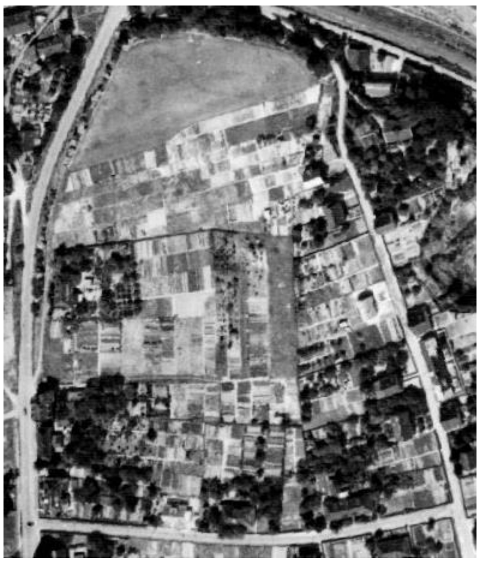
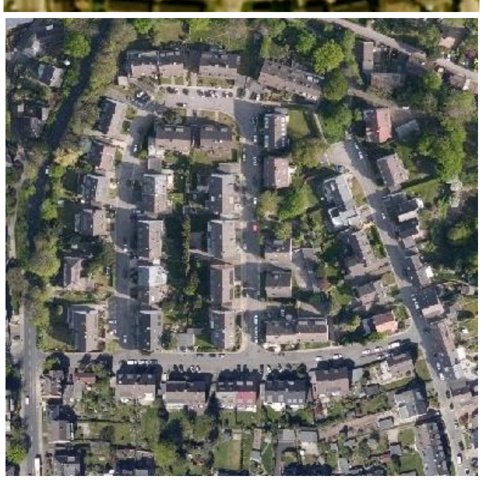





LARGE-SCALE BUILDING AGE CLASSIFICATION FOR URBAN ENERGY DEMAND ESTIMATION

Geo and satellite data for building age identification

OANA-MIHAELA GARBASEVSCHI

COVER IMAGE 'EVOLUTION OF BUILDING CONSTRUCTION BETWEEN 1926 AND 2019 IN A NEIGHBORHOOD IN THE REGION OF ESSEN, GERMANY' COMPOSED OF IMAGES AT SCALE 1: 1000 RETRIEVED FROM HTTPS://LUFTBILDER.GEOPORTAL.RUHR/. COPYRIGHT 2020 REGIONALVERBAND RUHR.

LARGE-SCALE BUILDING AGE CLASSIFICATION FOR URBAN ENERGY DEMAND ESTIMATION

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Engineering and Policy Analysis

Faculty of Technology, Policy and Management

by

Oana-Mihaela Garbasevschi

Student number: 4741803

To be defended in public on March 27 2020

Graduation committee

Chairperson	: Prof. dr. Willem Korthals Altes, Architecture and the Built Environment
First Supervisor	: Dr.ir. Trivik Verma, Policy Analysis
Second Supervisor	: Dr.ir. Iulia Lefter, Systems Engineering
External Supervisor	: Dr. Jacob Estevam Schmiedt, Qualifications, DLR
External Supervisor	: Dr. Michael Wurm, Remote Sensing Data Center, DLR
External Supervisor	: Dr. Björn Schiricke, Qualifications, DLR

TU Delft

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center

ACKNOWLDGEMENTS

I have benefited during my work for the graduation thesis from guidance on all aspects of the research and I would like to acknowledge in this foreword these individual contributions. Prof. dr. Willem Korthals Altes had the difficult task of anchoring the analytical results to the building policy context and his insights have helped me build a critical approach towards the usability of the model for decision making. D.ir. Trivik Verma has been an invaluable help in formulating a scientific work that respects academic standards and his keen sense for balancing the technical and policy aspects of research have led to a thesis that fits well with the Engineering and Policy Analysis goals. Dr.ir. Iulia Lefter's advice on machine learning techniques and formulating better research questions have been pertinent, well-timed and very perceptive. I would like to thank all three of them for their flexibility in dealing with the inherent mishaps and, most importantly, for their availability and sound advice.

The thesis topic has been proposed by the German Aerospace Center (DLR) and I undertook the majority of the research activity on the premises in Jülich, Germany. I have encountered there a professional and flexible working environment where I have been given the liberty to develop my own research ideas while at the same time being guided by experts. I have benefited from Dr. Michael Wurm 's experience in working with urban data and machine learning and his guidance had a great impact in shaping and focusing my research. Dr. Jacob Estevam Schmiedt introduced me to the topic of building energy models and provided insightful reflections on relevant research activities. I would like to thank him especially for his availability for my supervision and for managing my integration with the team. Both him and Dr. Björn Schiricke have welcomed me in their research group and have ensured I had all the necessary means to carry on my research in good conditions. Moreover, I would like to thank the entire DLR team for naturally accepting me as a team member and making our collaboration an enjoyable experience.

Last but not least, my friends and family have been a constant support, and I would like to give special thanks to Aashna Mittal for always giving the right advice at the right time and generally keeping my motivation up throughout the duration of my Master courses.

Since at the moment of writing this entry, we are dealing with unusual and difficult times caused by the coronavirus pandemic, I will end this by expressing how grateful I feel for having been part of a well-managed environment, both at TU Delft and DLR, where the safety of the people is put first, and that I am luckier than millions of people that do not have this chance. Hopefully, by the work put forward by these institutions and many others throughout the world, we will be able in the future to give better chances to everyone.

Oana Mihaela Garbasevschi

Aachen, March 15th 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Urban areas are the biggest consumers of electricity and energy consumption is only likely to increase with rapid urbanization. Out of the urban building stock residential buildings require continuous supply of energy for space heating and appliances. To answer to this demand in a sustainable way policy maker need to design energy efficiency strategies that must rely on accurate and traceable models. These models estimate energy demand based on a series of building features, out of which building age is of prime importance because it predicts the insulation properties of the building.

To support the energy modelling process, we propose a method of automatically identifying building age from spatial data at a large scale. We identify features of buildings that are significant for age prediction and determine which set of features has best prediction power at national scale, in Germany. It is expected that the accuracy of classification will be strongly related to sampling design and data availability. The final results will be used to identify the impact of misclassification errors on estimating energy use in urban energy models, providing in this manner a measure of the reliability of such models.

1. INTRODUCTION

Sustaining basic human activity requires a continuous supply of energy. In order to be able to fulfill our growing energy needs in a natural world that is under pressure and close to resource depletion, we need to achieve a significant change in the manner in which we produce, distribute and consume energy. One direction towards sustainability is to diversify the sources of energy production, by increasing significantly the share of renewable sources (United Nations, 2019). Another direction deals with smart energy management and planning for increased energy efficiency.

Worldwide urban areas account for 71% of energy-related CO_2 emissions (International Energy Agency, 2016). In the EU, buildings account for 40% of energy consumption and 36% of CO_2 emissions (European Commission, 2019). It is expected that by 2050 end-use energy demand to triple compared to 2005 values because of urban population growth and economic development (Creutzig, Baiocchi, Bierkandt, Pichler, & Seto, 2015). This leads analysts to conclude that the "technological, sociopolitical, cultural and ecological drivers of energy transitions will increasingly be urban-related" (Marcotullio et al., 2018).

1.1. PROBLEM IDENTIFICATION

The drivers of urban energy demand are multifaceted and have been categorized into socioeconomic, behavioral, geographical and built environment drivers. The factors associated with the built environment include technologies and materials used, infrastructure, age, type and size of buildings, city planning, street connectivity and accessibility to jobs and services and last but not least, population and employment density (Marcotullio et al., 2018). It is estimated that it is possible to reduce city-related energy demand by up to 50% by 2050 through increasing the energy efficiency of buildings, appliances and distribution networks and also by changes in residents' usage behavior.

The strategies involving buildings in particular are directed towards new ways of constructing new buildings and to retrofitting old buildings. The former measure can lead to reduced energy consumption by 50-70% for singe family dwellings and by 50-90% for multi-family dwellings (Global Energy Assessment, 2012). Building energy efficiency strategies such as renovation must take into account multiple factors such as climate, energy use and building age especially since 90% of EU building stock has been built before 1990 and 75% of it is energy inefficient according to current building standards (Filippidou & Jimenez Navarro, 2019).

Urban energy modelling helps policy makers understand the energy needs of city buildings and infrastructure, plan energy saving programs and analyze the future effect of the planned strategy (Reinhart & Cerezo Davila, 2016). Urban building energy models are constructed by applying "physical models of heat and mass flows in and around buildings to predict operational energy use as well as indoor and outdoor environmental conditions for groups of buildings" (Reinhart & Cerezo Davila, 2016).

The operating principle behind these energy models is that energy consumption in buildings is determined by weather, usage pattern, building geometry, construction techniques and building regulations existing at the time of its construction (Economidou et al., 2011). Building age is an essential parameter for any scenario for modelling energy consumption or refurbishment needs since building techniques and construction materials are an indicator of building energy consumption and these vary between historical periods.

Building age is used in energy modelling to estimate the insulation properties of the building and to inform assumptions around the thermal performance and mechanical system attributes of a building (Tooke, Coops, & Webster, 2014). Thermal insulation varies per building element, and is expressed by the materials' U-values. U-values measure the capacity pf the building element to prevent the transfer of heat between the inside and the outside of a building (Designing Buildings, 2019). Another energy indicator associated with building age is air leakage (Economidou et al., 2011). A building envelope with reduced air leakage performs better in internal temperature control and energy conservation. For these reasons, the ability to identify the age of the building stock is an important factor in estimating energy use and consequentially, energy saving potential.

1.2. KNOWLEDGE GAP

Scientific policy advising requires models that are transparent, unbiased and realistic (Alhamwi, Medjroubi, Vogt, & Agert, 2017). These models should be based on accurate and complete data.

Building stock data is acquired usually by two methods: either through census data collection, a process of registering population and housing statistics, or through surveys, a process focused on a reduced number of buildings and with a specific purpose (Mata, Kalagasidis, & Johnsson, 2014). Statistical information on the non-residential sector is sparser than for the residential sector. Moreover, energy modelling for non-residential buildings is a complex process since these buildings form a more heterogeneous group due to differences in floor size, usage pattern, energy intensity or construction techniques (Loga, Diefenbach, Stein, & Born, 2012). For these reasons, this research is concerned with the residential building stock.

Despite efforts of data acquisition through census and surveys, the year of construction or the building is not information readily available for modelling purposes. The reasons behind this lack of data are the high costs of surveys, un-uniform administrative procedures, and privacy concerns. Data availability varies both at national and regional level. One positive example is The Netherlands where the Dutch Land Registry and Mapping Agency provides information on the construction year for individual buildings (Kadaster, 2019). In Germany, this information is available sparsely, and whenever the information is collected, it is not public due to privacy laws protecting the publication of data on individual addresses, or it is available at an aggregated level (Zensus, 2011). Using incomplete or aggregated data in modeling energy consumption estimates introduces various degrees of uncertainty, depending on the resolution of the simulation, with higher uncertainties for simulation of single buildings (Reinhart & Cerezo Davila, 2016; Zirak, Weiler, Hein, & Eicker, 2020).

The need to reduce this source of uncertainty in energy modelling has led to various studies for the automatic identification of building age. The general trend is to deduce the age of a building from various physical characteristics, like shape (Alexander, Lannon, & Linovski, 2009; Biljecki & Sindram, 2017; Tooke et al., 2014), position (Rosser et al., 2019) or façade appearance (Li, Chen, Rajabifard, Khoshelham, & Aleksandrov, 2018; Zeppelzauer, Despotovic, Sakeena, & Koch, 2018). The prediction is realized by learning these characteristics from buildings with known age and then applying the learning result on buildings for which the age is not known. The literature overview highlights a common feature of the cited studies, which is that they have been performed only for single cities or neighborhoods and the power of generalization of their method is either not discussed or only briefly addressed. The accuracy of predictions across all age classes varies between 50% and 77%. This suggests that is not yet possible to determine precisely the year of construction based on spatial or image-related characteristics.

The aim of this research is firstly an attempt to improve the accuracy of building age identification by exploring building characteristics that are related to their surrounding urban environment. Secondly, the potential of automatic prediction of building age at a large spatial scale will be investigated. The end purpose of the research is to support energy consumption modelling and facilitate decision making surrounding energy efficiency policies. Outside the context of energy applications, building age is a factor to be considered also in scenarios considering: material stocks and flows in the built environment (Ortlepp, Gruhler, & Schiller, 2018), urban resilience, seismic vulnerability, building thermal performance under climate change conditions (Nahlik, et al., 2017), and real estate market valuation. Furthermore, investigating building age constitutes an opportunity for understanding the built environment, its spatial and temporal patterns.

1.3. Research objectives

The research has been carried under the tutelage of the German Aerospace Center (DLR) and its scope is the age prediction of the building stock in selected German cities, more specifically in the federal state of North Rhine-Westphalia. The actual end use of the proposed methodology is however broader than this specific context. It is expected that in the next two decades the highest impact from energy saving policies will be achieved in non-OECD countries (Marcotullio et al., 2018). In the absence of complete or reliable data, the access to automated tools that facilitate and accelerate the process of decision making will have a considerable impact.

To sum up the intent of this work, we aim to identify a set of building features that can be obtained exclusively from public data, that are representative at multiple spatial scales and that conduce to good accuracy of prediction of residential building age. The core of the proposed investigation can be further on summarized by the following research question:

How can urban energy models be improved by accurate and automatic residential building age identification at a large scale?

Starting from the main research question, the workflow is structured in several phases surrounding a central sub-question:

1) What is the influence of construction year on building energy efficiency?
2) What model parametrization is most suitable for the automatic classification of building age?
3) What features are relevant for the classification of building age and what is the prediction success for different groups of features?
4) How accurate is the classification and what is the model's power of generalization across different spatial scales?
5) What is the effect of misclassification of building age on energy demand estimation?

1.4. THESIS OUTLINE

The report has been introduced by an overview of the complex sociotechnical problem which forms the context for our research: transition towards energy efficiency in urban areas. The motivation behind the focus of this work, building age prediction, has been briefly addressed, as well as the knowledge gap that makes our investigation scientifically relevant. These aspects are further on developed in Chapter 2 with a broader introduction to the field of building energy efficiency and a review of the literature that supports and informs the research task at hand. The research methodologies employed are presented in Chapter 3 with an emphasis on modeling technique and data collection and interpretation. The discussion surrounding data processing continues in Chapter 4, with an overview of data sources used and their integration into a homogeneous dataset. The extraction of building attributes, also called classification features is then presented. The chapter also contains an analysis of the optimal parameters and sampling strategies for the chosen model. Chapter 5 summarizes the validation tests performed in order to judge the accuracy and generality of the model. Chapter 6 consists in a discussion of results and their relevance in answering the proposed research questions. The perceived limitations and suggested directions of future research conclude this chapter. Lastly, Chapter 7 is an overview of the study's main findings, both in terms of technical achievements and importance for model-based policy analysis.

2. LITERATURE REVIEW

In order to understand the connection between building age and energy policies a deeper dive is needed into the fields of building energy regulations, energy models and energy policies tailored for residential buildings. Whenever the context allows for country-specific explanations, the spotlight will be on Germany. After giving an overview of the policy context where we situate our research, we focus on the available research methods for investigating building age. These methods include both the traditional avenues of research where building age is the study target and studies of neighborhood urban structure where the age of the buildings is merely implicit.

2.1 BUILDING ENERGY EFFICIENCY

2.1.1. BUILDING ENERGY CODES

Building energy codes are a key policy instrument that enables governments to achieve their energy sustainability targets by improving building energy efficiency while at the same time ensuring comfortable living conditions to building inhabitants (UNECE, 2018). The building elements and energy systems that are the main focus in building energy codes and standards are: building envelope, water heating, lightning, and heating, ventilation, and air conditioning systems (HVAC). More elaborated standards refer also to: level of daylight or solar gains, air tightness, renewable energy, indoor and outdoor temperatures, passive solar systems and solar protection, performance of boilers and air-conditioning systems (Economidou, 2012; UNECE, 2018). The measures imposed are either voluntary of mandatory, and can be of prescriptive nature of performance-based. Prescriptive measures require a minimum energy efficiency level for the individual building components while performance measures refer to an integrated energy assessment of the whole building. The prescriptive approach is easier to implement compared with the performance approach, and refers to regulations of heat loss (U-value) of building component such as windows, roofs and walls, and to the efficiency levels of HVAC systems (UNECE, 2018). The object of both types of measures are new or existing (after refurbishment) residential and non-residential buildings.

Specific regulation on the thermal insulation of the building envelope has been introduced in European building codes in the late 1970s (Filippidou & Jimenez Navarro, 2019). The evolution of these energy-related standards has been unequal for EU member states since then.

