling	C	United States
[31]	2D	Building, City, Neighbourhood	Machine Learning	Energy use modelling	C	United States
[11]	2D	City	Statistical	Estimating energy savings	R	Netherlands
[51]	2D	Building, City, Neighbourhood	Engineering-based, Statistical	Heat consumption models	R	Netherlands
[52]	2D	City	Statistical	Energy consumption	B	United States
[53]	2D, 3D	Building, District, Neighbourhood	Statistical	Energy consumption patterns	B	Switzerland
[25]	2D, 3D	Building, District, Neighbourhood, City	Engineering-based	Analysis and optimization	B	Switzerland
[23]	3D	Building, City	Engineering-based	Building energy efficiency	C	United States
[24]	3D	Building, City	Engineering-based	Energy use	B	United States
[54]	2D	City	Engineering-based	building energy models	B	United States
[55]	2D	Building, Neighbourhood	Statistical	Energy inefficient residential properties	R	Ireland
[34]	2D	City	Machine Learning	Energy use	R	United States

Note: R, Residential; C, Commercial; B, Both Residential and Commercial;

3. Methodology

One significant challenge for urban planners and policy makers is to analyze and visualize large datasets and extract meaningful information from the data [6]. GIS-based modeling provides a framework for gathering, managing, and analyzing large scale data in a geographic context. Thus, GIS-based building energy modeling and planning helps to capture, store,and visualize in-depth information [6]. Hence, a generalized GIS-based methodology would allow for a wide variety of analyses, thereby, helping the stakeholders to maximize the analytical power of energy planning and modeling techniques [4].

The devised approach accounts for GIS-based building energy performance at multiple scales. The GIS-based mapping of multi-scale residential building energy performance follows seven steps (Fig. 2).

1. The initial step involves data collection from different resources (building stock, census, GIS and geographical data);
2. The next step focuses on geocoding of building stock data;
3. The pre-processing and feature selection step follows the geocoding procedure and employs data-driven approaches to improve the quality of the building stock data;
4. The next step, building archetypes development, uses pre-processed building stock data to identify archetypes representative of the building stock;
5. The data-driven model development step predicts building energy performance at large scale using a bottom-up approach;
6. The multi-scale GIS mapping step maps the building energy performance results; and
7. Finally, the energy planning step analyzes the modeling results for planning or decision making. This step analyzes and identifies the priority areas for implementation of long-term and sustainable energy related decisions.

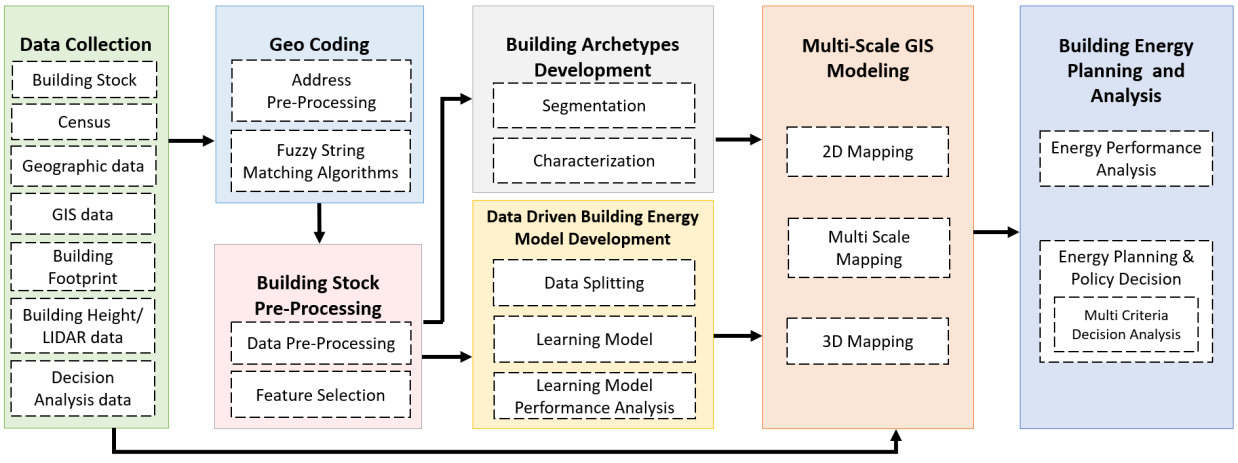

Figure 2: Overarching methodology for GIS-based mapping of multi-scale residential building energy performance using data-driven approaches

The following sections describe the individual steps of the methodology in further detail.

3.1. Data Collection

The data collection process gathers the datasets required for GIS mapping. These datasets include census, geographical, building geometry, and non-geometry information. The data collection process further merges the data from these resources, which can be represented as a visualization aid to inform energy policy decisions. At the large scale, existing building databases are often a major source of information about a building stock. These existing building stock databases could be in the form of a building energy certificates database that comprises geometric as well as non-geometric information. Geometric data consist of information about the building shape, building type, building fabric, number of floors, and window-wall ratios. Non-geometric building data includes envelope U-values, construction assemblies, and Heating Ventilation and Air Conditioning (HVAC) systems properties. Furthermore, energy consumption prediction also depends on building energy performance metrics. EPC data usually provides an overview of geometric and non-geometric information in addition to the building performance metrics. Furthermore, the dataset includes building quantification data (national statistics or census data) required to determine the number of buildings present in a specific area.

Similarly, 3D GIS data modeling requires building footprint and building height data. The most appropriate standard format model is a geospatial vector data format, also known as a shapefile. The building footprint and boundary data are usually available in a shapefile format that contains points, lines, and polygons. These data can be collected for the desired area from OpenStreetMap or national geography survey, which comprise geographical data of sufficient quality. Building height can be formulated as the product of the number of floors and the average building height for a specific area [39]. Light Detection and Ranging (LIDAR) data can also be used to infer the building height. However, LIDAR data is often unavailable at the district area level [39]. Finally, the geocoding process requires the national geographical database that contains the building address with spatial information.

3.2. Geocoding

This procedure follows data collection and involves the geocoding of building stock data. This study implements a data-driven approach that uses building stock and national geographic databases (Fig. 3). Partially geocoded building stock data is supplemented with the geographical dataset that includes the geocoded addresses of the residential building stock. To effectively reduce the search space, this process segments the dataset based on cities and counties. Segmentation increases the search accuracy by ensuring that the fuzzy string algorithms only use the search spaces where the address is located.

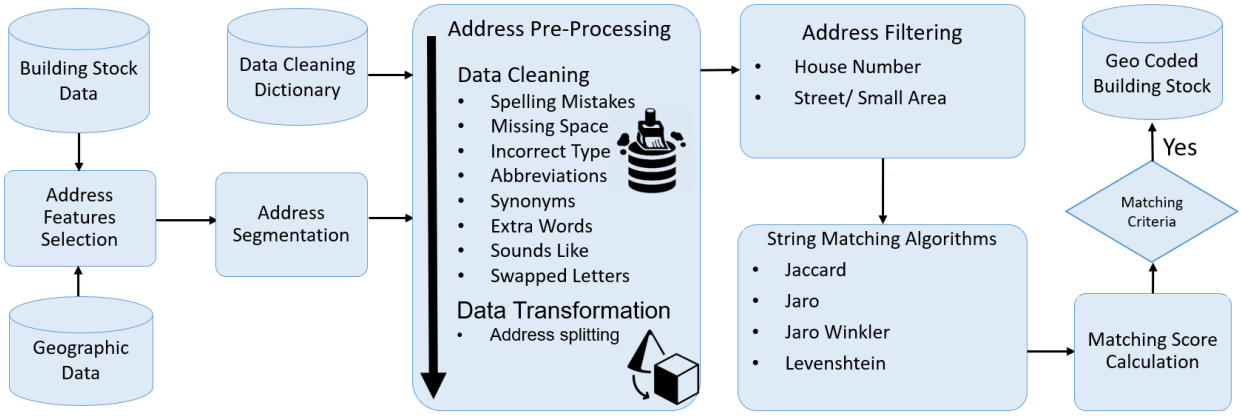

Figure 3: Residential building address geocoding process using data-driven approaches

As the data collected through surveys such as EPC data, generally contain irrelevant, incomplete, noisy, redundant, and inconsistent information, this process implements an address pre-processing procedure. The discrepancies mainly arise when the user does not follow a standardized procedure for reporting the addresses. Address pre-processing eliminates these inconsistencies using data cleaning and data transformation before the implementation of address matching algorithms. Data cleaning removes and replaces incorrect, incomplete, duplicate, and unstructured addresses with relevant words that increase the performance of a fuzzy string matching algorithm. The data cleaning process deals with spelling mistakes, missing spaces, incorrect types, abbreviations, synonyms, extra words, sounds like, and swapped letters (Table 2). The address cleaning process uses a data cleaning dictionary that comprises a list of predefined incomplete or irrelevant words along with their replacement option. Finally, the data transformation process extracts information such as street, county, city, and postal code from the addresses. This process further aids the address filtering process.

Table 2: Address cleaning task with set of examples for geocoding process

Cleaning task	Input Address	Ouput Address
Spelling Mistakes	123 Avvenue, Dublin	123 Avenue, Dublin
Missing Space	123Avenue, Dublin	123 Avenue, Dublin
Incorrect Type	123 Avenue Street, Dublin	123 Avenue, Dublin
Abbreviations	123 Ave Rd, DB	123 Avenue Road, Dublin
Synonyms	Baile Atha Cliath	Dublin
Extra Words	Top Apartment No 123 Court, Dublin	123 Court, Dublin
Sounds Like	123 Sqaare, Dubln	123 Square, Dublin
Swapped Letters	123 Avenue Raod, Dublin	123 Avenue Road, Dublin

Note: The above entries represent dummy addresses. Original addresses are not used due to privacy issues.

The geocoding process further implements address filtering at multiple levels, namely, house/ apartment number, street, and small area (cluster buildings nearby a street). The

geocoding process uses fuzzy string matching algorithms for address comparison with existing available national geocoded addresses databases. This study compares four different fuzzy matching algorithms including Jaro, Jaro-Winkler, Levenshtein, and Jaccard, based on a matching score. All of these string matching algorithms performed well for complex string matching based on existing literature [56, 57]. These algorithms can be mathematically formulated using equations 1, 2 and 3 respectively.

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{(m-t)}{m}\right), & \text{otherwise} \end{cases} \tag{1}$$

where $|s_i|$ is the length of the string s_i; m is the number of matching characters; t is half the number of transpositions. Two characters from s_1 and s_2 respectively, are considered matching only if they are the same and not farther than $\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1$. The number of transpositions defines as the half of matching characters that are not in the same index.

$$sim_{jw} = sim_j + \ell p\,(1 - sim_j) \tag{2}$$

where sim_j is the Jaro similarity for strings s_1 and s_2; ℓ is the length of common prefix at the start of the string up to a maximum of four characters; p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the similarity could become larger than 1. The standard value for this constant in Winkler's work is $p = 0.1$.

The Jaro-Winkler distance d_w is defined as $d_w = 1 - sim_{jw}$.

$$lev_{a,b}(i,j) = \begin{cases} max\,(i,j) & \text{if } min\,(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + k \end{cases} & \text{otherwise} \end{cases} \tag{3}$$

where $k = 0$, if $(a_i = b_i)$, 1 otherwise. $lev_{a,b}$ is the distance between the first i characters of a and the first j characters of b.

Jaccard is token based string matching algorithm. The calculation is to find the number of common string tokens and divide it by the total number of unique string tokens. Its expressed in the mathematical terms by in Equation (4).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4}$$

where, the numerator is the intersection (common string tokens) and denominator is union (unique string tokens).

The fuzzy string matching process matches the address based on two levels. The process initially matches the addresses based on house/apartment numbers at the individual building level. In the absence of house/apartment numbers, the process compares the addresses based on street names at the neighbourhood level. The matching process assigns scores between 0 to 1 to different string matching algorithms. These scores then determine the

least matching criteria to be considered as a geocoded address. The least matching criteria can be determined manually using a sample of the dataset. These geocoded addresses are then stored in the residential building stock database.

Selection of spatial projection often involves various reference coordinates that define the location of individual buildings in the stock. Geographical Coordinate Reference Systems (CRS) define spatial projection reference x, y points on the earths surface, such as longitude and latitude values. The common map projections in current use include the Universal Transverse Mercator (UTM) and the Military Grid Reference System (MGRS). The national geographic database usually contains spatial projection references (x, y coordinates) for addresses. Therefore, this study considers that the coordinate reference system is similar to the one used in the national geographic database while geocoding the addresses [21].

3.3. Building Stock Pre-Processing

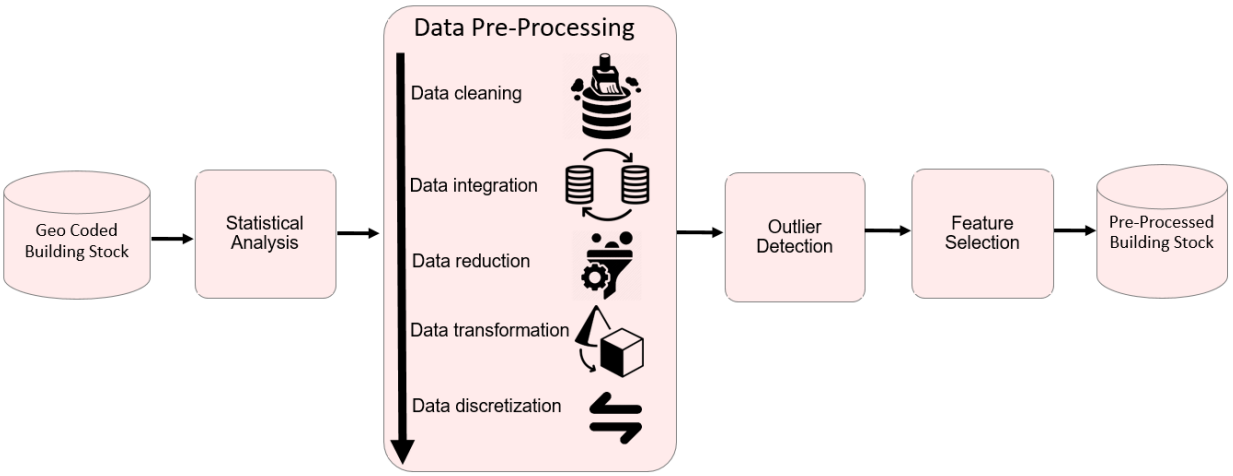

Figure 4: Building stock data pre-processing workflow to enhance the quality of building stock data

The building stock pre-processing employs four sequential steps, namely, statistical analysis, data pre-processing, outlier detection, and feature selection (Fig. 4). Statistical analysis aids in extracting initial inferences and summaries from data. This analysis involves the implementation of arithmetic operations (mean, median, and mode) and subsequent visual representations (histogram, density plots, charts). Data pre-processing involves transforming real-world or raw data into an understandable format. During pre-processing, the data goes through a series of operations such as data cleaning, data integration, data reduction, data transformation, and data discretization [14].

Outlier or anomaly detection is an essential step before implementing a learning algorithm. Outliers are observation points that lie at an abnormal distance from the majority of the other values in a data sample space. Generally, the outlier detection procedure implements distance-based, density-based, and Local Outlier Factor (LOF) methods [14].

The feature selection process identifies a subset of most relevant variables or attributes for the archetype representation and learning model development. This process removes irrelevant, redundant, and less important features that do not influence the learning model performance and thereby, reduce the input dimensionality, complexity, and computational load of the learning model [14].

Considered as one of the essential machine learning concepts that hugely impact learning accuracy, feature selection usually employs engineering or data-driven methods. Engineering methods use engineering judgment and existing practices in the literature [58]. Data-driven methods use various statistical approaches to develop learning models [14]. Generally, data-driven selection methods identify and rank features based on multiple statistical tests such as information gain, variance/standard deviation threshold, correlation coefficient, and chi-square tests [59]. For instance, the correlation coefficient filters those features that closely mirror the target feature. Similarly, a variance/standard deviation threshold filters the features that have the most or extremely different values. This study uses both engineering and data-driven methods to identify a subset of most relevant features. In the first step, the engineering method determines optimal features based on existing studies. In the next step, the data-driven selection method identifies features using multiple statistical tests. However, the type of feature selection depends upon the total number of features and data quality.

3.4. Building Archetypes Development

Building archetypes development requires two major sub-steps such as segmentation and characterization. The segmentation process determines the number of archetype buildings required to represent the residential building stock at multiple scales. There are various criteria for the segmentation of the building stock, for instance, building type, construction year, climate zone, or spatial information [7].

The characterization process determines the physical properties of each building archetype, such as building fabric, heating system, lighting, and hot water equipment [8]. This process estimates the values of the building archetype features on the basis of segmentation criteria using a data-driven approach. The segmentation criteria groups the data, and then performs the aggregation operation on each cluster to retrieve the properties of each archetype. The aggregation could be done by applying arithmetic or geometric mathematical operations (mean, median, or mode). The resulting aggregated value represents the characteristics of one building archetype.

This study generates the archetypes at the local level rather at the national or city level for fine-grained analysis using the average virtual building approach based on the statistical building data. Therefore at the local level, the formulated archetypes represent the entire cluster of buildings in that local area, which aids in the formulation of the entire building stock data. Local area archetypes provide a twofold advantage. Firstly these archetypes tackle the problem of data availability for modeling. Secondly, local level archetypes indirectly address the data privacy issues by using small areas for GIS mapping and model development.

3.5. Data Driven Building Energy Model Development

The large scale machine learning model development for building energy performance using a data-driven approach requires multiple steps (Fig. 5). The model development process uses pre-processed building stock data and begins with data splitting for training and testing purposes, followed by the implementation of learning algorithms. The process then analyzes the performance of the developed learning model.

Data splitting is the process of dividing the dataset into training and testing sets. The training dataset is a subset of the data that is used to develop the trained model. The testing dataset is a subset that evaluates a model to estimate the unbiased final performance of models. The most common approach for data splitting is using random data sampling, which splits the data randomly into 80%-20% split for training and testing, respectively [14].

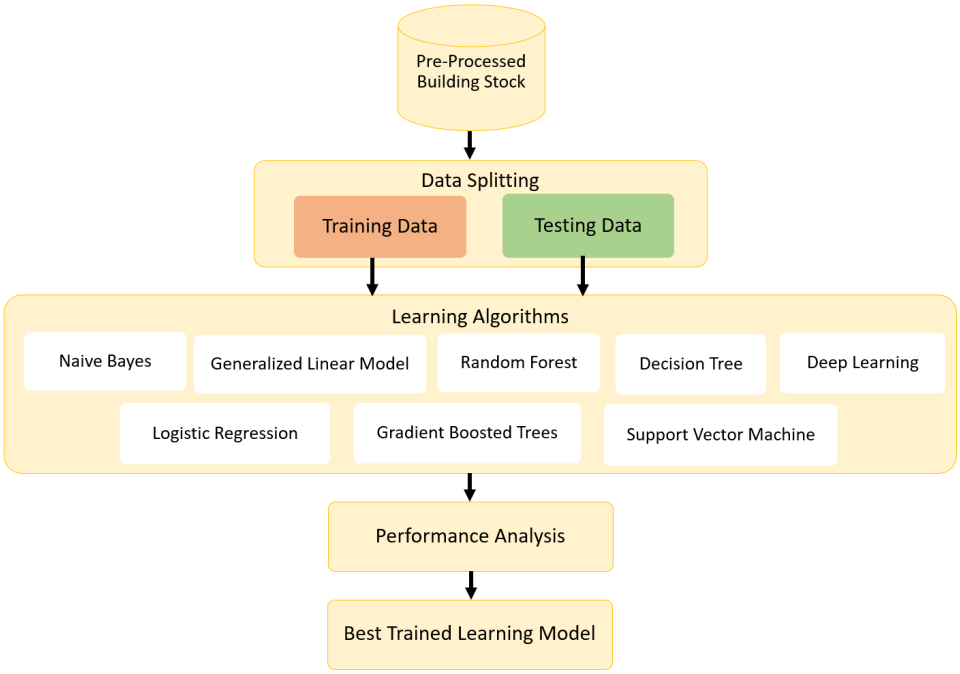

Figure 5: Machine learning model development process to predict residential building energy performance using data driven approach

The machine learning model development process implements classification algorithms to formulate a learning model. Classification forms part of a supervised machine learning algorithm that predicts the class of the given set of data points; classes are also known as labels or categories. This study employs eight different classification algorithms for energy prediction, namely, Naive Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Trees, Random Forest, Gradient Boosted Trees, and Support Vector Machine. These eight algorithms offer excellent performance when used for energy classification, prediction, or forecasting, as evident from previous studies [29, 12].

Model evaluation tests the effectiveness of the classification models. Some of the evaluation metrics include ACCuracy (ACC), precision, recall (Equations 7, 5, 6), and execution time [29].

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

where True Positives (TP) are the cases, which are predicted positive and are actually true. True Negatives (TN) are the cases, which are predicted negative but are true. False Positives (FP) are the cases which are predicted positive and are actually false. False Negatives (FN) are the cases, which are predicted negative but are false

Generally, the classification algorithm performance can be evaluated using accuracy. Accuracy represents the ratio of correctly predicted observations to the total observations. Larger values of accuracy signify better model performance. However, accuracy gives inappropriate results for imbalanced output labels. Therefore, this study considers alternative performance measures such as precision and recall metrics in addition to accuracy. Precision represents the number of positive class/label predictions that actually belong to the positive class/label. Similarly, recall is the number of positive class predicted out of all positive results in the dataset. Precision or recall can be used in the form of a confusion matrix that shows the overall performance summary of the class prediction results. A confusion matrix represents a specific table layout that provides a visualization of the classification algorithm performance. This study also considers the computation time for learning model development because the model development process should be efficient for large scale building stock data. Finally, the best-trained learning model based on performance indices is further used for predicting building energy performance for the entire building stock.