Building regulations in Germany are defined in the building control laws, which are divided into material and formal control laws. The material laws describe the building construction standards while the formal laws regulate supervisory procedures and the enforcement of the planning law. They were created at the end of the 18^{th} century based on the laws of the Prussian States, and their first major overhaul took place after the Second World War (Pahl-Weber & Henckel, 2008). The war brought about large-scale destruction of cities, towns and villages and at the same time, a large influx of refugees. The laws developed to deal with the immediate issue at hand were state laws that dealt with firstly with the removal of rubble, with the reconstruction of the infrastructure and then with housing. This phase was followed, in the 60s and 70s, with the construction of "large-scale housing estates and new developments in the urban fringes" (Pahl-Weber & Henckel, 2008). An important moment in the evolution of building codes has been the end of 1970s. Higher energy costs paired with increasing societal concerns for environmental protection, lead to the adoption of prescriptive building energy efficiency measures, in 1977 and 1979. Measures concerning the authorization, research and development and use of renewable energy were introduced in 1996 (Pahl-Weber & Henckel, 2008).

The introduction at European level, in 2002, of the Energy Performance of Building Directive (EPBD) was an important step in the evolution of energy efficiency regulations from prescriptive to performance-based measures. The latest amendments to EPBD were brought in 2019.

Germany, as well as other member states of the European Union, has adjusted its national laws in 2002, in conformity with EU regulations. The Energy Saving Ordinance (EnEv) has been introduced that year, together with a standard method for calculating the heat requirement of a building. The elements under regulation that are related to the age of the building are the U-values of construction elements: external wall, floors, basement, roof, upper ceiling, windows and entrance doors.

Primary energy demand is estimated based on the energy profile of reference buildings, which are representative of the national building stock and are defined as 'typical' buildings for which specific energy performance requirements exist in national legislation. In Germany, in 2016, the maximum primary energy demand equaled 75% of the reference value for 2014 value and the thermal envelope requirements were constrained by a further 20% from the 2014 value (Schettler-Köhler & Ahlke, 2016). These energy performance requirements are cost-optimal according to EU standards (Schettler-Köhler & Ahlke, 2016).

2.1.2. Energy saving measures

For countries in cold climates the first important measure for energy saving is to ensure insulation. Then, when heating or cooling is required, tight air sealing with mechanical ventilation should be considered (UNECE, 2018). For countries in warmer climates moderate insulation would be sufficient while the focus should fall on efficient cooling and heating systems (Filippidou & Jimenez Navarro, 2019). It has also been shown that heat pumps are a technology that is both highly efficient and essential in the process of heat decarbonization (Economidou, 2012). Other measures include insulation of heat distribution pipes, using condensing boilers, heat recovery ventilation systems, or solar thermal systems (Loga et al., 2012).

The greatest part of the European building stocks has been built before 1979, when the first building energy codes have been implemented. Renovations should play an important role in achieving energy targets but their rate is low, at 2.5-3 % and they consist in their greater part only in minor adjustments, thus making it doubtful that their full energy-saving potential will be reached (Filippidou & Jimenez Navarro, 2019). In Europe between 0.4 and 1.2% of buildings are renovated each year which means that more efforts must be consolidated into identifying energy-inefficient buildings. In Germany, the process of refurbishment differs per construction element. For buildings constructed before 1978, approximately 50% had their roof thermally upgraded, while only 20% had walls insulated, with an annual rate of less than 1% (Loga, 2012).

2.1.3. Building age and energy characteristics

In order to understand the need of renovation measure, this section presents the status of the construction year of the building stock across Europe and the energy attributes associated with building age.

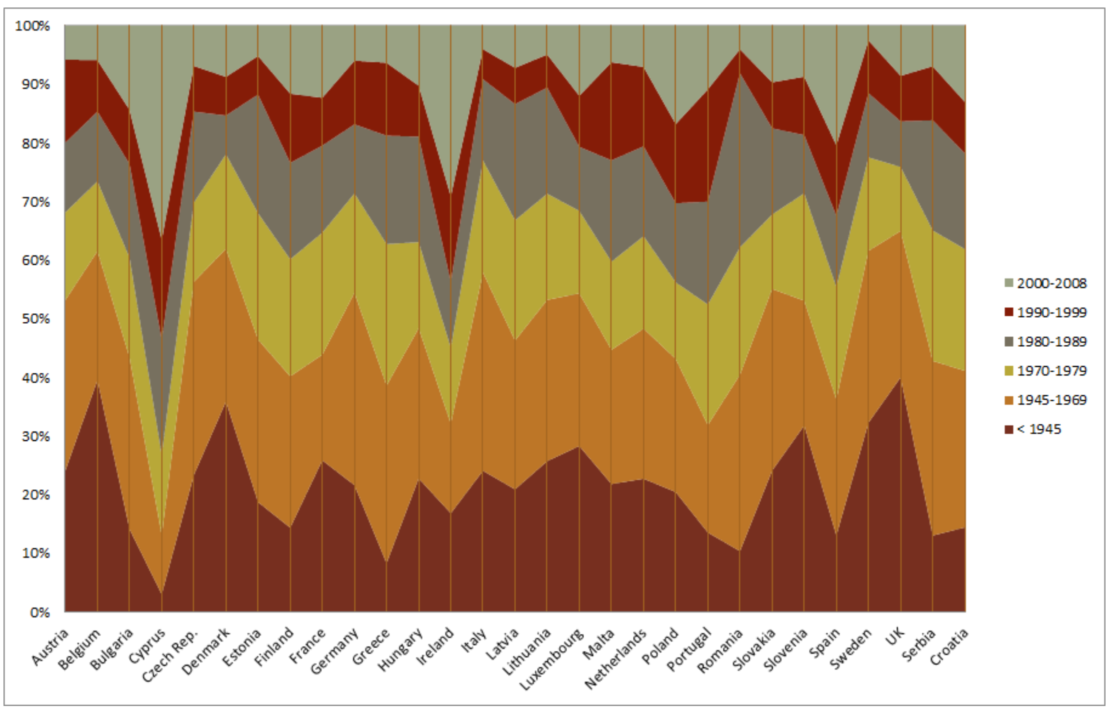


Figure 2.1: European building stock by construction year. Data retrieved from ENTRANZE, http://www.entranze.enerdata.eu/. Copyright Enerdata 2012-2020 (Data tool, 2019).

The introduction in the 70s of thermal insulation requirements in building codes has resulted in a significant decrease in thermal transmittance values (U-values) across all countries (Filippidou & Jimenez Navarro, 2019). All over Europe the U-values of building elements have decreased in time, as it can be observed in figure 2.2 but there are also exceptions to this trend, most notably Germany and Bulgaria, where buildings built in the 1960s are less well thermally insulated than buildings built before that time (Economidou et al., 2011), due to a combination of inferior material technology and lack of maintenance.

Another energy indicator associated with building age is air leakage (Economidou et al., 2011). The decreasing trend of air leakage over construction periods is less pronounced as thermal insulation, and there is a marked difference between old building stocks from Western and Eastern Europe from this point view, where the former has an increased level of air tightness than the later.

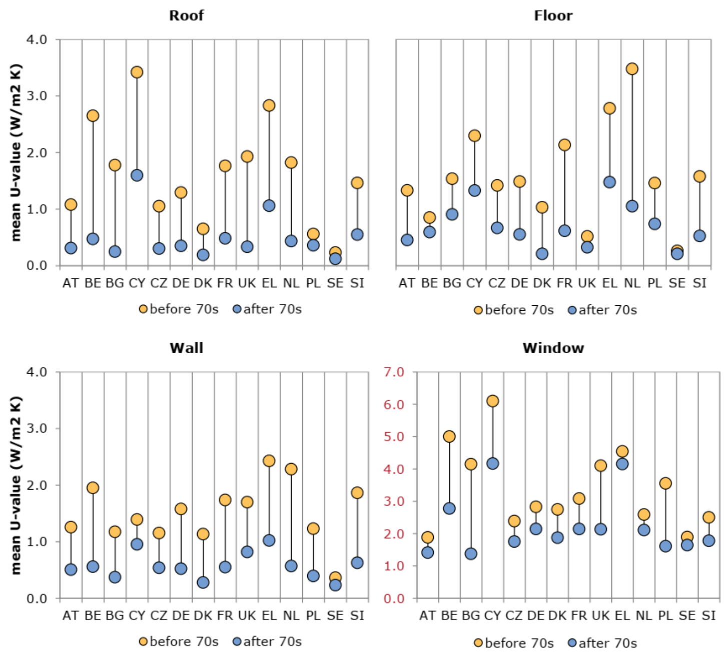


Figure 2.2: Mean U-value of the building's envelope elements (roof, floor, wall and window) calculated as a weighted average per surface area of each element for two subsets (built before and after the 1970s) of the national building stocks. Reprinted from *Achieving the cost-effective energy transformation of Europe's buildings* by Filippidou & Jimenez Navarro, 2019, retrieved from https://ec.europa.eu/jrc/en/publication/achieving-cost-effective-energy-transformation-europes-buildings. Copyright European Union, 2019 (Filippidou & Jimenez Navarro, 2019). Source of data EPISCOPE, http://episcope.eu/welcome/.

2.2 BUILDING ENERGY MODELS

The working principle of a building energy model is to express the requirements for energy as a function of selected input parameters with the goal of "quantifying the consumption and predicting the impact or savings due to retrofits and new materials and technology" so that "decisions can be made to support energy supply, retrofit and technology incentives, new building code, or even demolition and re-construction" (Swan & Ugursal, 2009). Two broad classes of models have been identified, each with their own set of input parameters and simulation methods: top-down and bottom-up. Bottom-up models estimate energy requirements for individual groups of houses and extrapolate the results at national or regional level. These models generally require a high-level of detail in the specification of their input parameters and use complex calculation and simulation methods.

Many of the most commonly used bottom-up models start with the in-detail simulation of energy requirements for individual buildings that represent each of major classes of buildings that form a national building stock. Such buildings are called *archetypes* and their parameters are derived from grouping together different features of the building stock by using available data at national level (Mata et al., 2014). The simulation results for archetype buildings are scaled-up then at regional or national level by multiplication with a weighting factor to achieve the total number of houses in each class.

There are several ways in which these archetypes can be defined and at the present date there is no single international or even European-level uniform characterization of the building stock through archetypes. The largest-scale initiative is the European Tabula Project which identifies an energy-relevant building typology for 13 member states of EU (Loga et al., 2012). For Germany the final result of the Tabula Project is a building classification based on building type (4 classes - single-family houses, terraced houses, multi-family houses and apartment blocks) and construction year (11 classes). Mata et al. (2014) propose another method of segmentation of building stock, based on building type, construction year, main heating system and climate zone and they apply it for 4 European countries: France, Germany, Spain and UK.

In their review of urban energy modeling Reinhart & Davila (2016) show that the combination of building type and construction year is a common factor in most of the proposed building typologies. Clustering construction years into building age classes is normally done based on building regulation changes and is country specific due to the evolution of these regulations in the national historical context (Firth, Lomas, & Wright, 2010).

As mentioned before the purpose of classifying buildings into archetypes is to extract a common set of input parameters for modeling energy requirements of buildings of same type. These input parameters include: heated floor area, external surface, window area, glazing type, ventilation rate, fuels used for the heating system, indoor temperature, and outdoor climate data. Extended parameters may include: "building form and orientation; daylight, solar gains and shading; thermal bridges; internal loads from appliances, equipment and occupants; the performance of different building components and equipment; and the use of renewable energy sources and automatic controls" (UNECE, 2018).

The parameters that are associated with construction year depend on the setup of calculation and simulation used by the energy model employed but mainly refer to the U-values of the construction elements of the building. A short overview of energy models and input parameters inferred from building age is presented in table 2.1.

Table 2.1: Building energy model input parameters inferred from building construction year.

Model	Construction Year Input parameter	Reference
Energy, Carbon and Cost Assessment of Building Stocks (ECCABS)	Average U-value of the building Ventilation rate	(Mata et al., 2014)
Community Domestic Energy Model (CDEM)	Average wall U-value and average roof U-value	(Firth et al., 2010)
SimStadt	Building storey height	(Zirak et al., 2020)
CitySim+	Floor type Glazing ratios	(Rosser, Long, Zakhary, Boyd, & Mao, 2019)

2.3 BUILDING AGE PREDICTION

A building is an individual unit of built environment, positioned in time and in space. It can be described by its interior and exterior, its position and the surrounding environment. It can have cultural, emotional and monetary value attached to it. The construction epoch determines many of the physical attributes of the building as well as the structure of the enclosing neighborhood. Discerning from all the building features the ones that are most representative for the year of its construction is a task that requires subject-matter expertise.

Building and neighborhood characteristics change with the construction epoch and particular attention should be paid to country-specific characteristics.

2.3.1. GERMAN BUILDINGS CHARACTERISTICS

There are approximately 18.9 million buildings with living space in Germany. A great majority of these buildings (38%) have been built between 1949 and 1978, as it can be seen in figure 2.3.

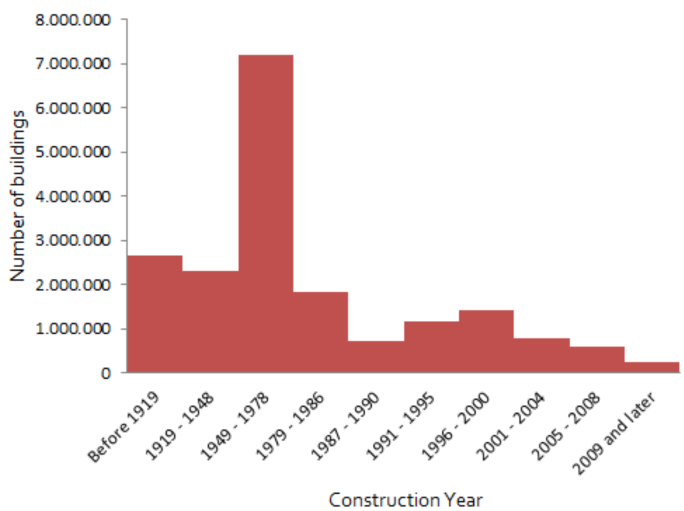


Figure 2.3: German building stock by construction year.

In Germany, between 1850 and 1920 buildings usually had 4 to 6 storeys, arranged along streets and enclosing a block area. Ribbon developments of the 1950s usually consist of 4–5 floors. In the 1970s and 1980s high-rise buildings with large open space between them were constructed (Mueller, Segl, Heiden, & Kaufmann, 2006). World War II changed the shape of many historic cities in Germany and some cities reconstructed their center altogether. The focus after the war was the densification of city centers (Braun, 2015). Another important characteristic of the German building stock are the differences that emerged from the separation between East (German Democratic Republic) and West Germany (Federal Republic of Germany). Eastern Germany followed an intense reconstruction plan driven by the need of social housing which has led to the construction of many high-rise pre-fabricated buildings (Grothe, 2010). These buildings are specific to the East but have also been built more sporadically in the Western part of the country. All these particularities make Germany an interesting and challenging case study for this particular research.

The challenge is visible by merely inspecting a set of representative buildings from different construction periods. With the purpose of providing nationwide building typologies for enabling large-scale energy modeling, the TABULA project classified the German residential building stock into 11 age classes and 4 type classes (Loga et al., 2012). Besides energy data and calculations, the project's web portal (Institut Wohnen und Umwelt , 2020) also offers an overview of representative images of buildings from each category, as displayed in figure 2.5. These images are a proof of the similarities in shape and appearances between age periods and are an indication of the complexity of the task at hand.

Construction Year	Single Family House	Terraced House	Multi Family House	Apartment Block
Before 1859				
1860 to 1918				
1919 to 1948				
1949 to 1957				
1958 to 1968				
1969 to 1978				

1979 to 1983	
1984 to 1994	
1995 to 2001	
2002 to 2009	
After 2010	

Figure 2.4: Building facades representative for the German building typologies identified in the European project TABULA. Data retrieved from TABULA WebTool, http://webtool.building-typology.eu/#bm. Copyright Institut Wohnen und Umwelt GmbH 2012-2016 (Institut Wohnen und Umwelt , 2020).

2.3.2. BUILDING SHAPE AND STRUCTURE

In detail knowledge of floor types, construction elements, location of kitchens, bathrooms, elevators, disposition of balconies, symmetry of the façade structure, decoration tiles, windows shape and location could be an indicator of the year of construction age. However, machine-readable data referring to construction materials or to the plan of buildings is generally unavailable. Real estate databases could contain this information, but their records are most likely to be incomplete, at a small spatial scale and intended for private use, which will not make them a reliable data source for large-scale applications.

Building attributes that are related to shape or location are increasingly easy to find and compute from open spatial data sources due to a growing number of open city data initiatives and the emergence of crowd sourced GIS maps.

There are two classes of studies that have dealt with building age prediction, age being either a continuous numerical value or an age periods, usually counted in decades. The first is defined by the use of spatial datasets and various supervised learning algorithms while the second relies on datasets of images of buildings and deep learning algorithms for image feature extraction and classification.

From the first class, Tooke et al. (2014) have used for predicting building age features related to building shape, 2D and 3D, and information derived from cadaster data and, in a lesser measure, LiDAR data. The authors relate their findings on building age with energy demand estimation and conclude that categorizing building ages into age groups and then defining common energy consumption values per group is not necessarily a good approximation and they suggest individual ages to be better proxies.

Rosser et al. (2019) used a similar approach of classification, based on spatial datasets, while adding a second phase of prediction optimization. The features they used for classification were related not only to building shape but also to their position with respect to other buildings, in a buffer area of 30 m. Starting from the principle that buildings were most likely built together in blocks, they optimize the initial classification results by aggregating results over sets of nearby buildings, where close buildings are determined using three types of proximity measures.

Alexander et al. (2009) explored several possibilities of using spatial information for building age classification. On one hand, they used as predictors the building area and position with respect to other buildings and also to the street, and on the other hand they tested a method based solely on building shape, where the shape is encoded according to its topology. The latter method appeared to be less successful than the former, which is the most common approach in the literature.

From the second class of studies, Li et al. (2018) and Zeppelzauer et al. (2018) have used convolutional neural networks to learn characteristics of a building epoch from patches extracted from images of building facades. The data used was taken either from Google Street View in the first case and from real estate databases in the second. Both studies used only images of single-family dwellings.

Another source of findings that are relevant for building age prediction is literature on the topic of building type predictions. Both directions of research investigate building shape and position, employ supervised or unsupervised learning models and their case studies have similar spatial resolutions.

Amongst these studies, we highlight the research of Wurm, Schmitt, and Taubenböck (2016) whom have used a large set of shape indicators, 2D and 3D, to determine building type. The types they have identified are: perimeter block development, terraced houses, detached and semi-detached and halls. Their conclusion is that more complex geometric features are better predictors of type.

One direction of research in the prediction of building age could include the analysis of the relationship between building age and roof material type and condition. The spectral characteristics of the roof can be extracted from aerial or satellite images. Roof material classification is however difficult due to the wide range of materials that may be used, the steep roof geometry, coloring (red tiles and red colored roofing are not separable), illumination conditions and other properties of the roof surface, such as age, dust or moss that lead to buildings with the same type of roof to appear to have different materials (Armesto Gonzalez, Docampoa, & Canas Guerrero, 2006; Mueller et al., 2006). For properly identifying roof materials hyperspectral sensors are used. Studies that investigate for example urban seismic vulnerability have made use of sensors that capture up to 503 spectral bands (Constanzo et al., 2016). In addition to hyperspectral sensors, for identifying houses within an image, the spatial resolution of the data should be ideally 0.25 m–0.5 m according to Jensen and Cowen (1999) as cited in Mueller et al. (2006). This type of data is costly to be acquired for extended regions.

2.3.3. Urban morphology

Buildings are the essential components of cities and the evolution of city planning is in close relation to the evolution of construction epochs. Investigating building age from the perspective of urban growth is a natural investigation step. Although growing needs of urban settlement, combined with the destruction caused by major historic events such as World War I and II, have led to mixed-age neighborhoods throughout Europe, the history of city planning does follow certain epoch specific considerations.

The 19th century industrial city was an overcrowded space, receiving within its growing boundaries both industrial workers and members of the bourgeoisie. The cities were planned along an orthogonal grid of streets and blocks, with buildings that were compact and similar in size and height (Oikonomou, 2014). Most of the 20th century development has tried to counter-balance this trend, by reducing density, increasing dispersion of settlements and planning for open and green spaces. Towards the end of the last century environmental concerns and a reconsideration of architectural principles lead to increased emphasis on compactness of the urban block for increased community spirit and shorter travel distances (Marshall, 2005). The evolution of the city form was accompanied by the development of new construction materials and techniques and enhanced building design (Oikonomou, 2014).

The urban space is the combination of *blocks*, composed of plots and buildings, and the public space and streets interspersed between them (Oikonomou, 2014). City planning consists in designing the optimal physical form of urban space through a combination of ""size, density, structure and built form" that best accommodate the various urban functions (Marshall, 2005). Identifying, understanding and describing in a quantitative manner the existing urban structures is the purpose of growing body of work, under the term of *urban morphology*. These studies focus on the analysis of individual spatial features, and more recently also to the interrelation between them and between different spatial scales (Berghauser Pont et al., 2019).

There are numerous studies that have defined metrics that can describe urban patterns. We will refer presently to studies that connect explicitly the notion of age of a building or block with the spatial characteristics of the surrounding urban space.

Hermosilla, Palomar-Vázquez, Balaguer-Beser, Balsa-Barreiro, and Ruiz (2014) analyzed the spatial characteristics of different blocks in Valencia, Spain and extracted an age-based classification. The historical center of the city is defined by buildings of different heights, blocks of different geometries, streets that are narrow and reduced green space. Another type of historical blocks is those spread into the fabric of the city and that have buildings of lower heights. At the end of the 19th century a type of grid-shaped block was constructed. The blocks developed in the 50s or the 60s have irregular shapes and buildings of average heights and are bounded by narrow streets. In the 70s and the 80s high-rise buildings were built in block which privileged greater open spaces. The trend of building tall residential buildings continued after 2000 with blocks delimited by wide avenues and an increased share of green space. The last type of residential blocks identified are suburban areas, situated farther from the city center and composed of detached and semidetached houses surrounded by vegetation. When investigating the metrics that best characterize these neighborhoods, they found that street width and area compared to building coverage, as well as the size of the green space are the best differentiators.

Another example of age-based classification for urban blocks has been done by J.H. Lowry and M.B. Lowry (2014) for the region of Salt Lake County, Utah, US. Their classification is limited to 3 types of neighborhoods: pre-suburban (1891-1944), suburban (1944-1990) and late suburban (1990-2007). Their most important of their finding is associated with the fact that World War II marked an important change in real estate development. The suburban neighborhoods are characterized on average by lower house density, increased building lot size and less fragmentation of land uses. From a demographics point of view they also observed an increase in percentage of house ownership but also increased time to commute to work.

Gil, Beirão, Montenegro, and Duarte (2012) compared the characteristics of two neighborhoods in Lisbon, a newly built one and one with continuous development since 1920. By computing a set of block and street metrics they concluded that the biggest differences among the two types of neighborhoods were: the new neighborhood contained long streets with wide pavements while the other contained highly connected streets; with respect to urban density, the old neighborhood contained medium density block with houses with private courtyards, while in the new one there was a mix of high density compact blocks and low density blocks.

Berghauser Pont et al. (2019) researched the block, street and plot patterns in 5 European cities, 3 of which Swedish, investigating measures of urban density and street connectivity. Their typology does not involve a temporal aspect. They found that streets that are well connected in the network (both at local and global scale) are mostly found in dense neighborhoods. These dense neighborhoods occupy however a small share of the city surface, with more than 65% of the surface being either sparsely or compactly occupied by low-rise buildings. They also noted a difference between Swedish cities on one hand and Amsterdam and London on the other hand, where the former has a more of tree-like pattern of streets, with a higher share of dead-end streets, and the later a grid-like pattern.

To sum up the relation between the reviewed literature and the current study, our approach extends previous work in building age detection. The research goes further by putting a stronger emphasis on urban morphology traits concerning street and urban block as mean of identifying age class and also by considering large scale classification. Our interest into block features is in the same vein as the idea of spatial autocorrelation introduced by Rosser et al. (2018). We do not however plan to assign uniformly the same age to the entire block since the urban fabric is not homogeneous and buildings are constructed in gaps in older neighborhoods, especially in areas that suffered from conflict-related destructions. The importance of street features, another novelty point in our research, has not yet been addressed in the building age prediction literature, with the exception of one study (Alexander et al., 2019) that mentioned distance to road as a minor building attribute that has proven irrelevant.

3. RESEARCH METHODOLOGY

The following chapter highlights the methodology used for this study. Particular emphasis is placed on the description and motivation behind the main activities of research, data collection and modeling, and the chapter concludes with an illustration of the research workflow.

3.1. RESEARCH APPROACH

The proposed investigation method is based on a case study and modeling approach. The core of the research work will be of quantitative nature and consists of data gathering, analysis and modelling. Literature review and desk research provide a starting point and context to future findings. The modelling technique employed is multilabel classification of building age.

3.1.1. MACHINE LEARNING

Machine learning is the process through which a machine can change its structure, program or data with the goal of improving future performance (Nilsson, 1996). One practical aspect of machine learning is the extraction of useful information from raw data according to Witten, Eibe, and Hall (2011) who define it as a body of techniques for extracting structural descriptions from examples with the purpose of prediction, explanation and understanding. One important subset of machine learning algorithms is *supervised learning* where labelled input data is used to make predictions for unlabeled new data based on data attributes also called *features* or *predictors* (Géron, 2019). The predictions can be either continuously numeric, in which case the task is that of *regression*, or a finite set of categories, in *classification*. Whenever the label to predict is categorical and non-binary we are dealing with *multilabel classification*.

In their review of machine learning methods used for smart cities and urban sustainability (Nosratabadi, Mosavi, Keivani, Ardabil, & Aram, 2019), the authors show that machine learning is widely used in topics such as energy, transportation, health, environment and city management. Out of all possible applications, smart energy is the field that has made use of practically all methods available in machine learning, including its most recent development, deep learning (Nosratabadi et al., 2019). In another review of machine learning for estimating building energy consumption (Seyedzadeh, Rahimia, Glesk, &

Roper, 2018) the authors conclude that these techniques show great potential for forecasting building energy performance and that they are a viable alternative for classical methods of building energy modelling and assessment.

The literature review we performed on topics concerning building age, building shape and urban morphology highlights a recurring number of machine learning methods employed: *Random Forest* (Biljecki & Sindram; 2017; Rosser et al., 2019; Tooke et al., 2014), *clustering* (Berghauser Pont et al., 2019; Gil et al., 2012), *Convolutional Neural Networks* (deep learning) (Li et al., 2018; Zeppelzauer et al., 2018).

The classification method used for this study is *Random Forest*, an ensemble-based supervised learning algorithm. Its advantages over other types of learning algorithms are robustness to noise, computational efficiency, feature importance estimation and treatment of both categorical and continuous data (Breiman, 2001). It is a method conceived to deal out-of-the-box with multilabel classification and it handles well high data dimensionality and multicollinearity of features (Belgiu & Dragut, 2016). Random Forest is also widely used in urban remote sensing for classification of hyperspectral images for the investigation of urban land cover (Tooke et al., 2014).

3.1.2. DATA COLLECTION

Some of the main challenges in data-driven decision-making are the availability, quality and interpretability of data. Data unavailability is an obvious reason for weak modeling results. Data interpretability is essential for obtaining relevant results and avoids misleading the decision-making process. Data quality impacts both the accuracy of the end result and the amount of time and effort to invest in a project. The quality aspect refers not only to accurate information but also to homogeneous standards and formats.

Data availability was the key driver in establishing the scope of research. After investigating the availability of open urban data in various states across Germany, North Rhine-Westphalia has been chosen as test case owing to its open data policy. Data interpretability has been a challenge in two aspects. Firstly, the language barrier has proven a retardant in the data collection process given that the majority of data portals were only in German. Secondly, for some data sources the metadata was either lacking or displaced and assumptions about the data attributes had to be made by closely inspecting the actual data. Data quality has not been an issue as far as completeness and consistency are concerned but it can quickly become one with respect to uniformity in data formats and standards especially when extending the analysis across geographical regions.

The notion of spatial scale is an important one for the generality of the method proposed. The aim is to investigate the potential of large-scale classification of building age with reproducible research. Consequently, national and international open data sources were given priority. Whenever the data was not available or unusable due to its low resolution, local data sources have been used, i.e. data repositories provided by the state North Rhine-Westphalia.

Inspection of the built environment can be achieved through a variety of measuring means (Lemmens, 2011): aerial surveys, hyperspectral imaging, LiDAR, thermal imaging, oblique digital images. In detail examination of the urban landscape is resource and time consuming and most research projects have to deal with a low cost-high data resolution tradeoff. The types of data sources used are: digital topographic maps, digital surface models, buildings models, Open Street Map or similar map products, census data, satellite images. The data types, sources and formats can be consulter in table 3.1. An overview of the process of data integration and classification setup is illustrated in figure 3.2.

Table 3.1: Data types and sources.

Data type	Data source	Spatial extent	Spatial resolution	Format	Technology
Building age	Census 2011	National	100m grid	Tabular Vector (polygon)	Excel Shapefile
Spectral images	Sentinel-2 (Copernicus)	International	10 m	Raster	GeoTiff
Building models		State		3D model	CityGML
Block	ATKIS[1]	State	1:10 000	Vector (polygon)	Shapefile
Address	ALKIS[2]	State		Vector (point)	
Street	OpenStreetMap	International		Vector (point, line)	Shapefile

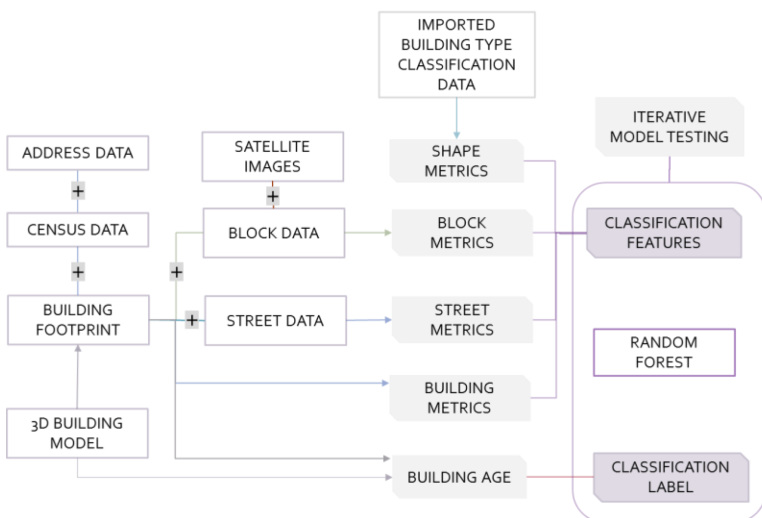


Figure 3.1: Overview of data integration workflow and classification setup.

The advantage of scoping our research to a single state with a good open data policy across all cities within the state was that homogeneity was insured within-data types for all cities. A process of data integration and transformation has to be applied between data types in order to merge all type of data into a single dataset per city analyzed. The city dataset contains building attributes that are considered

[1] Amtliche Topographisch-Kartographische Informationssystem (ATKIS) is the Official Topographic-Cartographic Information System, a basic information system for topographic geodata created by the Working Group of the Surveying Authorities of the States of the Federal Republic of Germany (AdV).

[2] Amtliche Liegenschaftskataster Informationssystem (ALKIS) is the Official Real Estate Cadastre Information System developed by the Working Group of the Surveying Authorities of the States of the Federal Republic of Germany (AdV).

relevant for building age prediction. The combination of building metrics and building age represents the input of the classification model. The research hypothesis identified by the research questions are tested against the model, in different setups that express: variations in spatial scale of classification, complexity of the model features, sampling strategy and parametrization.

3.2. Research workflow

The study is structured by the relevant research sub-questions identified in chapter 1 and the research activities follow closely this outline. For reference, the sub-questions are:

1) What is the influence of construction year on building energy efficiency?
2) What model parametrization is most suitable for the automatic classification of building age?
3) What features are relevant for the classification of building age and what is the prediction success for different groups of features?
4) How accurate is the classification and what is the model's power of generalization across different spatial scales?
5) What is the effect of misclassification of building age on energy demand estimation?

The research flow diagram in figure 3.2 illustrates the connection between research phase, methodology and the question under investigation.