3.6. Multi-Scale GIS Modeling

This process maps the building energy performance results at multiple GIS scales, ranging from the individual building level to the national level. Due to the limited availability of individual building data on a national scale, the learning model predicts building energy performance by using building archetypes. Therefore, the learning model uses input features from the developed building archetypes at the local level in Section 3.4. These building archetypes help to generate individual building's data on a national scale. This study devises the building archetypes at a small area/neighborhood scale to conduct finely grained analysis. The data-driven multi-scale GIS modeling process uses the concept of a bottom-up approach for modeling the entire building stock. The modeling process comprise two major phases, namely, building modeling and multiple-scale modeling (Fig. 6).

The first phase implements GIS modeling at the building level using the developed learning model (described in Section 3.5). The process begins with the collection of input feature data for the entire building stock. This process extracts input features from multiple sources such as building archetypes, geographical, and census data and feeds these features to the best-trained learning model for building energy performance prediction. The building archetype data helps to gather input feature values for the entire building stock. The input features include the original features used to create the best training model. The geographical or census data comprise the quantification data of buildings (number of buildings) at each geographic scale. In the next step, the best learning model predicts the building energy performance for the entire building stock. Finally, the predicted results are further used for 2D or 3D GIS modeling of each residential building. 2D GIS modeling requires the building footprint to map the building energy performance. 3D GIS modeling is done by extruding the building footprint through building height data.

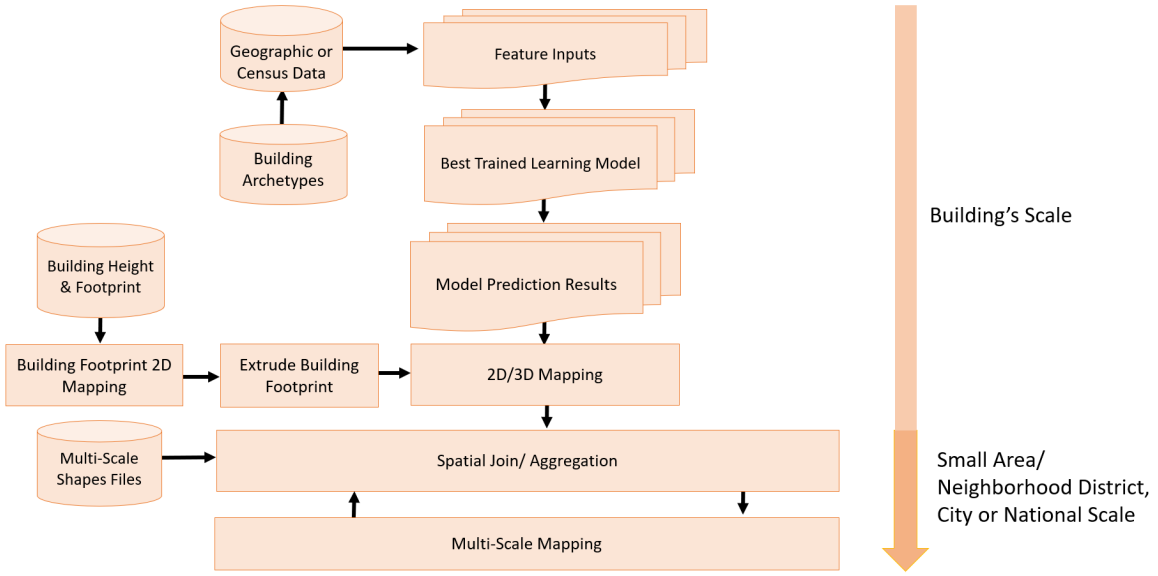

Figure 6: Multi-scale GIS mapping of residential building energy performance using data driven approach

In the second phase, building scale energy performance prediction results could be further extended to multiple scales such as small area, district, city, and county scales. The process uses the bottom-up concept to aggregate the building level results to a higher geographical level. Therefore, the spatial join or aggregation approach is used for multiple scale GIS modeling. A spatial join is a GIS operation that aggregates data from one geographical layer to another from a spatial perspective. The spatial join mapping at large scale requires shapes file for each scale. The spatial join process includes individual buildings, small areas and neighborhoods. Aggregated individual buildings represent small areas. All buildings in the small area aggregated to districts. On a similar note, district level predictions could be used to model cities or counties. Finally, the process combines the predicted 2D and 3D building layers to formulate the complete map for building energy planning and analysis.

3.7. Building Energy Planning and Analysis

This process implements the mapping of results for energy planning. The generated maps aid interested stakeholders when analyzing and identifying the priority areas for implementing energy-efficient strategies. The results can further be used to identify areas where energy policymakers can run targeted community-based events/campaigns to increase retrofitting activity. Compared to broad mass campaigns, targeted community based retrofit campaigns are more likely to be successful in increasing the retrofit activity in an area [60].

As data integration from different resources is a significant challenge for large scale GIS mapping, this process also implements a GIS-based Multi-Criteria Decision Analysis (MCDA) approach to support complex decision-making with multiple sources of decision analysis data [9]. MCDA approach helps to facilitate decision-makers to make the best possible decision with the consideration of multiple criteria. Furthermore, this approach is useful when various stakeholders have conflicting goals, objectives, and interests. The GIS-based MCDA approach is commonly used for the assessment of renewable energy potential, waste management, forestry, agriculture, and the environment sector [61]. In this research, the GIS-based building energy performance prediction results can be integrated with decision analysis data such as social, economic, or environmental data for complex decision making at a large scale (Fig. 7). GIS-based MCDA involves multiple steps for decision making analysis, namely,

1. Define the problem and set the goal or objective for decision-making analysis. An objective could be in terms of strategic policy or higher-level project output, such as economic, social or sustainable development;
2. Collect decision making data and predicted results in GIS layers format based on the MCDA objective. The spatial decision analysis data can be collected from a national census or spatial database;
3. Determine the appropriate thresholds for decision-making criteria or factors for each spatial layer. This could be acquired from experts', stakeholders' opinions, or existing literature from relevant fields. Each layer criterion must be measurable so as to reflect the performance for individual objectives;
4. Standardize or transform the criterion layers onto a relative scale. The process allows the comparison between each of the criterion layers and expert knowledge with meaningful scores;
5. Determine the weight (as a percentage) of each criterion based on its priority, importance, and objective. Generally, the Analytical Hierarchy Process (AHP) method is used for determining the weight of each layer [62]. AHP is a pairwise comparison approach that uses the experiences of experts or stakeholders' to estimate the weight. Furthermore, such a method allows both experts and stakeholders equal opportunity to give their input to derive qualitative and quantitative importance of each layer;
6. Aggregate or combine the generated layers based on criteria with defined weights. The final multi-criteria aggregated map developed using the Weighted Linear Combination

(WLC) technique could be used to obtain the suitability (priority) index or score S_a of each area a as follows (Equations 8) [63]:

$$S_a = \sum_{i=1}^{n} w_i x_i \tag{8}$$

where w_i is the calculated weight of criteria i as defined in step 5; x_i is the score of the area, a with respect to i criteria determined in step 4, and $i = 1, 2, ..., n$ where n is the total number of weighted criteria; and

7. Validate, and analyze the final GIS map.

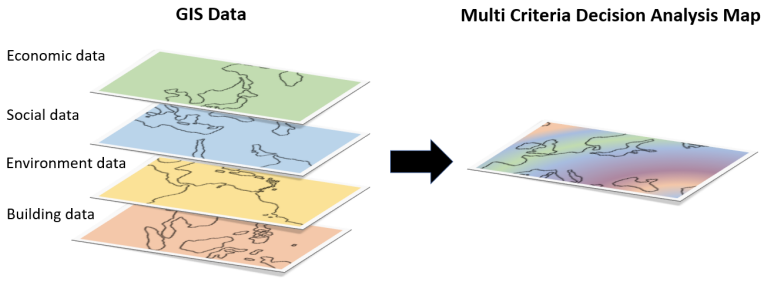

Figure 7: Multi-Criteria Decision Analysis (MCDA) approach using multiple GIS data for decision-making [64]

4. Case Study

The main objective of this case study is to develop a GIS-based building energy performance calculation methodology for the entire building stock of Ireland. The methodology integrates a data-driven approach with bottom-up modeling to predict (estimate) the building energy performance at multiple scales using spatial information. This study demonstrates the application of the devised approach using the Irish residential building stock. The geographical scale considered in Ireland at multiple levels including county, city, electoral district, and small area level. This allows for analysis of building energy performance at different spatial resolutions. This research proposes a GIS-based framework for multi-scale mapping of residential building energy performance that could act as a visual analysis tool for energy policymakers.

4.1. Data Collection

Collection of urban scale building stock data is quite challenging as individual building information is often unavailable. The data collection process involves the acquisition of raw building stock data from different sources, namely, EPC dataset, building census dataset, building footprint data, building geographical data, GIS data (shape files of small areas, districts, cities) and data from energy efficiency programs administered by the Sustainable Energy Authority of Ireland (Table 3).

Table 3: Building data requirements and associated data sources for Irish case study

Data Type	Case Study Data Source	Publisher
Building Stock	Irish EPC (BER) Database [65]	SEAI
Geographic data	GeoDirectory [66, 67]	An Post/ Ordnance Survey Ireland
GIS data	Ireland Counties, Districts, small area shapes files [68]	Central Statistics Office
Census	Irish Cenus database [68]	Central Statistics Office
Building Footprint	Dublin City Building footprint [67]	Ordnance Survey Ireland
Building Height/ LIDAR data	Dublin City Building Height Shape File	UCD School of Geography
Decision Analysis data	Irish Residential Energy Efficiency Program Data [69]	SEAI

Maintained by the Sustainable Energy Authority of Ireland (SEAI), the EPC (also referred to as Building Energy Rating (BER) certificate) dataset of the Irish residential stock represents the measured building stock and comprises more than 200 building features that include building fabric, heating systems, estimated end-use CO_2 emissions, estimated delivered, and estimated primary energy consumption. The Irish EPC dataset contains a building energy rating for each building which ranks the energy performance of the building on a graded scale from G to A1 based on the estimated energy consumption per metre squared per year [65]. The Irish EPC dataset contained approximately 695,000 residential buildings (at the end of year 2019) with the major proportion of building ratings lie within C1 and D2, with the highest percentage of building type being semi-detached and detached houses (Fig. 8).

The Irish census which is conducted every four years by the Central Statistics Office (CSO) collects a number of data points on the building in which the respondent lives . The census therefore provides the number of buildings in each geographical area [68]. According to the CSO 2016 dataset, there are approximately 1,983,715 residential buildings in Ireland, as opposed to the EPC dataset that consists of 695,000 residential buildings. This suggests that the EPC data is available for only $\approx 39\%$ of the residential building stock ([70]). This study employs machine learning algorithms to predict the energy rating of the remaining 61% of the stock by using limited variables.

The GeoDirectory database contains geographical information about the entire building stock of Ireland [66]. As the GIS mapping process requires geocoded buildings, a geocoding technique transforms the EPC building database using the GeoDirectory database. Published by An Post (Irish Postal Service) and Ordnance Survey Ireland, this database comprises geocoded addresses of 2,014,357 residential buildings.