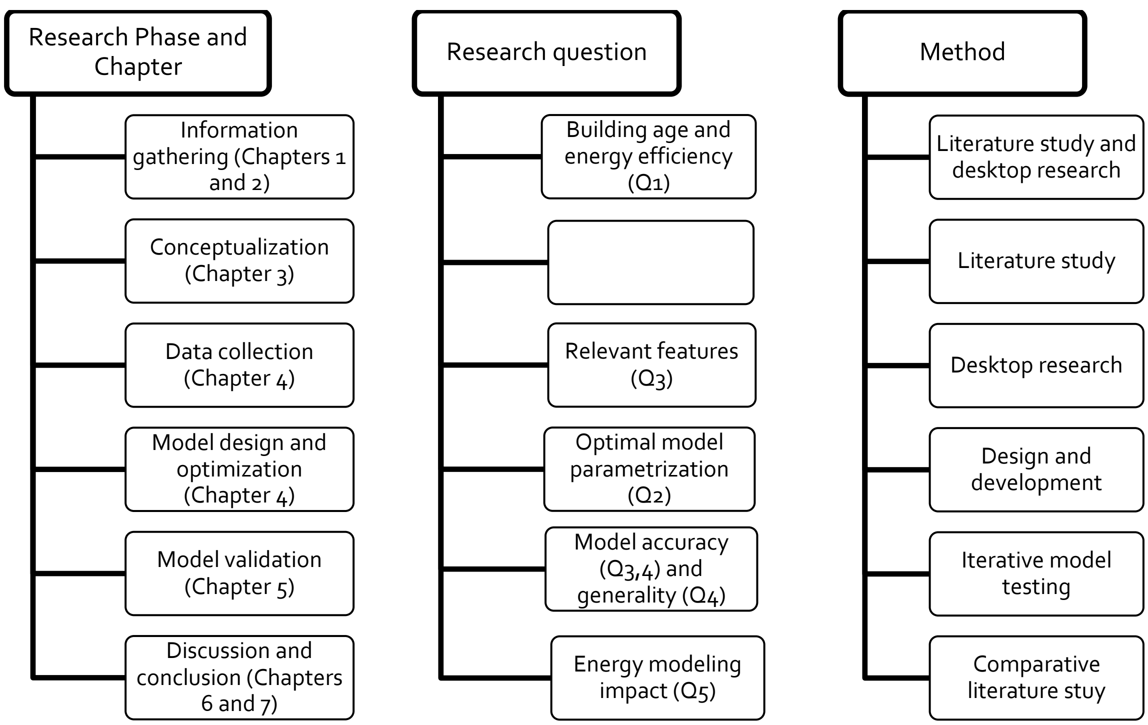


Figure 3.2: Research flow diagram.

4. MODELLING

The current chapter describes the data acquisition, data engineering and modelling workflow. Firstly, the data sources are introduced along with the most important processing steps. Then the features for building classification, either as original data or derived through computations, are presented. Lastly, the parameters of the model and the sampling design are described and evaluated. The best performing learning setup will be used in the next phase of obtaining the classification results.

4.1. DATA SOURCES

Germany is divided into 16 states. North Rhine-Westphalia (NRW) is the state with the highest population and the fourth largest by area. It has hosted the capital of the Federal Republic of Germany, in the city of Bonn, until the country's reunification in 1990.

Data availability differs largely per state in Germany. Since 2014 NRW is developing an open data policy through its Open Government strategy (European Data Portal, 2017). The state's open data web portal hosts more than 3800 datasets from 40 local authorities throughout the state (Open Government Germany, 2019), which makes it the first state in Germany in terms of size of published open data. Because of its size in area and population and mostly because of its data availability, NRW has been chosen as the test case for our modelling.

The data listed in this section is available for all municipalities in NRW. The scope of this analysis is restricted to a set of representative municipalities, for reasons of the limited time available for data processing. From the 10 most populous cities in the state we have chosen 7 that presented the potential of including in the learning phase buildings from all possible construction periods. The chosen cities (figure 4.1) are Cologne, Dortmund, Düsseldorf, Essen, Duisburg, Bielefeld and Münster. The number of inhabitants ranges from 1.08 million inhabitants (in Cologne) to 314 000 (in Münster).

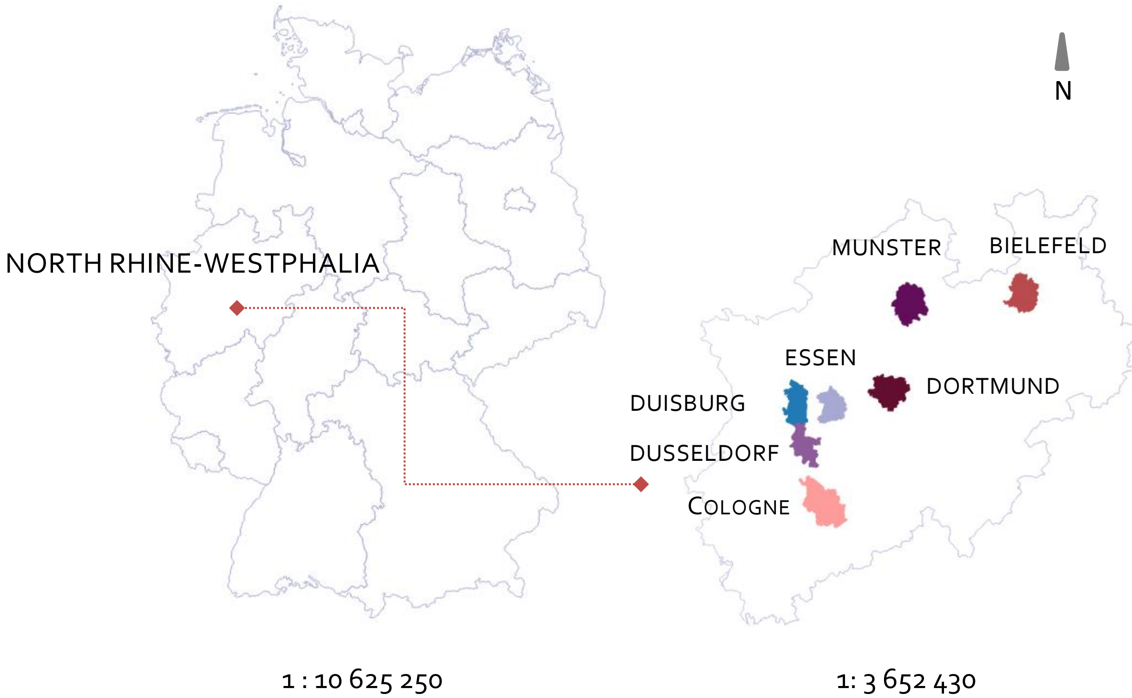


Figure 4.1: Geographical scope of research. On the left, the 16 states of Germany. On the right, the geographical location of the selected cities in the state of North Rhine – Westphalia.

4.1.1. BUILDING AGE DATA

A supervised learning algorithm like Random Forest requires a phase of learning from examples for which the class label is known. This information concerning building age is extracted from the 2011 Census.

The 2011 Census is a national population and housing statistical report (Zensus, 2011). The information is public and presented summarized at a municipality scale and also in a grid format of 100 m cells. The available information refers: buildings and apartments, population demographics and families and households' types. The construction year of residential buildings is given as a range of years and there are 10 groups identified, as depicted in table 4.1. Throughout the report, the concept of age will refer to an age class defined by a set of construction years and not to a single numerical value.

Table 4.1: Age classes defined in the 2011 Census.

Age class	Age description
1	Before 1919
2	1919 - 1948
3	1949 - 1978
4	1979 - 1986
5	1987 - 1990
6	1991 - 1995
7	1996 - 2000
8	2001 - 2004
9	2005 - 2008
10	2009 and after

For privacy reasons, the distribution of building ages over a grid cell is available only as percentage of the total number of buildings. In order to deduce accurate age information for individual buildings, for training we choose areas where a single age class is present, as illustrated in figure 4.2.

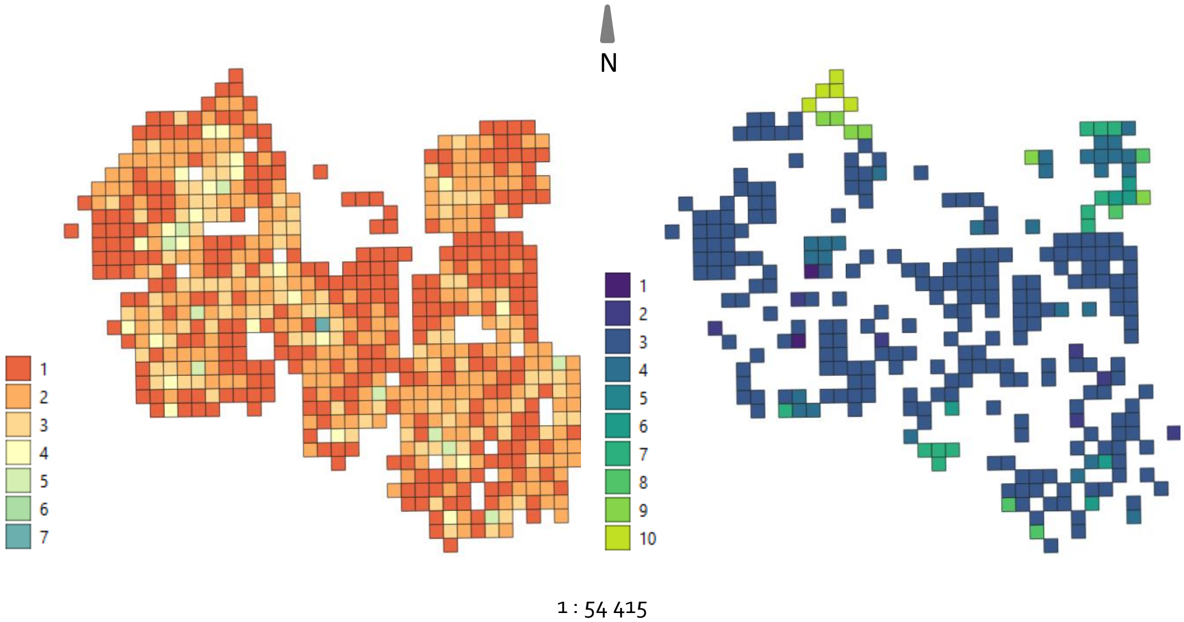


Figure 4.2: On the left, grid cells classified by the number of different ages of buildings enclosed in the area of the cell. On the right, grid cells containing only buildings of the same age, classified by age. Sample of a neighborhood in Cologne.

4.1.2. BUILDING DATA

The main data sources for building information are 3D building models. A 3D building model is a mathematical representation in three dimensions of the shape of a building, its position within the natural or built environment, coupled with other relevant non-geometric building attributes. 3D building models are increasingly being published and made available for open access due to their wide application in domains such as energy modelling, transportation or climate models for cities.

One of the mostly used open data 3D model is the *CityGML* standard, an XML-based format for the storage and exchange of virtual 3D city models (CityGML, 2019). The level of details of the model can go from the most basic one, which includes footprint shape and uniform building height (LoD1), to full specification of roof slopes, annexes and window opening in LoD3. For the purpose of this work we have used data at the LoD2 level, which includes besides the building footprint also the shape of the roof. Figure 4.4 illustrates the representation of building footprints while in figure 4.3 various examples of building 3D models are presented.

The LoD2 data is available for every city in the state from the open data portal of the administration of North Rhine-Westphalia state (Open NRW, 2019). For other cities of Germany either LoD1 or LoD2 data is available, but the practice is not yet spread at national level. In some cases, when the model is not openly available, it can be purchased for a fee, as is the case for the state of Baden-Wurttemberg.

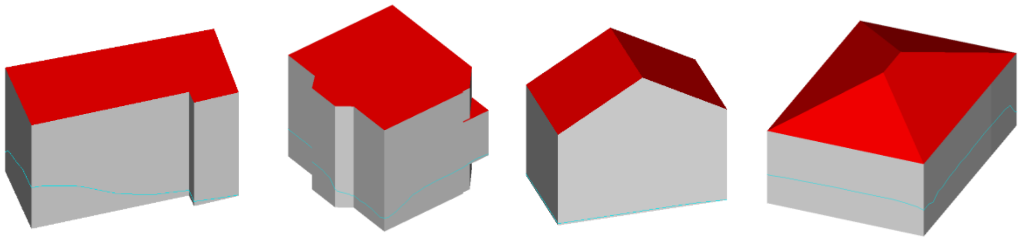


Figure 4.3: LoD3 building models that illustrate different roof shapes.

Other than data concerning building shape and position, this dataset also contains information such as building function, maximum building height, number of above or underground stories. Some information is not however consistently recorded, which made it unusable for our project (e.g. the data of the number of storeys).

The building function data allowed to filter out residential buildings and to compute some of the features needed for classification. The original information for all buildings within the city includes 512 functions. For the purpose of computing numerical metrics associated with blocks, we have grouped these functional classes into 10 overarching classes. There is no single typology of building functions for the non-residential German housing stock due to different "functional, morphological and structural characteristics" (Loga et al., 2012). The classification made for the purpose of this research can be consulted in table 4.2. The classification has been adapted from typologies proposed in a 2009 governmental study on benchmarks for non-residential buildings (BMVBS / BBSR, 2009) and a 2011 study on heated non-residential buildings (BMVBS, 2011).

Table 4.2: Building function classes.

Label	***Name***
1	*Public Facilities*
2	*Education and Research*
3	*Schools*
4	*Hotel, Accommodation*
5	*Public houses, Restaurants*
6	*Buildings for Events and Cultural Purposes*
7	*Sports Facilities*
8	*Retail and Services*
9	*Health Care*
10	*Transport Infrastructure*
11	*Office Buildings*
12	*Office and Administration*
13	*Factory and Industries*
14	*Workshop Buildings*
15	*Warehouses and Garages*
16	*Utility and Miscellaneous*
17	*Agriculture*
18	*House dependencies*
19	*Other*
20	*Residential buildings*

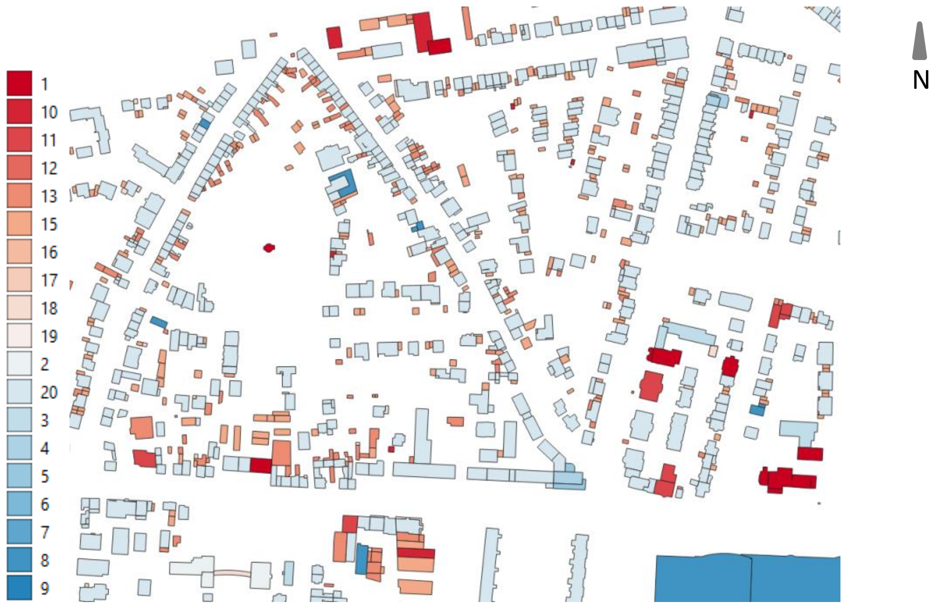


Figure 4.4: Building footprints classified by function class. Sample from a neighborhood in Cologne.

4.1.3. BLOCK DATA

Blocks are administrative areas enclosed by streets and have been extracted from ATKIS Digitales Basis-Landschaftsmodell, an open dataset accessible from the German Federal Agency of Cartography and Geodesy (Open Data, 2019). This digital landscape model describes the topographic objects of the landscape (roads, path, railways, settlements, vegetation) and the relief in vector format, with administrative boundaries up to the municipal level. The open data portal of NWR (Open NRW, 2019) provides freely the model for the entire state at a scale of 1:10 000. The division between urban blocks can be observed in figure 4.5.

Figure 4.5: Blocks of buildings. Sample of a neighborhood in Cologne.

4.1.4. ADDRESS DATA

Another dataset retrieved from the open data portal of NRW region is a collection of points in vector format that reference building addresses for a city region. The addresses have the purpose to uniquely assign a building to a grid cell, and consequently an age class, in cases when the building footprint overlaps two grid cells with different ages. Whenever this is not possible, i.e. no address can be assigned to a building, and the building overlaps two or more grid cells, the building is not added to the sample. For cases when multiple addresses are assigned to one building, if the addresses are situated in different age grid cells, the building is not added to the sample. An illustration of the process can be observed in figure 4.6.

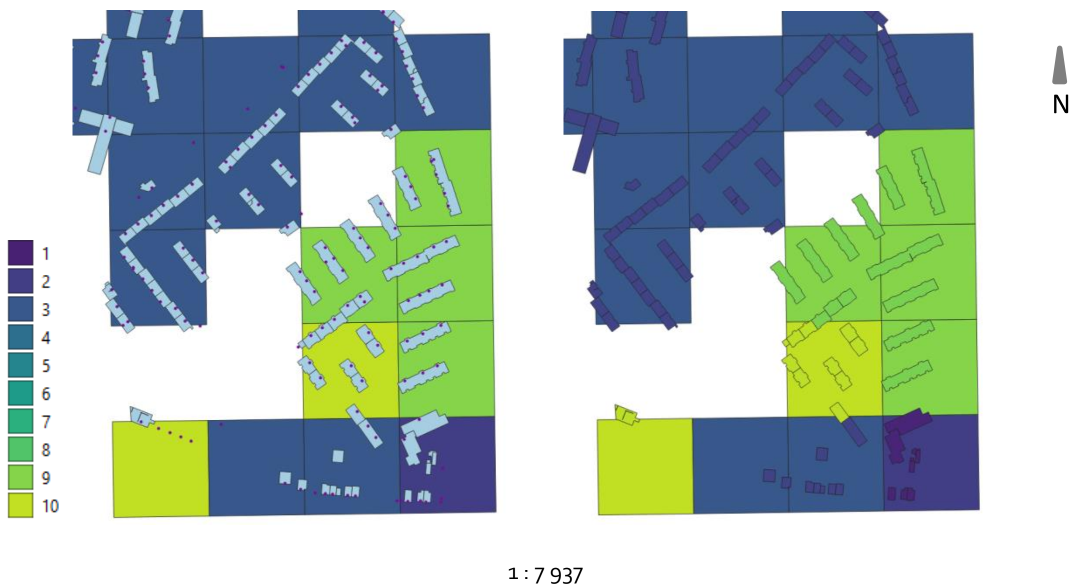


Figure 4.6: Assigning ages to buildings by overlaying vector layers (one layer with building footprints and one with age grid). Address points facilitate assignment when building is overlapping multiple grid cells.

4.1.5. STREET DATA

Street information has been extracted from Open Street Map (OSM) through the APIs provided by the *osmnx* Python library (Boeing, 2017). Open Street Map is a collaborative initiative for the creation of an open editable world map through crowdsources volunteered geographic information. In OSM, roads, streets and paths are identified using the keyword *highway*. The information available, although often incomplete, refers to road segment length, number of lanes, type of road, road name. The *osmnx* library provides a flexible method for extracting the street network inside a city by giving either a location name or location boundary. An example of street network inside a neighborhood is illustrated in figure 4.7. Besides easy access to street information, the library also provides a set of tools for analyzing the network of street nodes and edges. These methods can be applied either to the entire network for obtaining general statistics or to particular nodes or edges in the network. The library was used to extract network properties of the street nodes and segments closest to a given building since we consider the closest street as an indicator of the connectivity options available for the inhabitants of the building.

Figure 4.7: Street network and block limits. Sample of a neighborhood in Cologne.

4.1.6. REMOTE SENSING DATA

The remote sensing data available for this project, covering the entire region of Germany, is the Sentinel-2 data provided openly by Copernicus, the European programme for Earth observation. Sentinel -2 is a European wide-swath, high-resolution, multi-spectral imaging mission launched in 2015. Sentinel -2 carries an optical instrument payload that samples 13 spectral bands: four bands at 10 m, six bands at 20 m and three bands at 60 m spatial resolution.

Out of the four bands sampled at 10 m resolution, blue, green, red and near-infrared, we have only used the red and near infrared bands, for the purpose of calculating the NDVI index. The normalized difference vegetation index (NDVI) is a simple indicator used for assessing whether a surface being observed remotely contains live green vegetation or not (Normalized difference vegetation index, 2019). The value of the index is given by: $\frac{near\ infrared - red}{near\ infrared + red}$. Since the low resolution of 10 m does not lead to a significant description of the surface of a building, the NDVI index is used to give an estimate of the spectral characteristics of the surface of a block, built and non-built areas included.

The NDVI index has also been used to estimate the extent of areas covered with vegetation, as illustrated in figure 4.8. The main purpose of NDVI is indeed to distinguish vegetation from other types of coverage such as bare soil or water bodies, but there is no pre-defined threshold for the numerical value of the index that can be used to assign one type of surface. The general consensus is that values close to 1 are an indication of vegetation, values closer to 0 and slightly negative to water, and small positive values around 0.1 to soil (Wikipedia, 2019).

Figure 4.8: NDVI values (on the left) and vegetation areas computed based on NDVI values (on the right). Sample of a neighborhood in Cologne.

4.2. DATA EXPLORATION

As referred to in the beginning of the chapter, the analysis is focused on a set of 7 cities in NRW. After analyzing the distribution of construction ages in the building stock of the 10 most populated cities in the state, we have chosen cities the following due to the widest range of ages present: Cologne, Dortmund, Dusseldorf, Essen, Munster, Bielefeld and Duisburg. The distribution of building ages for each city is illustrated in figure 4.9. It can be easily noticed that when choosing the sample for classification, the proportions of ages are no longer the same as in whole city residential building stock. The period *"1949-1978"* is extensively represented in the sample, while the percentages of the other age classes and for buildings built after 1979 in particular, are significantly reduced.

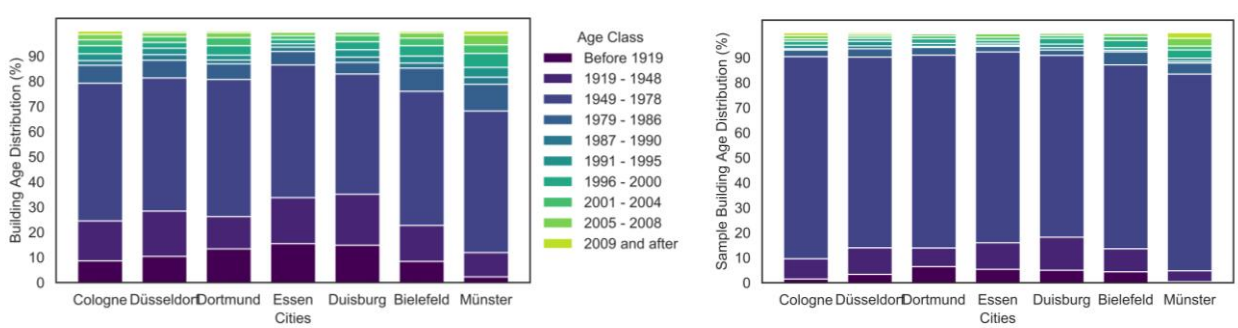


Figure 4.9: Building age distribution for selected cities. On the left, the distribution of ages for all residential buildings in the city. On the right, the distribution of age for the buildings in the sample available for classification.

Data exploration of the building features for Cologne does not indicate any significant pattern that could anticipate the results of building age classification. None of the 89 features exhibits correlation with the age class. The kernel density plots of selected features indicate that some attributes exhibit a larger degree of differentiation than others while still presenting a significant overlap between age groups. Figure 4.10 illustrates the kernel density plots for simple building attributes such as height, ground area, roof angle and ratio height to area.

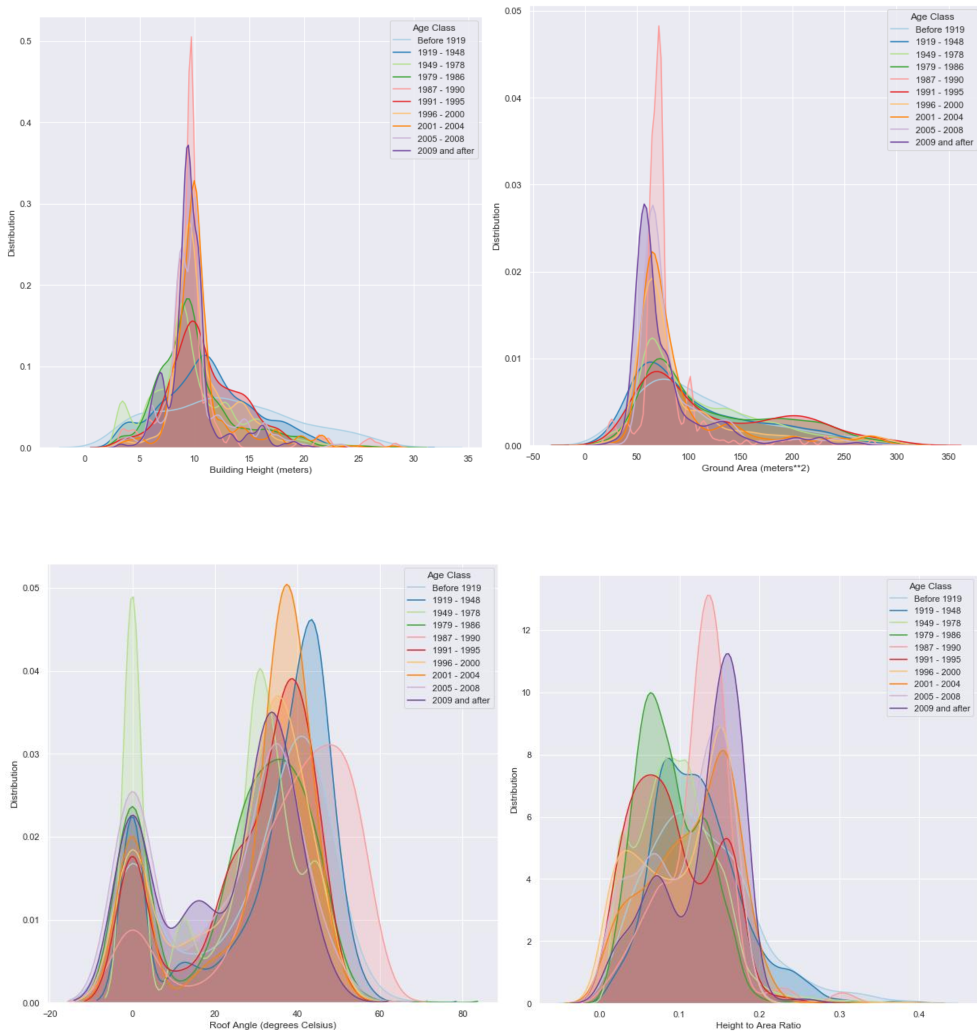


Figure 4.10: Kernel density plot of attributes of buildings in the city of Cologne, grouped by age class. From the top, in a clockwise traversal, the following building attributes are depicted: building height, building footprint area, roof angle and ratio of height to area.

4.3. CLASSIFICATION FEATURES

The selection of classification attributes is an important step for a classification problem and "it is essential to reach agreement on which attributes to use to describe the urban form, how they relate to performance and how to calculate them" (Gil et al., 2012). Ideally the prerequisite of performing a classification task is a "deep understanding of the learning problem" (Witten et al., 2011) and manually selecting features that are "meaningful to the community of experts or practitioners" (Gil et al., 2012).

In the absence of a civil engineering and architecture domain knowledge, the most relevant attributes for predicting building age have been selected through literature research and through consulting with experts into the physics of urban energy modelling and urban remote sensing. The 89 selected features are categorized into four classes: one category deals with the shape of the building, the second with more complex shape and position metrics, the third with characteristics of the streets close to buildings and the fourth with the urban block where the building is situated in. The complete list of features can be consulted in Appendix A.1. The terms *feature* and *metric* will be used interchangeably in the report, and both refer to the same notion.

The first category of metrics includes geometric features related to the shape of the building, 2- or 3-dimensional: height, footprint area, volume, perimeter, shape complexity, density and compactness.

The second set of features includes shape metrics that are more computationally complex, and also a few metrics concerning the building neighbors. Neighboring buildings are buildings that are connected. These metrics have been imported from a DLR MSc thesis project that dealt with the classification of building construction types (Droin, 2019). The final result of classification (building types) and all features have been copied as data attributes to the buildings in our sample, whenever there was more than 90% footprint area correspondence between the building samples in the original project and the current one. The shape metrics are extracted from the work of Angel, Parent, and Civco (2010). The authors argue that one of the most important spatial properties of geographic shapes is compactness and they propose a set of measures to characterize a circle, which they consider to be the most compact of all shapes (Angel et al., 2010).

Street metrics have been computed for street nodes that were closest to the footprint of a building, and more specifically, closest to the vertices of the footprint. Some street metrics have been computed from the characteristics of the street node itself, others from the characteristics of the street to which the node belongs to (e.g. street width or length). Whenever multiple streets or street nodes were retrieved for a building, the metric with the maximum value was chosen to represent the building feature. Street metrics are an estimation of the connectivity of the street network and for a particular building they express the ability for the inhabitants of the building to reach easily or on the contrary, with difficulty the urban transportation network. For this reason, we choose the maximum potential of connectivity from the closest available options.

Block metrics have been extracted from studies that analyzed different urban morphologies, for the purpose of either classifying neighborhoods by age or discover patterns in streets and blocks that characterize neighborhoods or cities.

4.4. CLASSIFICATION SETUP

The following sections discuss the chosen classification model and the methodology used for model validation.

4.4.1. RANDOM FOREST

Random Forest (RF) is a machine learning algorithm built on the ensemble principle, which means that its success rate is obtained by averaging the predictions from multiple models, and more precisely, from multiple Decision Trees.

A *Decision Tree* is a hierarchical model of supervised learning that splits the input space of labeled samples into smaller regions according to the values of the features. A tree is composed of internal decision nodes and terminal leaves. At each decision node a test function with discrete outcomes is used to label the outgoing branches. The process is repeated on each branch until a leaf node is reached. A leaf node contains only samples with the same label.

One reason for which RF is a popular algorithm is because they do not require prior feature processing such as feature scaling or centering. There are also no prior constraints on the data to be used for training; unlike for example a linear model which assumes linear input data. Random Forest has been used extensively in remote sensing applications for the high accuracy of classification and its ability to handle high data dimensionality and multi-collinearity (Belgiu & Dragut, 2016). Compared with single Decision Trees, RF is also less sensitive overfitting.

4.4.2. VALIDATION METHOD

There are several ways to validate a statistical learning model. One learning setup would be to split the dataset into a training set and a test set, train the model on the training set and then estimate the model error on the test set. Another method is to further on split the training set into an actual training set and a validation set. The validation set is used for all model optimization or model selection tests, and then the final error rate of the model is estimated using the test set (Witten et al., 2011).

Another learning scheme is the *K-fold cross-validation* (Kohavi, 1995). The data is split into *K* mutually exclusive and exhaustive equal-sized subsets. These subsets are called folds. The model is trained on the union of *K*-1 folds and tested on the *K*th fold. The process is repeated *K* times, for each fold, and the model error rate is the average over all repetitions.

A version of the cross-validation method is the *leave-one-out validation* where one sample of data is considered a fold, resulting in a number of folds equal with the number of observations. This is a more computationally intensive method which is designed for delivering more accurate results. The gain in accuracy compared to 10-fold cross validation is not always significant as studies have shown (Kohavi, 1995).

In practice, the standard method used, for the balance of accuracy and computational needs, is 10-fold cross validation (Witten et al., 2011).

In problems with multiple classes and an imbalanced class representation in the input data, stratified sampling is a method that ensures improved success rates. This technique is used in combination with *K*-fold cross-validation and consists in assigning to each fold samples from a class in the same proportion as in the input dataset.

All optimization and evaluation tests have been done using stratified 10-fold cross-validation. If splitting into training, test and validation sets or repeated validation is required, these aspects will be specified in the reporting of results.

4.4.3. EVALUATION METRICS

The success rate of a classification algorithm is generally defined as the ratio of correctly labelled observations to size of input data: $\frac{true\ pozitive + true\ negative}{all\ predictions}$. In order to determine if the success rate of the algorithm is different from a random prediction, Cohen's kappa coefficient is generally used. The kappa coefficient measures the agreement predicted and actual classes while correcting for any agreement occurring by chance (Witten et al., 2011).

For problems of multi-class classification where the representation of classes is not equal, the absolute success rate as defined above is a misleading evaluation metrics since it emphasizes the results for the majority classes. For this reason, another single numeric evaluation metric will be used, the *sensitivity* or *recall*, defined as the ratio of correct positive predictions to the total no. of positive predictions: $\frac{true\ pozitive}{true\ pozitive + false\ negative}$. Sensitivity is computed for each class on its own and also as an average for all 10 classes.

Other multi-class evaluation metrics have been computed and will be reported for specific tests: *precision*, defined as the ratio of correct predictions to the total no. of predicted correct predictions, $\frac{true\ pozitive}{true\ pozitive + false\ pozitive}$, and *F1* score, a metric that combines recall and precision into a single value, defined as $2 * \frac{precision * recall}{precision + recall}$.

We consider sensitivity as the most important metric since the target of the project is to be able to correctly identify the age of a building. Sensitivity and precision vary inversely proportional so attention will also be given to the F1 score, which offers a combined evaluation of the two metrics simultaneously.

4.5. Optimization tests

The optimization tests have been performed on the data extracted for the largest city in the region of study: Cologne. These initial tests include: choice of model hyper parameters and sampling design.

4.5.1. Hyper parameters

Random Forest is a nonparametric classification and regression algorithm, which means that the number of parameters is not determined before training and the structure of the model can fit closely to training data (Géron, 2019). In order to avoid this tendency to *overfit*, the algorithm uses hyper parameters to control the structure of the decision trees and of the forest, a process called regularization (Probst, Wright, & Boulesteix, 2019). The regularisation hyperparameters include: the number of trees in the forest, the maximum depth of a tree, the minimal number of samples in a node for the node to be split, the minimum number of samples in a leaf node and the number of features randomly chosen as candidates for a split.