The Irish retrofit housing scheme dataset contains quantitative data for residential buildings that have completed energy upgrades through one of SEAI's programs. Homeowners

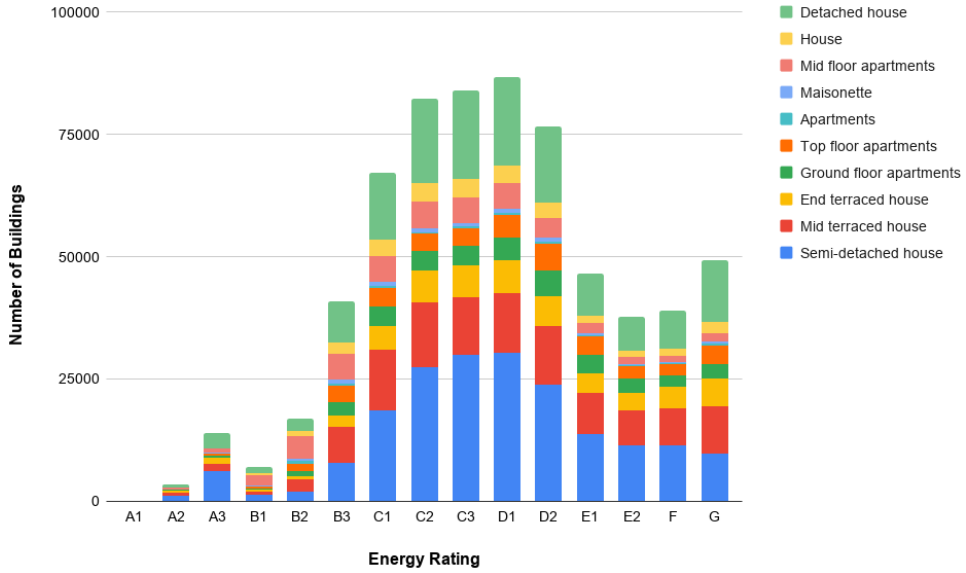

Figure 8: Irish EPC data distribution indicates that a considerable proportion of building energy ratings lie within the C1 to D2 range, with semi-detached and detached houses as the highest percentage of house type.

apply to SEAI for grants which subsidise the cost of their upgrades. Maintained by SEAI, the dataset comprises 265,182 retrofitted buildings and includes homes which have been upgraded through one of SEAI's energy upgrade programs such as Better Energy Homes, Warmer Homes, Better Energy Communities, and the Deep Retrofit pilot program [69].

This study uses the multi-scale concept for GIS mapping for individual Irish buildings, small areas, districts, cities, and counties. Each small area represents a group of buildings, and a cluster of small areas constitute one district. The mapping process maps the predicted building energy rating to the building stock. Based on information available from CSO, Ireland comprises 26 administrative counties, 5 cities, 139 municipal districts, and 18,641 small areas with more than two million residential buildings. The building footprint and boundaries of small areas, districts, cities, and counties levels are obtained from Ordnance Survey Ireland [67]. Published by the School of Geography at University College Dublin, individual building height data are only available for the residential stock of Dublin city. It is worthwhile to clarify the differences in 2D and 3D GIS mapping structures implemented in this case study. The 2D map represents the building energy performance at all multi-scale levels for Ireland. On the other hand, the 3D map only represents the building energy performance at the Dublin city level.

4.2. Geocoding

The lack of geocoded data poses a significant challenge when implementing GIS mapping. As the Irish EPC dataset does not include geocoded addresses, the geocoding process assigns a geocode to each residential building in the EPC dataset using a state of the art

Java based programming algorithm (Algorithm 1). As processing of the entire Irish building stock requires huge computational resources, this study implements a parallel programming method that uses multiple processes to improve the computational time. The geocoding process uses two datasets, namely, the Irish residential building EPC dataset (contains the addresses for geocoding) and the Irish GeoDirectory database (contains the geocoded addresses of residential buildings). In this case study, the GeoDirectory database contains three different geographic coordinate systems that assign the unique reference projections of each building, namely, Irish Grid (East, North), Irish Transverse Mercator (East, North), and ETRS89(longitude, latitude). The geocoding procedure then segments the data based on cities and counties. It is of paramount importance to implement a pre-processing technique on the aforementioned datasets as the EPC data collection process is manual and the data lacks geocoded features (longitude and latitude). Moreover, the EPC assessors may or may not follow a standardized procedure to fill geographical information such as address, postal code. Address pre-processing eliminates these inconsistencies before the data can be used to implement any address matching algorithms. The pre-processing procedure normally comprises data cleaning and data transformation tasks.

The geocoding process uses the processed EPC dataset to implement the fuzzy matching algorithms for string matching. This study compares four different fuzzy matching algorithms, namely, Jaro, Jaro-Winkler, Levenshtein, and Jaccard. The string matching process filters and compares the addresses in two levels. The first level compares the EPC addresses that contain house or apartment number with all the addresses in GeoDirectory database at individual building level. The second level compares those EPC addresses that do not contain house or apartment numbers with the nearest small areas in GeoDirectory database. The string matching process then compares the algorithms on the basis of a matching score, which determines the least matching criteria for geocoding addresses in the EPC dataset (Table 4). The results indicate that the minimum matching scores for Jaro-Winkler, Jaro, Jaccard algorithm, and Levenshtein are 0.90, 0.80, 0.50, and 0.50, respectively, for building level comparisons (Table 5). On a similar note, the minimum matching scores for Jaro-Winkler, Jaro, Jaccard algorithm, and Levenshteinare are 0.85, 0.75, 0.40, and 0.40 respectively for small area level comparisons (Table 5). Finally, the results of address matching are stored in the database for GIS mapping.

Algorithm 1: Algorithm pseudocode for geocoding of Irish EPC data

```
Result: Building stock geo coded database
Data: EPC database as epc
Data: GeoDirectory database as geo
Data: Cleaning dictionary as clean_dic
split epc based on counties as epc(county) ;
split geo database based on counties as geo(county);
while read all Irish counties as county do
    clean epc(county) using clean_dic as epc_clean;
    filter epc_clean as epc_num;
    filter epc_clean as epc_without_num;
    Geocoding(epc_num,geo(county));
    Geocoding(epc_without_num,geo(county));
end
Function Geocoding(nongeo_db,geo_db)
    while read all nongeo_db addresses do
        while read all geo_db addresses do
            call string matching algorithm ;
            calculate score;
        end
        sort and select highest score ;
        if highest score match criteria then
            add to geo coded database
        end
    end
end
```

Table 4: Fuzzy string matching algorithm example to explain geocoding criteria for Irish addresses

Address	Matched Address	JW	JO	JD	LV	Criteria
123 UCD Merville Reception, Belfield, Dublin 4	123 UCD Merville, Belfield	0.88	0.80	0.57	0.57	Pass
123 UCD Square, UCD road north UCD, Dublin 6W	123 UCD Square, 6W	0.85	0.75	0.40	0.40	Pass
123 UCD Avenue Dublin 4	123 UCD Avenue, Belfield, Dublin 4	0.90	0.84	0.67	0.68	Pass
UCD Avenue, UCD, Dublin 4	UCD Avenue, Belfield Campus, Dublin 4	0.87	0.78	0.59	0.62	Pass
12 UCD , Blackrock Campus, Co.Dublin	123 UCD, Belfield Campus, Dublin 4	0.76	0.70	0.59	0.68	Fail
12B, UCD Block B, Blackrock, Dublin 4	12B, Block B, Dublin 1	0.78	0.63	0.55	0.58	Fail

Note: The above entries represent dummy addresses. Original addresses are not used due to privacy issues.

Table 5: Geocoding criteria for fuzzy string matching algorithm of Irish EPC residential building dataset

Algorithm	Building Level Criteria	Small Area/ Street Level Criteria
Jaro Winkler (JW)	0.90	0.85
Jaro (JO)	0.80	0.75
Jaccard (JD)	0.50	0.40
Levenshtein (LV)	0.50	0.40

4.3. Building Stock Pre-Processing

The building stock pre-processing procedure extracts the Irish building characteristics and associated energy usage using data-driven methods. These methods include the initial statistical analysis, data pre-processing, outlier detection, and feature selection techniques.

An initial statistical analysis of density plots for the roof and floor U-values reveals that the entire spectrum of U-values contains a significant number of zeroes (Fig. 9). The data pre-processing step eliminates these inconsistencies; average values for features (identified using clustering of building type) missing values in the dataset. Data pre-processing also involves data filtering and data transformation. While data filtering removes irrelevant data instances, data transformation converts all categorical and nominal values into numerical values as data-driven techniques usually processes numerical values. Furthermore, the data transformation technique reduces several combinations of rating classifiers (for instance, A, B, C, D, and EFG) from the existing rating labels (A1, A2,..., E, F, G) [55]. The classifiers generate clusters of adjacent energy ratings. For instance, the classifier labeled EFG comprises the individual rating labels E, F, and G. This is done to determine whether reducing the number of classifiers will affect the learning model efficiency used in the prediction of building energy rating. Furthermore, the ten residential building types in the EPC dataset are merged into five major ones, namely, apartments (top, middle, ground, maisonette), semi-detached houses, detached houses, terraced (middle or end) houses, and bungalows (houses). This is essential to ensure the consistency of building types in the GeoDirectory dataset.

This study implements the LOF algorithm to remove the outliers from the EPC dataset because this algorithm is viable for large datasets [14]. The LOF algorithm uses the distance function to measure the density of objects amongst each other. The Euclidean distance measure is used with the LOF algorithm for this case study. The lower and upper bounds for minimum points for the distance measure are set to 10 and 20, respectively. The results indicate that the EPC dataset contains a significant number of outliers; for instance, in the building window, wall roof, and floor u-values in the EPC dataset (Fig. 9).

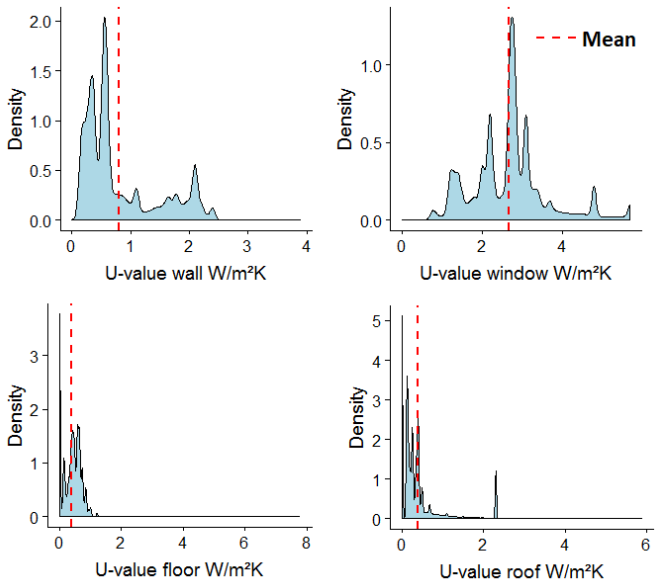

Figure 9: Density plots of building fabric (wall, window, floor and roof) U-values (W/m^2K) for Dublin city consisting of more than 250,000 residential buildings with the mean value to illustrate the initial statistical analysis.

While the EPC dataset contains more than 200 variables, this study considers only the influential variables for archetype and learning model development. The case study uses both engineering and data-driven methods to identify a subset of most relevant features. In the first step, the engineering method determines 63 features out of more than 200 features based on existing studies [14, 71]. In the next step, the data-driven selection method identifies 43 features out of selected 63 features based on multiple statistical tests such as variance/standard deviation threshold and correlation coefficient. The correlations coefficient removes those features that closely mirror the output feature. The output feature in the EPC data is building energy rating label expressed in terms of primary energy ($kWh/(m^2 \cdot yr)$). A correlation of less than 0.01% and more than 50% suggests that the feature has no significant influence on the building energy rating. The standard deviation threshold method eliminates features that are too similar or dissimilar. The removed features either include those that are more than 90% of all values being identical or features with lots of missing values. For instance, floor level features such as floor fabric U-values do not fulfill the set criteria as nearly all values are identical. EPC assessors, while performing surveys, often submit default values for floor level features due to the absence of accurate data. Other eliminated features include the date, ID, and target value.

The feature selection process lists 43 influential features out of the initial 200 features. These 43 features can be categorized based on building envelope, building fabric heating system, hot water, spatial and output labels (Table 6). The final processed data comprise only improved quantitative building stock information, which is used for archetype formulation and learning model development.

Table 6: Selected features and their types form Irish EPC dataset for data-driven model development

Features Type	Features
Building Envelope	No of stories, area (ground, floor, wall, door, roof, window and total physical), living area percent and year of construction.
Heating System	Efficiency (main, supply and adj factor), supply heat fraction, fuel (main, supply) and central heating boiler thermostat.
Hot Water	Fuel (main, supply) and water storage volume.
Building Fabric	U-value (wall, door, roof, floor, windows and fabric), total area (opening and loss fabric), percent open area, insulation thickness, insulation type, avg u-value openings, thermal mass category, primary circuit loss and most significant type (window, roof).
Spatial	Small area and county code.
Output Label	Energy rating

4.4. Building Archetypes Development

The building archetypes development procedure involves two essential processes, namely, segmentation and characterization. For the building stock under consideration, the segmentation, using building type criteria, identifies five types of buildings. As individual buildings in the structured dataset contain their own set of values for different variables (features), the characterization process aggregates (median) these values for buildings that belong to one particular segment (archetype). Thus, the aggregation results in a single set of values for associated variables. These aggregated values for each building type represent the characteristics of an individual building archetype. This study implements aggregation at the small area level using building type segmentation for granular level analysis. As per the analysis, there are 18,641 small areas in Ireland. Five different building types exist in the GeoDirectory database, namely, apartments, terraced houses, detached houses, semi-detached houses, and bungalows. This process resulted in the identification of 93,205 small areas building archetypes on the basis of building type segmentation that represent more than two million residential building in Ireland.

4.5. Data Driven Building Energy Model Development

This process involves the formulation of a building energy performance machine learning classification model. The process begins with data splitting, which randomly splits the Irish EPC dataset into two groups to create training and testing datasets.

This study employs and compares eight algorithms to devise the classification model, namely, Naive Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. The algorithms are compared on the basis of different classifications of energy ratings. The training process considers seven classifications of energy ratings. The results show that deep learning algorithms can effectively handle complex and high-dimension data with a large number of input features (such as the Irish EPC dataset used in this case study). On a similar note, the GBT algorithm effectively handles datasets with both categorical and numerical

values. Furthermore, both algorithms can handle missing entries in the dataset, thereby, enhancing the model accuracy. Although studies have shown that SVM is often the optimal choice for building energy prediction, the algorithm is certainly not suitable when handling large datasets. Deep learning algorithms deliver high accuracy for a significant number of classification scenarios when compared to other algorithms. Although the interpretability of deep learning algorithms is less compared to algorithms like classification and regression trees, the high number of input features reduces the interpretability of classification and regression trees by a significant proportion.

Classfication	Performance Measure	NB	GLM	LG	DL	DT	RF	GBT	SVM
A1,A2...E,F,G	ACC	17%	41%	23%	76%	60%	39%	67%	13%
	Time	0:03	2:06	0:14	05:55	0:02	0:01	04:22	2:25
A,B,C,D,E,F,G	ACC	36%	60%	50%	78%	56%	31%	77%	36%
	Time	0:02	0:35	0:07	04:47	0:01	0:03	0:21	2:38
A,B,C,D,EFG	ACC	56%	70%	45%	81%	70%	34%	81%	26%
	Time	0:02	0:17	0:06	05:23	0:01	0:02	0:18	2:53
A,B,CD,EFG	ACC	76%	84%	75%	88%	77%	33%	83%	61%
	Time	0:02	0:09	0:05	04:39	0:01	0:02	0:14	5:41
A,B,C,D,EF,G	ACC	44%	58%	49%	73%	53%	34%	77%	36%
	Time	0:03	0:33	0:08	04:56	0:01	0:02	0:18	3:00
A,B,CD,EF,G	ACC	65%	72%	64%	77%	72%	48%	84%	61%
	Time	0:02	0:14	0:05	04:42	0:01	0:02	0:18	9:33
A,B,C,DE,FG	ACC	49%	68%	61%	78%	70%	50%	84%	38%
	Time	0:02	0:18	0:10	06:22	0:02	0:03	0:15	8:48

Low High

Note: Naive Bayes, Generalized Linear Model (GLM), LogisticRegression (LG), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT), and Support Vector Machine (SVM).

Figure 10: Comparison of different building energy rating classification along with performance analysis of learning model for using Irish EPC dataset.

The classification A, B, CD, and EFG returns the highest accuracy of 88% using the deep learning algorithm (Fig. 10). It is worthwhile to mention that this classification is acceptable for stakeholders; the goal is often to identify the buildings with significantly poor performance. These findings indicate two significant conclusions. The developed data-driven model could calculate a building's energy rating using a limited number of input features with the highest accuracy. Furthermore, the model accuracy could be improved by aggregating lower energy rating labels. For instance, the highest model performance with actual energy rating classification (A1,A2,...E,F,G) is 76% and the model accuracy experiences an increase of 12% with aggregated classification (A, B, CD, and EFG) using the deep learning algorithm (Fig. 10).

The selected learning model comprises four energy rating classifiers, namely, A, B, CD, EFG. The model delivers the highest accuracy of 88% using the deep learning algorithm that uses 43 input units with 2 hidden layers, each of size 50 units and 4 output units (Fig. 11). A deeper investigation of the model using the confusion matrix indicates that the precision of four output classes is between 75% and 95%. A confusion matrix is a table that describes the performance of each label in the classification model. Similarly, the recall values are

between 76% and 99% for the four output labels. The precision values are the highest for the EFG class (95%), indicating that the model correctly predicts 14,522 out of all 15,254 prediction results. The recall values are the highest for the A-class (99%), showing that the model accurately predicts 1899 out of all 1920 actual results (Table 7).

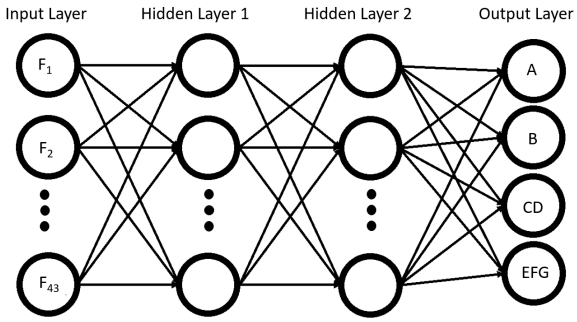

Figure 11: Deep learning model architecture based on 1 input, 2 hidden and 1 output layer for prediction of building energy rating using Irish EPC data.

Table 7: Confusion matrix depicting the overall summary of prediction results obtained using deep learning model for Irish EPC residential building data.

	true A	true CD	true B	true EFG	class precision
pred. A	1899	63	336	24	82%
pred. CD	4	40720	313	4452	90%
pred. B	17	1997	6312	38	75%
pred. EFG	0	626	106	14522	95%
class recall	99%	94%	89%	76%	

4.6. GIS Building Energy Performance Mapping

This process involves the mapping of EPC prediction results at multiple scales, which range from the individual building level to the national level. The developed building archetypes use 43 input features to represent the unique buildings in small areas. These archetypes represent the entire Irish building stock using the GeoDirectory dataset. The formulated deep learning model uses the values of 43 input features to estimate the building energy performance of the whole building stock. The mapping process maps these modeling results to obtain 2D and 3D GIS maps using ArcGIS tool. As mentioned earlier, the 3D GIS maps only consider the building stock of Dublin city as building footprint, and height data are only available for this particular region (Fig. 12).

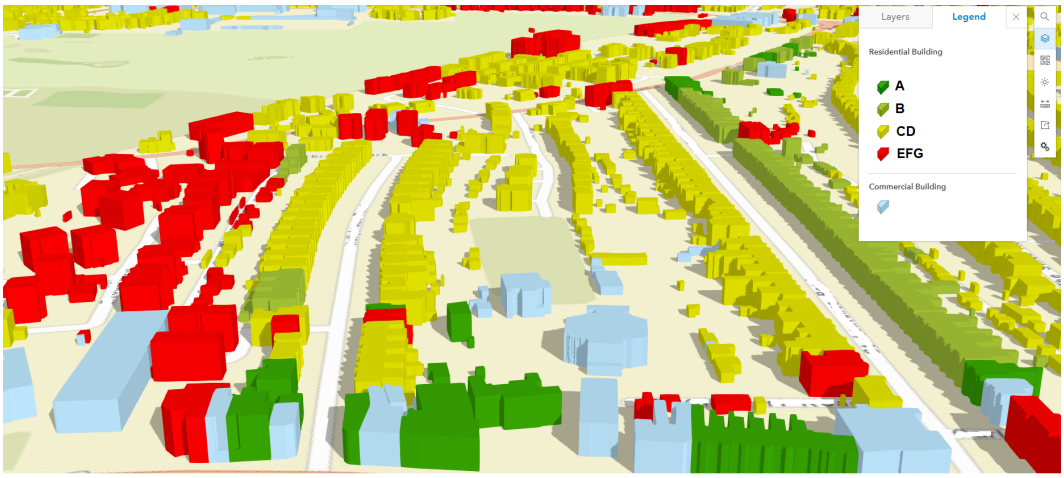

Figure 12: 3D map of Dublin city shows the predicted building energy rating across different buildings; each building represents a small area building archetype using ArcGIS online tool.