The opposite trend to overfitting is *underfitting*, where the algorithm is not complex enough to learn the structure of the data (Géron, 2019). Both overfitting and underfitting should be avoided.

Performing tests on various combinations of parameters led to the conclusion that, with the exception of number of trees and maximum tree depth, all default parameters suggested in the model's implementation in the Python *scikit-learn* library (Pedregos, 2011) were optimal for our classification problem. The maximum tree depth has been set to 25 to curb the overfitting tendency and the number of trees to 200 to ensure stable results over various input datasets. The test results are presented in Appendix A.2. The metrics followed are average sensitivity over validation sets and average sensitivity on the training set.

Table 4.3: Random Forest hyperparameters.

Name	Value
Number of trees	200
Maximum tree depth	25
Maximum number of features randomly chosen	square root of total number of features
Minimum number of samples for split	2
Minimum number of samples in a leaf	1

4.5.2. Sampling design

Despite the advantages mentioned in a previous section, Random Forest is an algorithm that is sensitive to sampling design, and requires certain conditions to be fulfilled in order to obtain accurate classification results (Belgiu & Dragut, 2016). The training and validation data should be statistically independent; for multi-class classification, the classes should be balanced in terms of number of samples of each class; a high number of data dimensions should be paired with a high number of samples. In this section we will address the issue of class imbalance, while the next section deals with the statistical independence of training and validation data.

The data exploration step (4.2) showed us that we are indeed dealing with an imbalanced learning problem. For the selected cities, the percentage of buildings of age class *"1949-1978"* in the final sample ranges between 73% and 80%. For this reason, age class *"1949-1978"* is considered to be the majority class and all other ages are the minority classes. The ratio of the majority class with respect to the other classes is 5 : 1 at best and 160 : 1 at worst. This is a direct result of the fact the majority class comprises of the three decades when building construction in post-war Germany has been the most prolific.

An imbalanced learning problem is generally approached in two ways: either by assigning different classification costs to classes while training or by resampling the dataset by undersampling and oversampling techniques (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). A RF-specific solution to counter imbalanced learning has been suggested by (Chen, Liaw, & Breiman, 2004) and consists of using balanced samples when boostrapping the samples used for training a decision tree within the random forest.

4.5.2.1. *Undersampling and oversampling*

Undersampling is a straight-forward concept that consists in removing samples of the majority class or classes from the training set. Oversampling is the reversed procedure of adding samples from the minority class or classes to the training set. The simplest way to oversample is by sampling with replacement from the dataset resulting in duplicated observations from minority classes. Chawla et al. (2002) propose an alternative that creates "synthetic" samples for minority classes, meaning the new samples resemble the original observations but are not duplicates. This method is the Synthetic Minority Oversampling technique (SMOTE) and has been shown to perform well for many applications such as network intrusions, sentence boundary in speech, breast cancer detection and in several bioinformatics applications (Blagus & Lusa, 2013).

SMOTE creates new instances of a minority class by a using a *K*-nearest-neighbor approach. A random number of original observations are chosen and for each of their *K* neighbors a new sample is created as a linear combination of the initial observation and its neighbor. The authors indicate that generally a combination of SMOTE and undersampling performs the best (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). We have used SMOTE's implementation provided in the *imbalanced-learn* Python package, which is a very flexible tool that offers the user's the possibility to choose the classes to oversample, for each class the number of new samples to be created and also, the number of neighbors to consider (Lemaitre, Nogueira, & Aridas, 2017).

A pairwise combination of 5 undersampling and 7 oversampling strategies has been tested through 10-fold cross validation on the building dataset of Cologne, for a total of 35 resampling tests. It has been shown that resampling is best to be performed inside the cross-validation loop (Santos, Soares, Abreu, Araujo, & Santos, 2018). This principle has been implemented in the following manner: after making the split between training – with 90% of the data – and validation – with 10% of the data – we undersample and then oversample the training set. Undersampling consists in reducing the number of samples of majority class by a percentage: 0%, 25%, 50%, 75% and 90%. Oversampling strategies include either multiplying the minority samples by a factor: 200%, 300%, 400%, or 500%, or increasing the number of sample for each class such that each minority class is an equal percentage of the training dataset: 5% or 10%. The values have been adjusted slightly for class *"1919-1948"* since it is the most represented class among the minority classes. The test results are shown in figure 4.11.

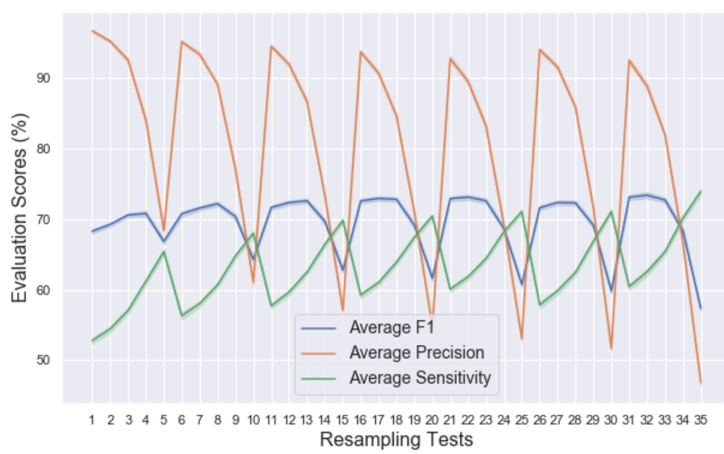


Figure 4.11: Evaluation of different sampling designs.

In figure 4.11 the difference in a group of 5 tests is given by the undersampling strategy. The tradeoff between sensitivity and precision is easily observable. Removing samples from the majority class improves the sensitivity of classification while decreasing the precision rate. The increase and decrease are however unequal. While sensitivity increases in a range of 10%, precision drops by up to 40%. In terms of oversampling, the impact on results is less pronounced, for both metrics.

By sorting the results on average F1 score and giving priority to a modest increase in minority samples, we select as best sampling design the strategy when undersampling is done by reducing the majority class samples by 50% and oversampling adds synthetic samples from the minority classes by multiplying the initial sample size per class with a factor of 400% (a factor of 150% for class *"1919-1948"*).

4.5.2.2. *PENALIZED CLASSIFICATION*

In penalized classification the misclassification of each class will have a different cost to force the algorithm to acknowledge minority classes. We have tested a cost-sensitive classification where the costs are inversely proportional with the distribution of ages in the training set. This means that the rarest of classes will yield the highest cost of misclassification. These costs, or class weights, as they are labeled in the *sklearn* implementation of RF, are multiplied with the probabilities obtained after normal classification.

The influence of classification costs has been tested through 10-fold cross validation on the building dataset of Cologne. The results of our test – displayed in table 4.4 – show that applying classification costs does not have an effect on sensitivity, but that precision is slightly improved.

Table 4.4: Evaluation of penalized classification.

Classification Costs	Average Sensitivity	Average Precision	Average F1
No	62,47 %	86,74 %	72,62 %
Yes	62,24 %	89,69 %	73,47 %

4.5.2.3. *BALANCED TREES*

In their paper, Chen et al. (2004) propose two methods for handling class imbalance when using Random Forest for a classification task. The first one consists in making the classification cost-sensitive, and the second one in undersampling the majority class when growing single trees for classification, the Balanced Random Forest technique.

We have tested the Balanced RF method both individually and in combination with class penalization through 10-fold cross validation on the building dataset of Cologne.

Table 4.5: Evaluation of Balanced trees classification.

Balanced RF	Classification Costs	Average Sensitivity	Average Precision	Average F1
No	No	62,47 %	86,74 %	72,62 %
Yes	No	67,87 %	41,93 %	51,82 %
Yes	Yes	68,08 %	42,56 %	52,37 %

The results of our test (displayed in table 4.5) show that while there is indeed an improvement in the sensitivity score of about 5%, the precision of classification is severely affected. A 40% precision is not acceptable in the context of our problem, where this would imply that numerous old or energy-inefficient buildings (before 1979) would be misclassified as new buildings, as it can be observed in the confusion matrix in figure 4.12.

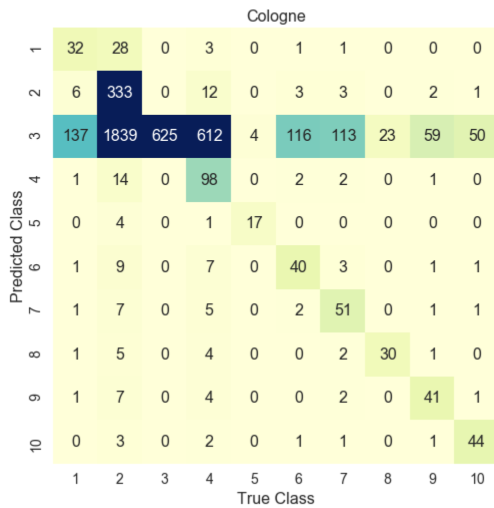


Figure 4.12: Confusion matrix for classification using Balanced Trees approach. Number of samples correctly labelled on the diagonal. Training and testing on Cologne dataset.

4.5.3. Spatial sampling

The idea that buildings close to each other tend to display similar features is an important one for our thesis. This enables us to explore the common features shared by buildings in the same block. The characteristic of data points to dependencies on each other based on their geographical proximity is called spatial autocorrelation and has been shown to lead to optimistically biased prediction results (Pohjankukka, Pahikkala, Nevalainen, & Heikkonen, 2017). The predictor values tend in these cases to be correlated with the underlying spatial structure and leads to model overfitting with non-causal predictors (David R. Roberts, 2016). Standard cross-validation in spatial models does not ensure statistical independence between training and validation data, a prerequisite for many learning algorithms, including Random Forest (Belgiu & Dragut, 2016).

One method of dealing with spatial autocorrelation is to use models that incorporate it into their learning structure as an autocovariate factor, for example autocovariate models, spatial eigenvector mapping, autoregressive models, or phylogenetic least squared regressions (Dormann, et al., 2007). Another is to realize a cross-validation split where there is a clear separation between the data points used for training and those used for validation. Pohjankukka et al. (2017) propose a method called spatial k-fold cross-validation where data points from the training set which are geographically closed to data points in the test set. Roberts et al. (2017) offer a similar technique called block cross-validation where the input dataset is split into non-overlapping

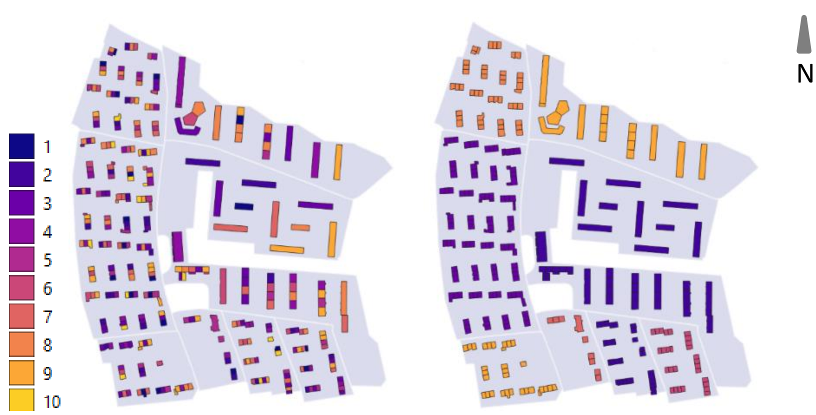
Studies have showed that whenever there is spatial correlation between data, the accuracy of classification is overly optimistic. Removing this correlation leads to more accurate results.

For the building dataset, the most obvious source of spatial autocorrelation is buildings' belonging to a block. Since all buildings in the same block have equal block attributes, buildings in the test set that belong to the same block as buildings in the training set will be classified with high accuracy only on the basis of their shared block attributes. Thus, the classification results will be highly optimistic. Besides block features, a visual inspection of the dataset is enough to highlight the fact that many times buildings of the same shape are found in close vicinity in the same block.

Another source of correlation for spatial reasons would be buildings that share the same street attributes. Buildings aligned on the frontline of a neighborhood, facing the street, are very likely to share the street attributes, since these attributes are computed based on the street closest to the building.

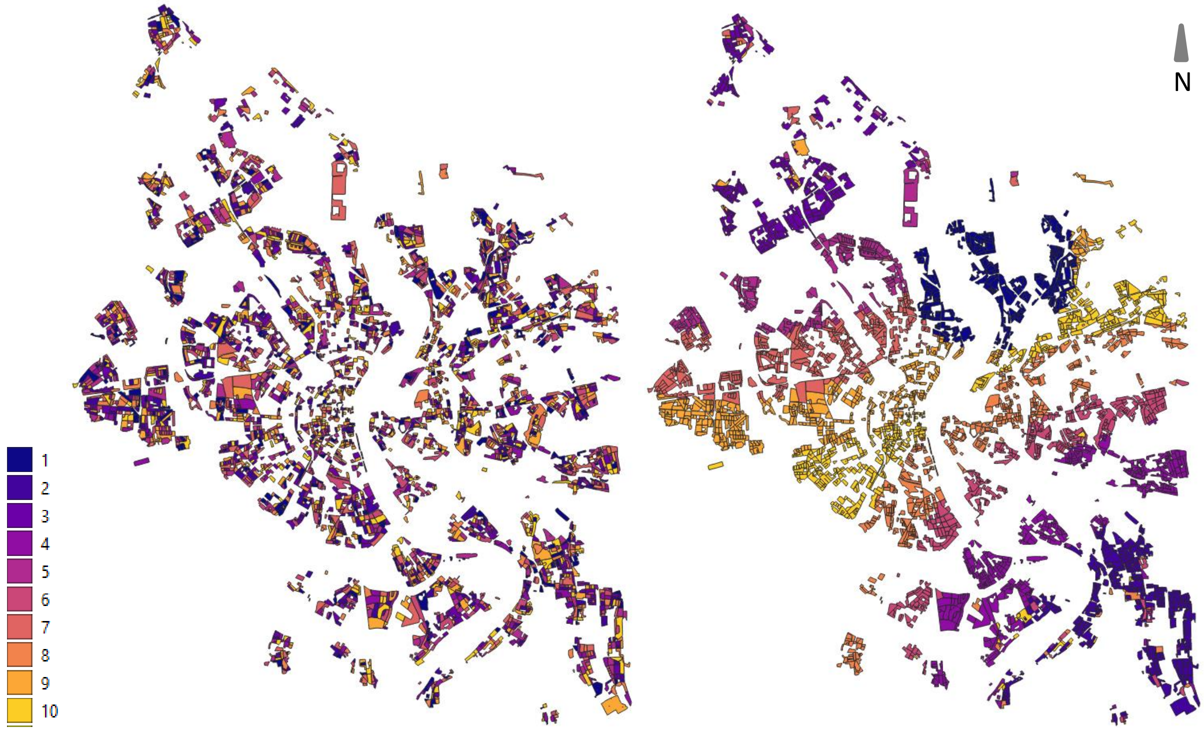
Our method to avoid the aforementioned issues is to ensure that no building in the test set will belong to a same block as a building in the training set. Since testing is done through 10-fold cross validation, the idea behind the spatial sampling is to split the number of blocks in the sample into 10 sets such that each set has approximately the same number of buildings. The challenge with this approach is to also obtain in all 10 sets a similar distribution of age classes.

We have compared four methods of sampling. The first is the normal stratified sampling, where the building stock is split into 10 folds (*standard* method). The second is a spatial sampling (a split of the blocks containing the buildings) where the distribution of ages in a fold is similar (but not equal) between 10 folds (*spatial stratified* method). The third is a random spatial sampling, with no concern for the distribution of ages in a fold, only requiring that each fold contains a minimum number of buildings from each age class (*spatial random* method). The fourth method is a random spatial sampling, similar to the third method, with the added constraint that clusters are formed from neighboring blocks (*spatial distance* method). The principle of spatial sampling is illustrated in figure 4.13 while figure 4.14 shows the difference between the first two methods of spatial sampling and spatial sampling with neighboring blocks.



1 : 2381

Figure 4.13: Grouping buildings into 10 subsets for 10-fold cross validation. On the left standard random sampling of buildings. On the right, spatial sampling with division per block.



1 : 112 848

Figure 4.14: Grouping buildings into 10 subsets for 10-fold cross validation. On the left, spatial sampling with random selection of blocks for folds. On the right, spatial sampling with selection of blocks by proximity.

We test these sampling strategies on 6 models: one with all features, one with all except block features, and one for each category of feature separately (building, shape, street, block). We perform repeated cross-validation (10 times) in order to eliminate the bias introduced by the method of clustering.

We observe how the accuracy of prediction drops between the normal sampling and the two spatial sampling for all models. The deterioration of results (displayed in figure 4.15) was expected in the case of block features, but it seems that for every type of feature there is an underlying correlation of attributes at a spatial scale. This issue is often overlooked in studies dealing with spatial data which leads to overly optimistic classification results.

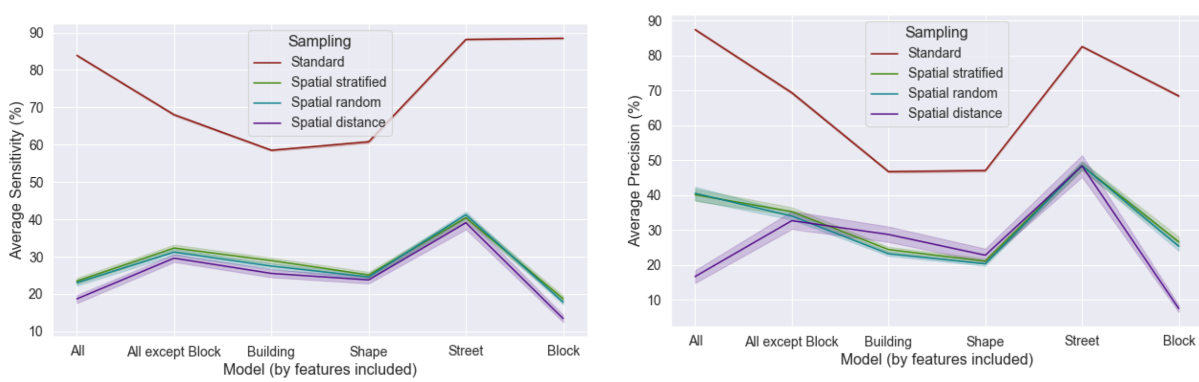


Figure 4.15: Evaluation of spatial sampling: sensitivity and precision average scores per sampling strategy and per model. Confidence interval displayed for 10 test results.

In order to eliminate the possibility that the model is overfitting on the training data and is unable to recognize new data, we have tested several combinations of hyperparameters that have been judged suboptimal in our initial tests (section 4.5) since they were producing underfitting models. The tests have been performed on a reduced model, containing only building features (figure 4.16). Testing the model with the spatial sampling technique shows little variation in evaluation scores, either sensitivity of precision. This proves that the difference in classification results when using normal stratified sampling and using spatial sampling is not a product of an overfitting model, but a consequence of spatial autocorrelation in the building features, which leads to more pessimistic results when this autocorrelation is accounted for and eliminated.

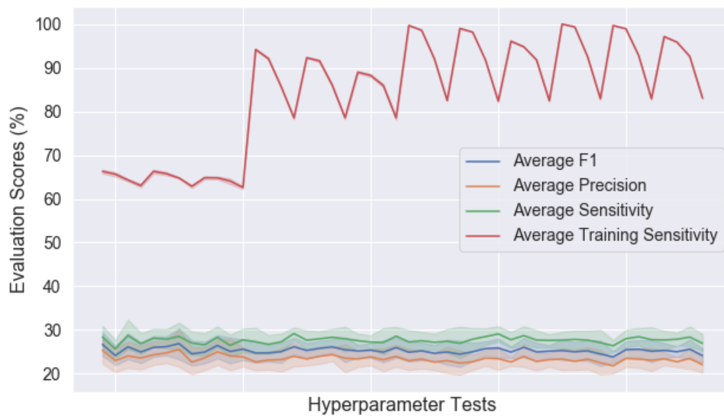


Figure 4.16: Evaluation of spatial sampling for different combinations of sub-optimal hyperparameters.

4.5.4. FEATURE SELECTION

One important step in a learning problem is to obtain the most relevant features for classification through a process called *feature engineering*. By eliminating superfluous attributes, the accuracy of classification is usually improved (Witten et al., 2011). Feature engineering supposes either to choose a subset of existing features (*feature selection*), to combine existing feature into new ones (*feature extraction*) or to gather more data for new feature computation (Géron, 2019). We have chosen to apply a feature selection technique in our study for two reasons: firstly, Random Forest (RF) is an algorithm that does not require preliminary feature processing such as normalization or discretization and also RF deals well with feature collinearity which is one of the main reasons for which feature extraction is used, and secondly, we wish to keep the model's readability in what concerns identifying best features for prediction.

Feature selection is the process of searching the attribute space for the best subset for data classification. Since the number of possible subsets increases exponentially with the size of the feature set, exhaustive search is impossible and multiple search methods have been proposed, with various levels of complexity, e.g. forward selection and backward elimination, best-first search, beam search and genetic algorithms search (Witten et al., 2011).

The method used for this study is a combination of *backward elimination*, the process of removing one attribute at a time from the full set of attributes, and feature ranking. We first obtain an ordered list of features by their importance for classification and then we proceed to eliminate the last feature in this list. The method used for judging feature importance is RF's built-in ranking procedure which consists in reducing *the Gini impurity index*. The index is computed at a node of a decision tree.

Gini impurity is a measure of how often an element chosen randomly from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset (Decision tree learning, 2020). The index is defined as:

$$I(G) = 1 - \sum_{i=1}^{K} p_i^2$$

where p_i is the fraction of items labeled with class $i$in the set.

Random Forest's feature importance method analyzes the decision tree nodes that use a particular feature for split and computes the impurity reduction, then averages the result over all trees in the forest; the average for each node is weighted by the number of training samples that were associated with that node (Géron, 2019).

A common malpractice in supervised learning is to perform feature selection and model testing on the same data which leads to biased accuracy estimates (Kohavi, 1995). When testing is done through cross-validation, it is advised that model selection should be done inside the cross-validation loop, reason for which both feature selection procedures will be done through cross-validation inside the iterations of the initial training-test validation split. The testing setup for model validation will be explained in more detail in the next chapter.

5. RESULTS

After the model's hyperparameters and the best method of sampling have been chosen, we proceed to two types of model validation: for each city, the model is trained either on data from the same city, or with combined data from all the other cities. The second test will be labelled as region classification to indicate the extended spatial scale used for model learning.

While *Random Forest* is the *mathematical model* employed for classification, the concept of *model* we will refer to for the remainder of the report is defined as the ensemble of data attributes the algorithm uses for training and classification. Three models are the starting point of our analysis, irrespective of geographical location, geographical scale or sampling strategy. These models are combinations of the general classes of features identified as relevant for building age prediction. *Model 1* comprises on one hand of building features related to the building appearance and geometry and on the other, on features that describe the shape of the building footprint and its relationship with neighboring buildings. *Model 2* extends the first model by including features of the street or street intersection that are closest to the building. *Model 3* further extends the two models with the addition of attributes that describe the urban block where the building is situated in.

These models can be applied at different spatial scales, where the scale is determined by the geographical location of the training data. When the training process uses data from buildings from the same city as the buildings to be classified, the model is called a *city model*. If training data originates from other locations (other cities), the model will be denominated as a *region model*. There are multiple ways in which to define a *region* for training. We have chosen the most exhaustive manner at our disposition which consists in training the model on the merged data from six cities and testing classification on the seventh.

5.1. CITY MODELS

Deploying individual city models enables the concurrent analysis of several research hypotheses. Firstly, the accuracy of classification of the chosen models will be estimated which consequently will allow observing whether the accuracy differs significantly per geographic location. Secondly, the impact of the spatial distribution of training samples on classification results can be analyzed. Thirdly, the differences in accuracy of classification for individual building ages can be compared across different locations. Furthermore, the confusion between building ages can offer insight into whether construction epochs present the same dissimilarities across geographic locations. Last but not least, the importance of specific building features for improved classification will be tested.

The validation of the model is done through 10-fold cross-validation. The chosen resampling strategy for testing city models is a 50% under-sampling combined with a 400% over-sampling strategy for minority classes. Figure 5.1 demonstrates how the resampling strategy improves the ratios between age classes for all cities.

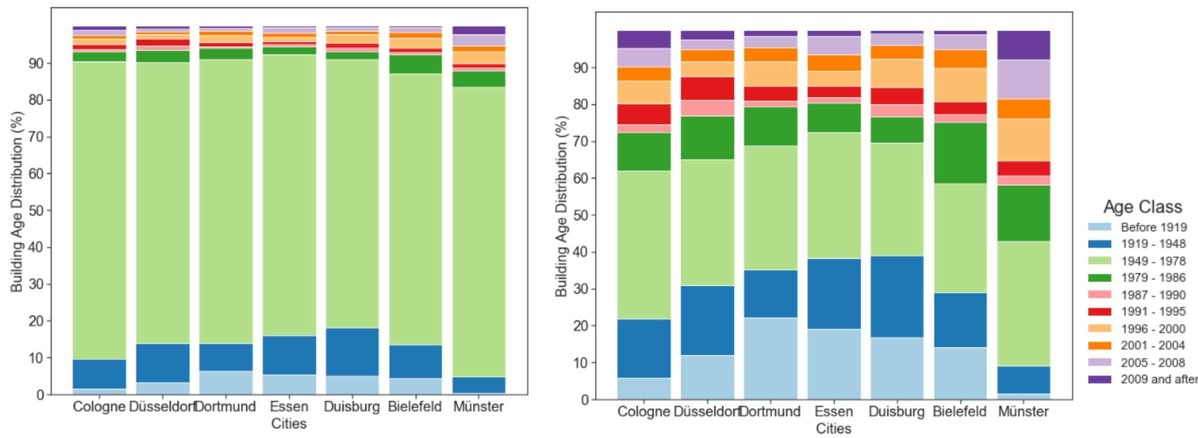
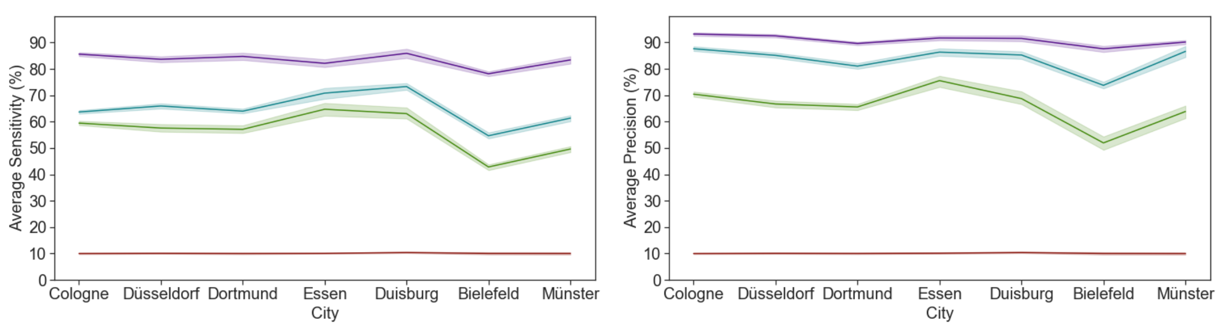


Figure 5.1: Distribution of age classes in the training dataset per city: on the left, distribution in the original datasets; on the right, the distribution after applying the resampling technique.

5.1.1. SAMPLING METHOD

One of the main goals of this investigation is the accuracy of building age classification. Figure 5.2 shows that across all cities the average sensitivity over all building classes varies between 45% and 85%. All three model have been compared with a *baseline model* that classifies randomly based on the distribution of ages in the training data. Across all cities, *Model 3* obtains the highest and most stable results. The other models also show similar trends in prediction across cities, with two exceptions. The results also show that more than 20% increase in accuracy can be obtained when using features with a strong spatial autocorrelation trend, such as the block features.



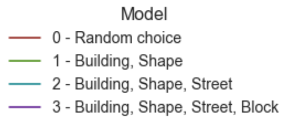


Figure 5.2: Average sensitivity and precision scores for building age classification per city. Normal stratified sampling is used for selecting training data. Confidence intervals are computed based on 10 test results.

In order to further on test the effect of spatial autocorrelation on classification results, a spatial sampling of training data is applied. This method constrains the training and test data to not belong to the same block. Figure 5.3 shows a sharp decrease in both sensitivity and precision for the particular method of sampling.

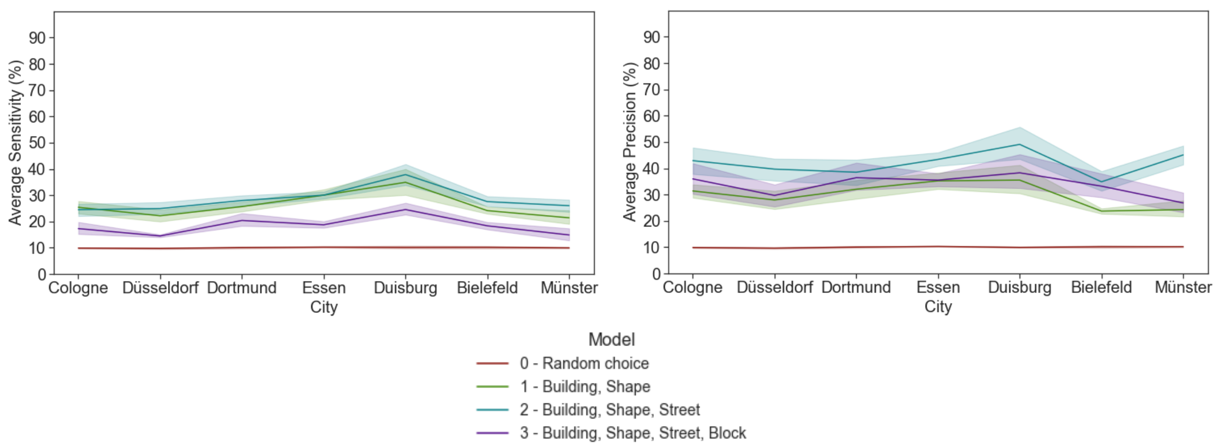


Figure 5.3: Average sensitivity and precision scores for building age classification per city. Spatial sampling is used for selecting training data. Confidence intervals are computed based on 10 test results.

The breakdown of classification accuracy per age class (figure 5.4) shows that age class *"1949-1978"* benefits of a very high accuracy of prediction, followed by the class of second oldest buildings, class *"1919-1948"*. The trends for the other age classes are less homogeneous with weak classification power, at an average of 20% per class.

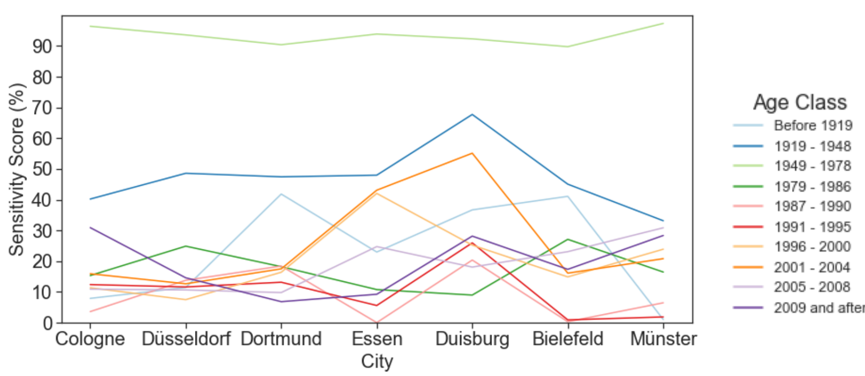


Figure 5.4: Sensitivity scores for each age class for building age classification per city. Normal stratified sampling is used for selecting training data. Results are obtained from classification with *Model 2*.

5.1.2. CLASSIFICATION ACCURACY

While we acknowledge the bias introduced by spatial autocorrelation and its effect on classification results, the following accuracy test are performed using normal stratified sampling. This setup corresponds to a plausible case study where the age of buildings in a neighborhood or city is partially known in the energy model (through house survey for example) and the missing building ages are to be predicted using a classification model. In this scenario, it is very likely that buildings from the same block will be both in the training and the test set.

The accuracies of classification for all models, as indicated by the sensitivity scores, over all age classes and across all cities can be observed in figure 5.5.