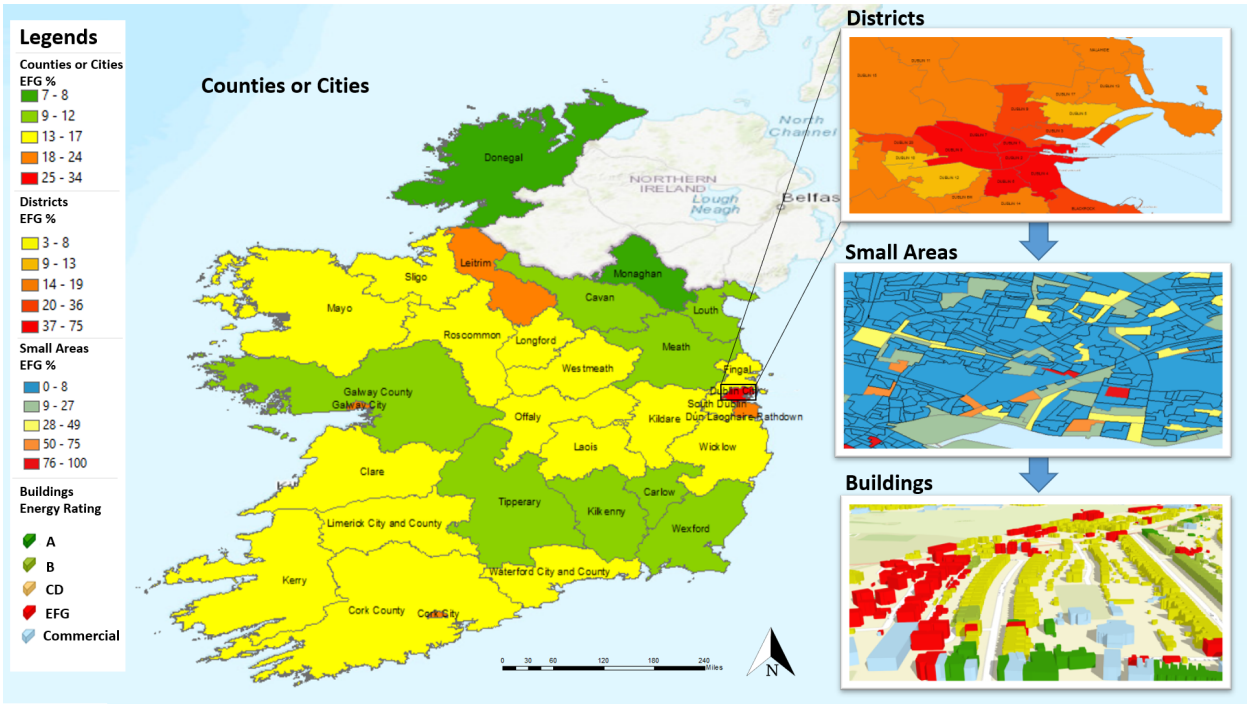

Figure 13: Irish multi-scale map shows the percentage of EFG building energy rating prediction results that helps to estimate energy demand distributions at multiple-scale

Finally, 2D GIS-based building scale energy performance prediction results are further extended to small area, district, city, and national scale by using a spatial join or aggregation approach. The process uses a bottom-up approach to aggregate the energy performance

prediction results from building to a national scale. The process starts with the spatial aggregation of building energy rating results at a small area level. The next step aggregates all small area predictions to district, city and county levels. The developed map can be used to visualize the distribution of energy ratings at multiple scales. For instance, the results can help identify the percentage of poorly performing buildings (EFG rating) for multiple levels (Fig. 13).

The results indicate that the highest percentage of predicted EFG ratings belongs to Dublin, Cork and Galway city councils that represent 34%, 24% and 24% of total residential buildings respectively. (Fig. 14). Moreover, when implementing the Dublin district-level analysis, results indicate that city center districts (Dublin 1 and Dublin 2) have the highest number of EFG ratings (Fig. 15). Similarly, stakeholders can identify the distribution of small area energy ratings of specific districts such as Dublin 1 or Dublin 2. Furthermore, for each small area, 3D building modeling results help to perform fine-grained analysis.

The developed multiple-scale map helps the decision makers to identify areas where there are a large number of energy inefficient buildings. This information can then be used to conduct targeted community based social marketing which can increase the rate of retrofit in the area. The map also identifies clusters of Irish residential buildings with a poor energy performance that further suggests which area has poor levels of insulation and heating systems performance. The results further reveal heating or electrical demand in a given area for energy planning purposes. For instance, the map can identify areas where district heating projects may be efficient.

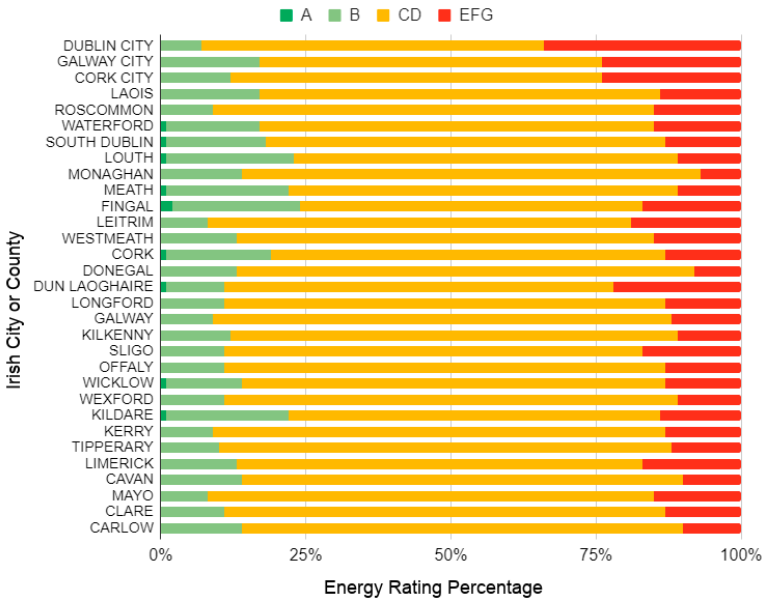

Figure 14: Proportion of energy rating predictions for Irish cities/counties expressed in percentage.

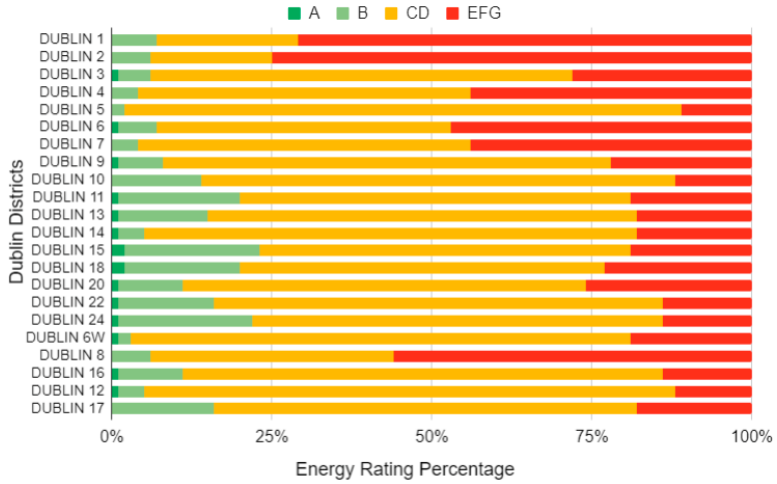

Figure 15: Proportion of energy rating predictions for different districts of Dublin city and counties expressed in percentage.

4.7. Building Energy Performance Planning and Analysis

This process involves the application of GIS maps for energy planning and decision making. For this particular case study, the main goal is to identify areas in the Irish counties where energy policymakers can run targeted community-based events/campaigns to increase retrofitting activity. Hence, this process implements the Multi-Criteria Decision Analysis (MCDA) approach to formulate planning decisions. It is worthwhile to note that the GIS-based building energy performance prediction results can be integrated with decision analysis data.

Three different decision analysis data layers, such as retrofit grant, population, and socio-economic (household income), are considered for implementing complex decision-making and decision analysis. The Irish retrofit grants refer to the financial support available for homeowners to upgrade their residential buildings. The support subsidizes the cost of energy efficiency upgrades. Population and socio-economic data are collected from the Irish census database. This analysis would aid in the implementation of retrofit activities that enhance the energy performance of buildings and thereby, reduce energy consumption. Furthermore, the results can also help the urban planners for energy-based planning and decision analysis.

This case study focuses on the small areas of Dublin Fingal county which contain more than 114,000 residential buildings to demonstrate the application of GIS-based MCDA analysis using the ArcGIS tool. Each layer clearly indicates the non-priority areas as well as the priority areas based on defined criteria. The first layer, predicted EPC energy rating, represents small areas that have more than 40% EFG rated buildings. The total land priority area for the EFG rating layer is estimated to be 66 km^2, which accounts for about 14% of the entire Dublin Fingal area. The second layer, population, represents the number of building inhabitants with greater than 300 inhabitants in a small area. The total land priority area for the population layer is estimated to be 373 km^2, which accounts for about 77% of the entire Dublin Fingal area. The third layer, represents the distribution of retrofit grants over

an area which received with a maximum coverage of 20%. The total land priority area for the retrofit grant layer is estimated to be 385 km^2, which accounts for about 80% of the entire Dublin Fingal area. Finally, the fourth socio-economic layer represents the criterion of greater than 60% of low-income households in the small area. The total land priority area for the socio-economic layer is estimated to be 433 km^2, which accounts for about 90% of the entire Dublin Fingal area (Table 8).

Table 8: Map layers criteria with weights and source for multi-criteria decision analysis of Dublin Fingal county.

Layers	Criteria	Weight	Source
Socio-Economic	Area having low income household % greater than 60	0.58	Central Statistics Office
EFG Energy Rating	EFG energy rating buildings % greater than 40 in small area	0.25	Prediction Results
Retrofit Grant	Retrofit grant received area % less than 20%	0.11	SEAI
Population	Greater than 300 inhabitant in small area	0.06	Central Statistics Office

The final map indicates the aggregated operation of the weighted criteria maps (socio-economic, population, retrofit grant) including the predicted energy rating map. The weights are assigned to each layer using the Analytical Hierarchy Process (AHP) method. The highest weight of 0.58 is assigned to the socio-economic layer, followed by the 0.25 weight assigned to the EFG rating layer. The 0.11 weight is assigned to the retrofit grant and 0.06 weight assigned to the population layer. The final aggregated map categorizes the priority areas into five classes: non-priority, low priority, medium priority, high priority, and extreme priority areas. The land area for the total extreme priority category is estimated to be 41 km^2, which accounts for about 9% of the total Dublin Fingal area. The high priority land area is estimated to be 274 km^2, which accounts for about 57% of the entire Dublin Fingal area. The medium and low priority land areas are estimated to be 88 and 35 km^2, which accounts for about 18% and 7% of the total Dublin Fingal area respectively.

The final developed maps from this analysis aid stakeholders to analyze and identify the priority areas for implementation of sustainable energy decisions. The map indicates the areas that are most beneficial for targeting community-based campaigns to increase the retrofit activity (Fig. 16). Furthermore, when considering the Irish building stock, stakeholders would be interested in identifying the potential areas where the need for heavily subsidized retrofit schemes would be high. For instance, such an approach could inform SEAI determine potential targets to increase the uptake of the Warmer Homes program.

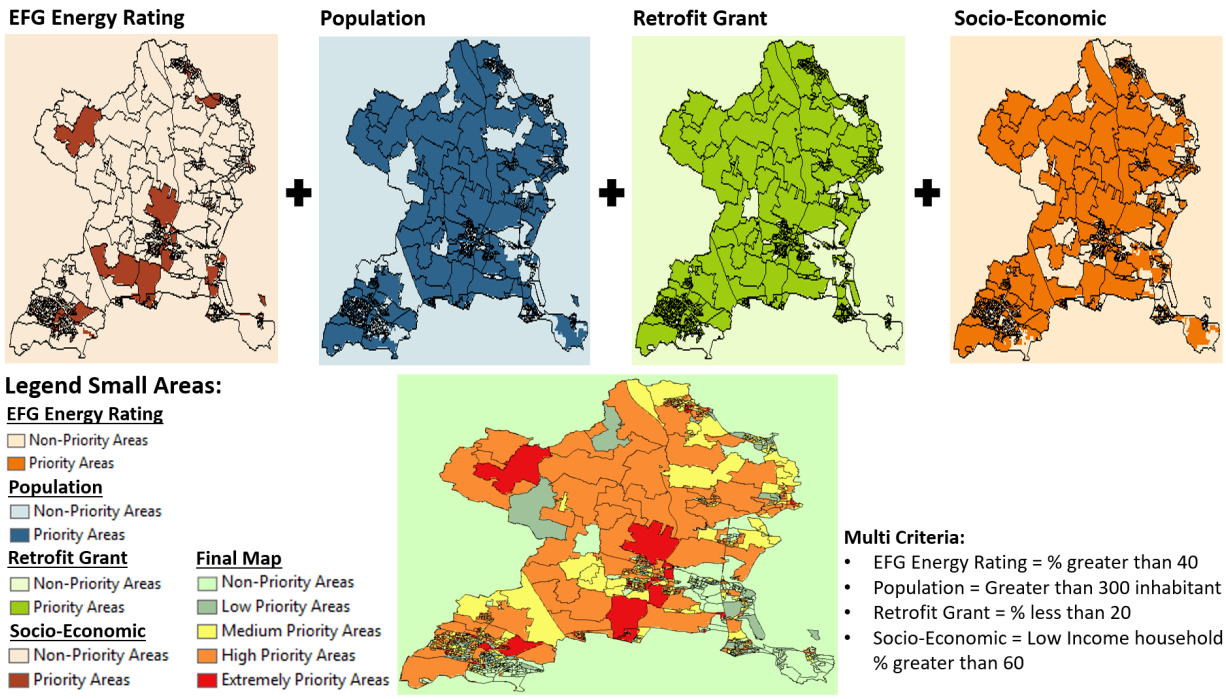

Figure 16: Dublin County (Fingal) small areas maps for multi-criteria decision analysis to help stakeholders analyze and identify priority areas to implement sustainable energy schemes

The real-life application of MCDA analysis could be further extended to support urban decision-making and facilitate energy planning and analysis in urban areas by minimizing CO_2 and energy usage. For instance, there are 4918 residential EPC rating buildings identified in the high priority area. Upgrading the buildings in the A-rated building will reduce $\approx$1751 MWh/m^2/yr EUI savings and $\approx$ 407 tonnes/m^2/yr in CO_2 reduction. These estimates would help urban planners in targeting renovations in areas of particular interest. Furthermore, the policymakers would keep track of the builder sector in terms of energy efficiency and carbon emissions.

5. Discussion

Planning urban energy systems often involves the use of limited building stock data to devise and implement energy policy decisions at multiple scales. The proposed data-driven methodology uses the limited building stock data to facilitate planning at various scales. Although the Irish Energy Performance Certificate database represents only 39% of the entire residential building stock, the devised methodology uses this limited data (650,000 Irish EPCs) to predict the energy performance of more than 2 million Irish residences using deep learning algorithms. Alongside this, the methodology identifies a list of influential building features that will further aid the planning process by substantially reducing the scope of the required analysis. Integrating the prediction results with a spatial aggregation

approach would further aid the stakeholders identify areas in Irish counties where energy policymakers can run targeted community-based campaigns to increase retrofitting activity.

Although the proposed methodology can be applied generally, the approach has limited applicability due to the availability of required data. As the approach is data-dependent, the scale and quality of the available data has a huge impact on the generated results. These limitations are expressed in further detail as follows.

- Data Quality: GIS-based modeling uses building stock data normally acquired through surveys. For instance, building energy rating assessors collect and assemble the EPC manually. Although each country measures data quality and mandates that assessors follow defined standards, the process is prone to human error. Similarly, as geocoding requires high resolution data, accessing GIS data at fine granular level would pose a significant challenge when implementing this methodology.

- Computational Time: This study uses a national building stock database that contains a massive amount of data. Implementation of processes, namely, geocoding, data preprocessing and learning model development with such databases would required significant computational power. This study uses a server that comprises 2 x Intel Xeon E5-2697 CPUs with 30 MB cache on each processor, 48 cores and 256 GB of RAM. The computational time of the eight learning algorithms considered in the case study might be different for a different system.

- Building Archetype Development Approach Bias: In this case study, the archetypes are developed at a small area level for fine grained analysis. However, the availability of data required for archetypes development at lower scales could pose a significant challenge. It is worthwhile to mention that the case study uses only the building type as the segmentation criterion for archetype development. Construction age has also been extensively used as the segmentation criterion and plays a crucial role in archetype development. However, construction age data was not available in the quantification dataset for GIS mapping. Hence, the archetypes are formulated using only the building type characteristic.

- EPC Coverage Bias: The representativeness of building performance data is hugely dependent on the rules governing the EPC in each country. For instance, the EPC regulation in Ireland mandates that every house on the market must complete a EPC before the house can be sold/rented. These houses could significantly differ from the total building stock. Henceforth, model built on areas with a low overall EPC coverage will lead to somewhat inaccurate estimates of the total quality of the building stock.

6. Conclusions and Future Work

The research conducted in this paper identifies a generalized data-driven methodology based on the bottom-up approach for mapping of residential building energy performance at multiple scales. Urban planners and energy policy makers often face significant challenges when implementing sustainable energy analysis and planning at a large scale due to

the complexity of the energy system. From the individual building level to the national level, system complexity increases exponentially, mainly due to a significant increase in the number of buildings and associated data resources. This research formulates a methodology to effectively integrate various data resources using machine learning algorithms; this can facilitate energy planning at multiple scales. The proposed method can estimate building energy performance from local to national scale with limited knowledge of building dynamics. This study deploys a bottom-up approach for detailed qualitative and quantitative analysis. Modeling results help in the development of an energy efficiency footprint of buildings at the urban scale.

The methodology devised in this study proposes a generalized solution for energy planning and decision making at multiple scales. The solution uses limited available resources such as energy performance certificates, geographical, spatial, census, and retrofit project data to predict the building energy performance. The study uses a data-driven technique to geocode building stock data for spatial mapping. The results further indicate that the data-driven approach coupled with the spatial data could enhance the quality of existing data and extract meaning full knowledge for decision making. The bottom-up data-driven approach predicts the building energy performance using available limited building stock data. A comparison of eight different learning algorithms concludes that no single learning algorithm yields perfect results. The selection of learning algorithms is usually case-specific and depends on the data used for training the learning model.

When dealing with complex energy systems, urban planners and energy policymakers face several challenges for the implementation sustainable planning and energy efficiency solutions at the urban scale. The generalized data-driven methodology could produce maps of residential building energy performance at multiple scales, which will aid the planning process. Moreover, planning at the urban scale often involves a significant amount of building data. Urban planners could use the proposed approach to integrate various data resources using machine learning techniques; this eventually facilitates planning at multiple scales. Furthermore, urban planners could use the GIS modeling results to develop an energy efficiency footprint of the urban building stock and henceforth, identify priority areas to implement energy efficiency solutions.

Future work will investigate the integration of building stock energy performance and social science based research (such as occupancy and demand patterns). For instance, it would be worthwhile to investigate how buildings with the same Energy Performance Certificate differ statistically in terms of their measured consumption. This research could be further extended to identify the applicability for commercial buildings. The results achieved by using the proposed methodology could also be improved by using detailed building quantification data. It is worthwhile to mention that data licensing and computational resource requirements play a crucial role in GIS-based building stock modeling. However, these aspects are not included in the scope of this study and will be further investigated as a part of future research. Furthermore, as the proposed methodology deals with national scale databases that contain huge amount of data, integration with cloud-based or big data approaches using services such as Google Cloud Platform or Amazon Web Services would add a significant value to the research.

7. Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under the SFI Strategic Partnership Programme Grant number SFI/15/SPP/E3125. We acknowledge the Sustainable Energy Authority of Ireland (SEAI) for access to anonymised Building Energy Rating (BER), Better Energy Homes (BEH) and Better Energy Warmer Home (BEWH) datasets. The opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the SFI and SEAI. The authors would also like to acknowledge IBPSA Project 1, an international project conducted under the umbrella of the International Building Performance Simulation Association (IBPSA). Finally, we would like to acknowledge GeoDirectory and UCD School of Geography to provide spatial data for Ireland.

References

[1] X. Cao, X. Dai, J. Liu, Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade, Energy and buildings 128 (2016) 198–213.

[2] Energy for Europe by European Commission, `https://ec.europa.eu/energy/en`, [Online; accessed 02-Mar-2020].

[3] E. Recast, Directive 2010/31/eu of the european parliament and of the council of 19 may 2010 on the energy performance of buildings (recast), Official Journal of the European Union 18 (06) (2010) 2010.

[4] C. Li, 2.09 Gis for urban energy analysis, in: Comprehensive geographic information systems, Elsevier Oxford, 2018, pp. 187–195.

[5] A. Alhamwi, W. Medjroubi, T. Vogt, C. Agert, Gis-based urban energy systems models and tools: Introducing a model for the optimisation of flexibilisation technologies in urban areas, Applied energy 191 (2017) 1–9.

[6] U. Ali, M. H. Shamsi, C. Hoare, J. O'Donnell, Gis-based residential building energy modeling at district scale, in: BSO 2018: 4th IBPSA-England Conference on Building Simulation and Optimization, Cambridge, United Kingdom, 11-12 September 2018, International Building Performance Simulation Association, 2018.

[7] U. Ali, M. H. Shamsi, C. Hoare, E. Mangina, J. O'Donnell, A data-driven approach for multi-scale building archetypes development, Energy and Buildings 202 (2019) 109364.

[8] C. F. Reinhart, C. C. Davila, Urban building energy modeling–a review of a nascent field, Building and Environment 97 (2016) 196–202.

[9] S. T. Moghadam, P. Lombardi, An interactive multi-criteria spatial decision support system for energy retrofitting of building stocks using communtiyviz to support urban energy planning, Building and Environment 163 (2019) 106233.

[10] T. Hong, Y. Chen, X. Luo, N. Luo, S. H. Lee, Ten questions on urban building energy modeling, Building and Environment 168 (2020) 106508.

[11] A. Mastrucci, O. Baume, F. Stazi, U. Leopold, Estimating energy savings for the residential building stock of an entire city: A gis-based statistical downscaling approach applied to rotterdam, Energy and Buildings 75 (2014) 358–367.

[12] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205.

[13] A. Nutkiewicz, Z. Yang, R. K. Jain, Data-driven urban energy simulation (due-s): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow, Applied energy 225 (2018) 1176–1189.

[14] U. Ali, M. H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, Applied Energy 267 (2020) 114861.
[15] U. Ali, M. H. Shamsi, M. Nabeel, C. Hoare, F. Alshehri, E. Mangina, J. O'Donnell, Comparative analysis of prediction algorithms for building energy usage prediction at an urban scale, in: Journal of Physics: Conference Series, Vol. 1343, IOP Publishing, 2019, p. 012001.
[16] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, et al., Energyplus: creating a new-generation building energy simulation program, Energy and buildings 33 (4) (2001) 319–331.
[17] P. Fritzson, V. Engelson, Modelicaa unified object-oriented language for system modeling and simulation, in: European Conference on Object-Oriented Programming, Springer, 1998, pp. 67–90.
[18] S. A. Klein, Trnsys-a transient system simulation program., University of Wisconsin-Madison, Engineering Experiment Station Report (1988) 38–12.
[19] Building Performance Database by Department of Energy, `https://bpd.lbl.gov/`, [Online; accessed 02-Mar-2020].
[20] Energy Performance Certificates across europe from design to implementation, BPIE, `http://www.buildup.eu/en/practices/publications/energy-performance-certificates-across-europe-design-implementation`, [Online; accessed 02-Mar-2020] (2010).
[21] T. Lan, G. Lansley, J. Van Dijk, P. Longley, Geocoding historical census records in england and wales, in: Proceedings of the 27th Conference on GIS Research UK (GISRUK), Vol. 2019, Geographic Information Science Research UK, 2019.
[22] Q. Tian, F. Ren, T. Hu, J. Liu, R. Li, Q. Du, Using an optimized chinese address matching method to develop a geocoding service: a case study of shenzhen, china, ISPRS International Journal of Geo-Information 5 (5) (2016) 65.
[23] T. Hong, Y. Chen, S. H. Lee, M. A. Piette, Citybes: A web-based platform to support city-scale building energy efficiency, Urban Computing 14.
[24] C. Reinhart, T. Dogan, J. A. Jakubiec, T. Rakha, A. Sang, Umi-an urban simulation environment for building energy use, daylighting and walkability, in: 13th Conference of International Building Performance Simulation Association, Chambery, France, 2013.
[25] J. A. Fonseca, T.-A. Nguyen, A. Schlueter, F. Marechal, City energy analyst (cea): Integrated framework for analysis and optimization of building energy systems in neighborhoods and city districts, Energy and Buildings 113 (2016) 202–226.
[26] N. Abbasabadi, M. Ashayeri, Urban energy use modeling methods and tools; a review and an outlook for future tools, Building and Environment (2019) 106270.
[27] Y. Sun, F. Haghighat, B. C. Fung, A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energy and Buildings (2020) 110022.
[28] A. Scarlat, Machine Intelligence Primer for Clinicians: No Math Or Programming Required, Amazon Digital Services LLC - KDP Print US, 2019.
[29] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047.
[30] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, Energy and Built Environment 1 (2) (2020) 149–164.
[31] N. Abbasabadi, R. Azari, A framework for urban building energy use modeling, ARCC Conference Repository 1 (2019) 386–394.
[32] W. Li, Y. Zhou, K. Cetin, J. Eom, Y. Wang, G. Chen, X. Zhang, Modeling urban building energy use: A review of modeling approaches and procedures, Energy 141 (2017) 2445–2457.
[33] J. Ma, J. C. Cheng, Estimation of the building energy use intensity in the urban scale by integrating gis and big data technology, Applied Energy 183 (2016) 182–192.

[34] C. E. Kontokosta, C. Tull, A data-driven predictive model of city-scale energy use in buildings, Applied energy 197 (2017) 303–317.
[35] S. J. Quan, Q. Li, G. Augenbroe, J. Brown, P. P.-J. Yang, Urban data and building energy modeling: A gis-based urban building energy modeling system using the urban-epc engine, in: Planning Support Systems and Smart Cities, Springer, 2015, pp. 447–469.
[36] S. T. Moghadam, G. Mutani, P. Lombardi, Gis-based energy consumption model at the urban scale for the building stock, in: 9th International Conference, Improving Energy Efficiency inCommercial Buildings & Smart Communities Conference (IEECB & SC16), in print, 2016.
[37] S. Yamamura, L. Fan, Y. Suzuki, Assessment of urban energy performance through integration of bim and gis for smart city planning, Procedia engineering 180 (2017) 1462–1472.
[38] Y. Chen, T. Hong, M. A. Piette, Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis, Applied Energy 205 (2017) 323–335.
[39] S. T. Moghadam, J. Toniolo, G. Mutani, P. Lombardi, A gis-statistical approach for assessing built environment energy use at urban scale, Sustainable cities and society 37 (2018) 70–84.
[40] D. Groppi, L. de Santoli, F. Cumo, D. A. Garcia, A gis-based model to assess buildings energy consumption and usable solar energy potential in urban areas, Sustainable cities and society 40 (2018) 546–558.
[41] Y. Zheng, Q. Weng, Modeling the effect of climate change on building energy demand in los angeles county by using a gis-based high spatial-and temporal-resolution approach, Energy 176 (2019) 641–655.
[42] Y. Ahn, D.-W. Sohn, The effect of neighbourhood-level urban form on residential building energy use: A gis-based model using building energy benchmarking data in seattle, Energy and Buildings 196 (2019) 124–133.
[43] K. Adam, V. Hoolohan, J. Gooding, T. Knowland, C. S. Bale, A. S. Tomlin, Methodologies for city-scale assessment of renewable energy generation potential to inform strategic energy infrastructure investment, Cities 54 (2016) 45–56.
[44] Z. Li, S. J. Quan, P. P.-J. Yang, Energy performance simulation for planning a low carbon neighborhood urban district: A case study in the city of macau, Habitat International 53 (2016) 206–214.
[45] M. Wetter, C. van Treeck, L. Helsen, A. Maccarini, D. Saelens, D. Robinson, G. Schweiger, Ibpsa project 1: Bim/gis and modelica framework for building and community energy system design and operation–ongoing developments, lessons learned and challenges, in: IOP Conference Series: Earth and Environmental Science, Vol. 323, IOP Publishing, 2019, p. 012114.
[46] ODYSSEE and MURE data bases, European-Union, `http://www.odyssee-mure.eu`, [Online; accessed 02-Mar-2020].
[47] R. Braun, V. Weiler, M. Zirak, L. Dobisch, V. Coors, U. Eicker, Using 3d citygml models for building simulation applications at district level, in: 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), IEEE, 2018, pp. 1–8.
[48] J. F. Rosser, G. Long, S. Zakhary, D. S. Boyd, Y. Mao, D. Robinson, Modelling urban housing stocks for building energy simulation using citygml energyade, ISPRS International Journal of Geo-Information 8 (4) (2019) 163.
[49] J. Schiefelbein, J. Rudnick, A. Scholl, P. Remmen, M. Fuchs, D. Müller, Automated urban energy system modeling and thermal building simulation based on openstreetmap data sets, Building and Environment 149 (2019) 630–639.
[50] U. Pont, D. Latzer, R. Giffinger, A. Mahdavi, Assessing energy profiles of urban neighborhoods: A streamlined gis-based approach, in: Applied Mechanics and Materials, Vol. 887, Trans Tech Publ, 2019, pp. 264–272.
[51] R. Nouvel, A. Mastrucci, U. Leopold, O. Baume, V. Coors, U. Eicker, Combining gis-based statistical and engineering urban heat consumption models: Towards a new framework for multi-scale policy support, Energy and Buildings 107 (2015) 204–212.
[52] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, V. Modi, Spatial distribution of urban building energy consumption by end use, Energy and Buildings 45 (2012) 141–151.
[53] J. A. Fonseca, A. Schlueter, Integrated model for characterization of spatiotemporal building energy

consumption patterns in neighborhoods and city districts, Applied Energy 142 (2015) 247–265.
[54] C. C. Davila, C. F. Reinhart, J. L. Bemis, Modeling boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets, Energy 117 (2016) 237–250.
[55] J. Curtis, N. Devitt, A. Whelan, Using census and administrative records to identify the location and occupancy type of energy inefficient residential properties, Sustainable Cities and Society 18 (2015) 56–65.
[56] G. Recchia, M. M. Louwerse, A comparison of string similarity measures for toponym matching., in: Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place, 2013, pp. 54–61.
[57] N. Gali, R. Mariescu-Istodor, P. Fränti, Similarity measures for title matching, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 1548–1553.
[58] J. Egan, D. Finn, P. H. D. Soares, V. A. R. Baumann, R. Aghamolaei, P. Beagon, O. Neu, F. Pallonetto, J. O'Donnell, Definition of a useful minimal-set of accurately-specified input data for building energy performance simulation, Energy and Buildings 165 (2018) 172–183.
[59] F. Wang, Y. Yu, X. Wang, H. Ren, M. Shafie-Khah, J. Catalao, Residential electricity consumption level impact factor analysis based on wrapper feature selection and multinomial logistic regression, Energies 11 (5) (2018) 1180.
[60] M. Fowlie, M. Greenstone, C. Wolfram, Do energy efficiency investments deliver? evidence from the weatherization assistance program, The Quarterly Journal of Economics 133 (3) (2018) 1597–1644.
[61] O. A. Omitaomu, B. R. Blevins, W. C. Jochem, G. T. Mays, R. Belles, S. W. Hadley, T. J. Harrison, B. L. Bhaduri, B. S. Neish, A. N. Rose, Adapting a gis-based multicriteria decision analysis approach for evaluating new power generating sites, Applied Energy 96 (2012) 292 – 301, smart Grids.
[62] K. Shiraishi, R. G. Shirley, D. M. Kammen, Geospatial multi-criteria analysis for identifying high priority clean energy investment opportunities: A case study on land-use conflict in bangladesh, Applied Energy 235 (2019) 1457–1467.
[63] T. Ayodele, A. Ogunjuyigbe, O. Odigie, J. Munda, A multi-criteria gis based model for wind farm site selection using interval type-2 fuzzy analytic hierarchy process: The case study of nigeria, Applied energy 228 (2018) 1853–1869.
[64] A. Rikalovic, . OSI, S. Popov, . LAZAREVI, Spatial multi-criteria decision analysis for industrial site selection: The state of the art, in: XI Balkan Conference on Operational Research-Balcor, 2013.
[65] Building Energy Rating Certificate Database by SEAI, `https://www.seai.ie`, [Online; accessed 02-Mar-2020].
[66] K. McDonagh, Geodirectory technical guide, an post and ordnance survey ireland (2019).
[67] Ordnance Survey Ireland, `https://www.osi.ie`, [Online; accessed 02-Mar-2020].
[68] Census of Population 2016 - Profile 1 Housing in Ireland by Central Statistics Office, `https://www.cso.ie/en/releasesandpublications/ep/p-cp1hii/cp1hii/hs/`, [Online; accessed 02-Mar-2020] (2016).
[69] Better Energy Homes Scheme and Better Energy Warmer Homes Scheme by SEAI, `https://www.seai.ie`, [Online; accessed 02-Mar-2020].
[70] SEAI, Energy in the residential sector report, `https://www.seai.ie`, [Online; accessed 01-February-2019] (2018).
[71] U. Ali, M. H. Shamsi, C. Hoare, F. Alshehri, E. Mangina, J. O'Donnell, Application of intelligent algorithms for residential building energy performance rating prediction, in: 16th IBPSA International Conference and Exhibition Building Simulation 2019, IBPSA, 2019.