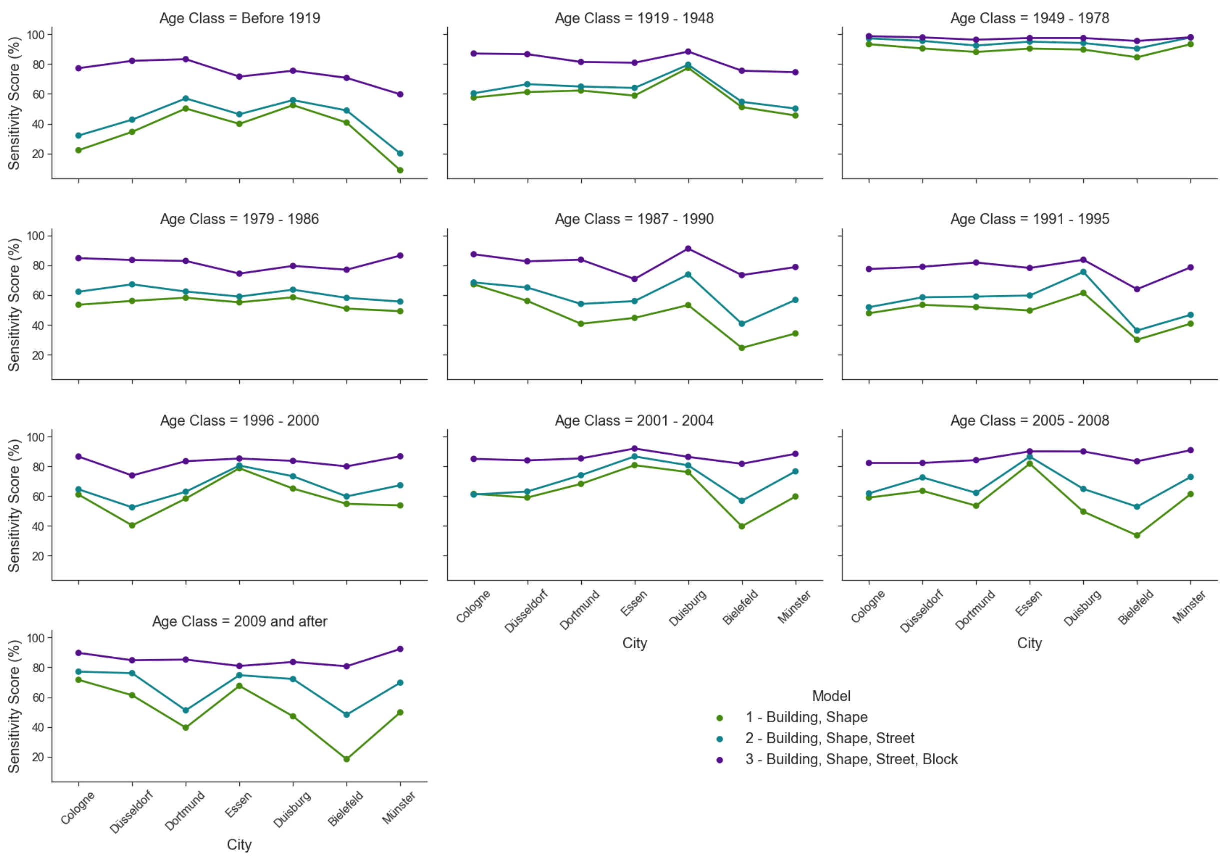


Figure 5.5: Sensitivity scores for each age class for building age classification across cities. Normal stratified sampling is used for selecting training data.

The first observation is that the models perform in a correlated manner for the majority of age classes, even though the accuracies differ in absolute values. In other words, for most classes the hierarchy between classification accuracies is preserved across all models. The building features that extend the first and second models – street and block features – do not provide contradictory information with respect to the core set of building features. This observation is particularly true for *Model 1* and *Model 2* since for *Model 3* there is a reduced variation of results between cities.

The second observation concerns the differences between accuracies across age classes. *Model 3* leads to the best results in classification and the smallest variation in accuracy per age class, with sensitivity scores ranging from 65% to 98% across all cities. For the other two models, the sensitivity of classification per age class ranges from 20% to 95%, with Model 2 performing better by 8% on average. The age class that is best classified is the *"1949-1978"* period, while the weakest prediction results are obtained, on average, for the *"before 1919"* period.

The third observation refers to differences between accuracy of age classification per city. For the majority of age classes, accuracy of classification differs from city to city. The exceptions to this rule are the periods of *"1949-1978"* and *"1979-1986"*. The age classes for which the largest accuracy dissimilarities between cities occur are the period of oldest *"before 1919"* and newest *"2015 and after"* buildings. The differences reach up to 45%, for example when using *Model 1* for the classification of newest buildings between Cologne and Bielefeld. One explanation for the poor accuracy of identifying these buildings in Bielefeld could be the very small sample size in the training dataset (1.5% of the training data). The sample size is not however a determinant of age class accuracy as the classification results for Münster and Duisburg prove it. For both cities the age class *"2015 and after"* is classified by *Model 1* 45% of times correctly, while the training sample sizes are 7.8% for Münster and 0.09% for Duisburg. The differences of results between age classes within a city are best observed in figure 5.6

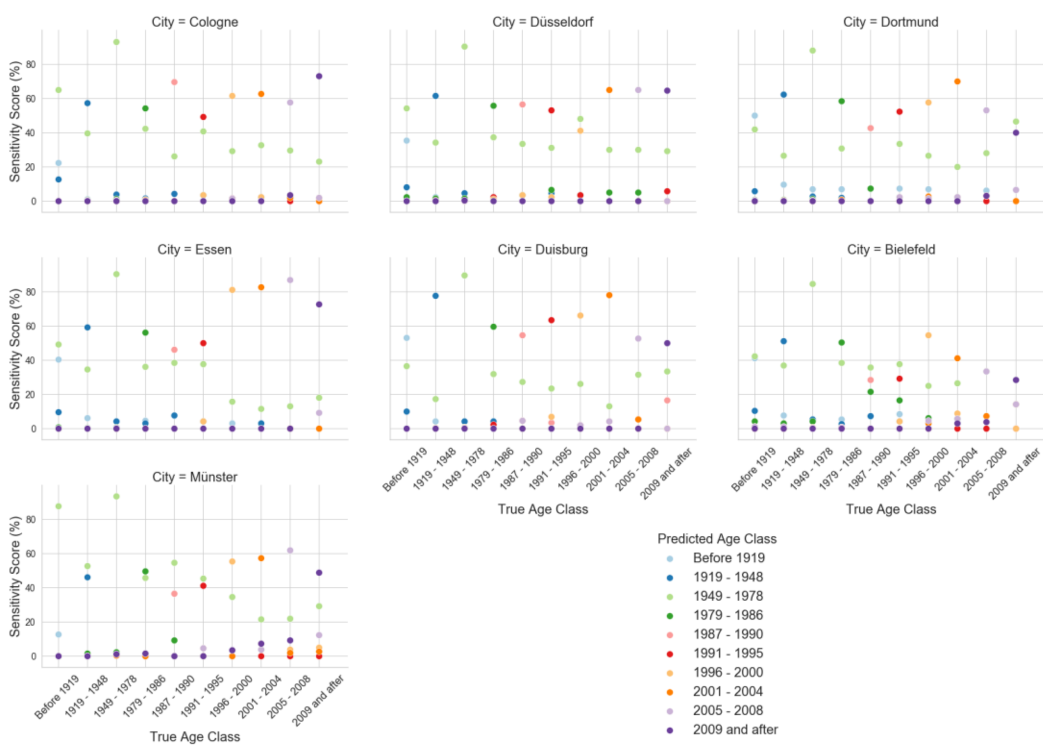


Figure 5.6: Distribution of predicted classification labels for buildings in every age category per city. Normal stratified sampling is used for selecting training data for classification with *Model 1*.

Figure 5.6 illustrates for each city and for each age class the proportion of buildings being labeled with an age class. This alternative representation of a classical confusion matrix enables us to observe which classes are most often misclassified and which classes are they confused with. Misclassification analysis is especially important for *Model 1* results, since it could lead to insight on whether buildings from different construction periods share similar physical features.

It can be observed that the main source of misclassification is buildings being labeled as age class *"1949-1978"*. On average 32% of buildings constructed in the other time periods have been wrongly classified as belonging to this class. For a majority of cities, the other sources of misclassification result in wrong predictions of up to 5% of the tested buildings. An exception is the city of Bielefeld where a greater confusion between classes has been identified. Dortmund is a unique example where misclassification occurs between the oldest buildings and all other age classes, in the sense that about 7% of buildings built after 1919 irrespective of the actual class are classified as being built before 1919. For the two classes of buildings built before 1948, misclassification occurs either with the *"1949-1978"* class or among each other. These buildings are rarely classified as being constructed after 1978. For the buildings built after 1978, the results vary between cities. Bielefeld and Münster for example are two cities where there is significant confusion between these buildings, especially between those built after 1995.

5.1.3. FEATURE SELECTION

Experiments have shown that eliminating data attributes that do not contribute to the classification process better accuracy results are obtained. For this reason, we adopt an inside-the-validation-loop feature selection strategy that consists in two steps: feature ranking by importance and then feature selection by backwards elimination. For reasons of time availability, the technique has been applied only to the first two models, consisting of 44 and 57 features respectively.

The backwards elimination consists in recursively removing the last feature in ranking from the feature set and comparing the classification results for each of the reduced models thus obtained. The inside-the-loop aspect refers to the fact that both steps – ranking and elimination – are performed in every one of the 10 validation-folds through 5-fold cross-validation on the training data. In other words, after splitting (10 times) the dataset into 90% training and 10% test, we split the training data further on (5 times) into 80% training and 20% validation twice, independently one of the other: the first time feature importance is computed, and the second time feature backwards elimination is performed.

The method of feature ranking is the Random Forest built-in feature importance evaluation through Gini impurity reduction. One example of feature ranking by importance for a cross-validation fold of *Model 1* can be observed in figure 5.7. We observe that importance is severely reduced only for the categorical variable *"building type"*.

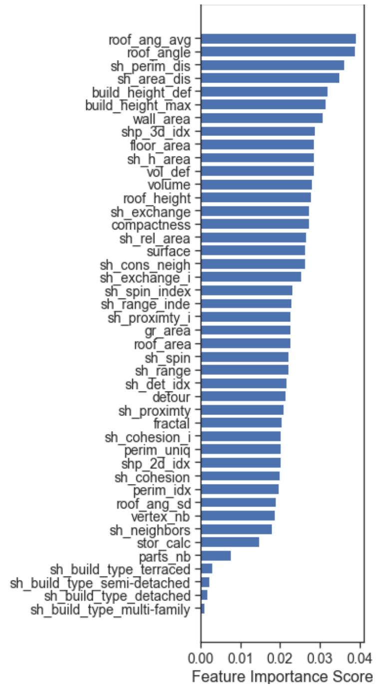


Figure 5.7: Ranking features by importance. Training and testing on Cologne data set with *Model 1*

After the backward elimination accuracy evaluation, it can be noticed that the classification results do not vary significantly for a model with more than 15 features. As a general rule, we choose as best model the model that contains more than 10 features and outputs best accuracy. This selection is done with the purpose of obtaining a flexible model that can accommodate new data. For example, the best result corresponds in the case depicted in figure 5.8 to a subset of 24 features extracted from Model *1.*

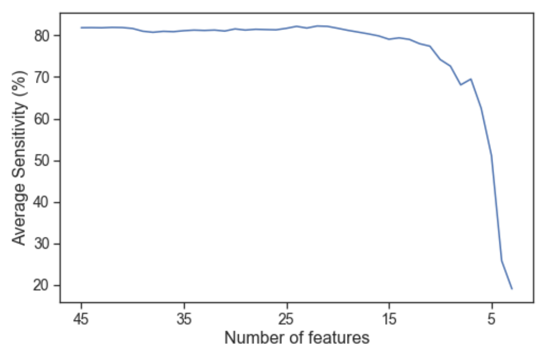


Figure 5.8: Classification results for backward feature elimination. Training and testing on Essen data set.

After the phase of feature selection inside the cross-validation loop, we obtain for each city 10 models with different number of features. The common features of the 10 models are considered as the best model for city classification. Since the number of features in the best model differs per city, we also compute the intersection of best models per city to obtain a unique reduced model for all cities. The single models obtained after applying the procedur to *Model 1* and *Model 2* contain the features listed in table 5.1.

Table 5.1: Common features shared by best models for all cities.

	Model 1
1	*Building Height*
2	*Shape Compactness*
3	*Floor Area*
4	*Average Roof Angle*
5	*Roof Height*
6	*Shape Area Neighbors*
7	*Shape Perimeter Neighbors*
8	*Shape Exchange*
9	*Volume*
10	*Wall Area*

	Model 2
1	*Street Length*
2	*Street Close Centrality*
3	*Street Connectivity*
4	*Street Intersections*
5	*Shape Area Neighbors*
6	*Shape Perimeter Neighbors*

To evaluate the success of the feature selection procedure the original models are tested against the best models for each city and against the model with shared features across cities. As it can be observed in figure 5.9 the results diverge for the two models. The unique reduced model obtained by selecting features from *Model 2* improves accuracy of classification, both sensitivity and precision, for all cities when compared with the full-feature model or the individual city model. For *Model 1* however no difference is observed between the three different model versions. One possible explanation for this phenomenon is that the backwards elimination method has chosen a local optimal model instead of a global optimum. This indicates a limitation of the chosen feature selection technique.

Another explanation can be derived from the nature of the features in the two models. *Model 1* contains only building related features, while *Model 2* contains also street features that have an important spatial autocorrelation component. The reduce version of *Model 2* is comprised mostly of these street features which leads to the conclusion that classification can be done successfully based on attributes that connect the buildings to a location and that are related to their spatial distribution. This conclusion is enforced by the high accuracy results obtained with *Model 3*, where the combination of street and block features make the spatial correlation the main driver of classification. The results of feature selection for *Model 1* indicate that features related only to building shape have a limited potential for correctly predicting the age of the building.

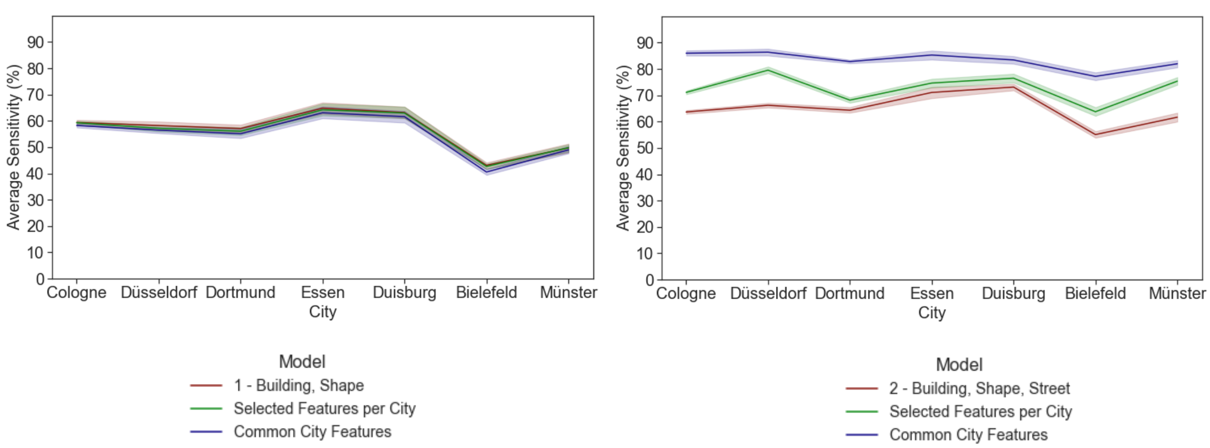


Figure 5.9: Average sensitivity scores for different model versions obtained through feature selection. On the left, versions of *Model 1*. On the right, versions of *Model 2*. Confidence intervals computed based on 10 test results.

5.1.4. Feature importance

For identifying the features most relevant for age classification a simple method will be employed. The method consists in averaging the feature importance results over all tests performed on city models for each of the three base models. For 10-fold cross-validation tests for ten cities by applying two sample methods, normal and stratified sampling, we obtain a total of 140 tests. The subset of 10 features that rank highest after averaging their importance over the complete set of tests are displayed in table 5.2.

Table 5.2: Most important features for classification across all cities and for both standard and spatial sampling.

	Model 1		***Model 2***		***Model 2***
1	*Average Roof Angle*	***1***	*Street Close Centrality*	***1***	*Block NDVI Average*
2	*Shape Perimeter Neighbors*	***2***	*Street Length*	***2***	*Street Close Centrality*
3	*Roof Angle*	***3***	*Street Connectivity*	***3***	*Block Average Volume*
4	*Shape Area Neighbors*	***4***	*Street Intersections*	***4***	*Block Average Height*
5	*Roof Height*	***5***	*Average Roof Angle*	***5***	*Block Maximum Height*
6	*Building Height*	***6***	*Shape Perimeter Neighbors*	***6***	*Block Vegetation Percentage*
7	*Wall Area*	***7***	*Roof Angle*	***7***	*Block Simpsons Diversity (Number)*
8	*Volume*	***8***	*Shape Area Neighbors*	***8***	*Block Simpsons Diversity (Area)*
9	*Shape Compactness*	***9***	*Building Height*	***9***	*Street Connectivity*
10	*Shape Exchange*	***10***	*Roof Height*	***10***	*Block Shape Compactness*

5.2. Region models

The power of generalization of the model is tested by a training-test setup which defines as test data one of the 7 cities and as training data the combination of observations from all the other 6 cities. The re-sampling strategy used for this test is a 75% under-sampling combined with a 500% over-sampling strategy for minority classes. Three models have been tested: 1) all features; 2) all features except block features; 3) selected features that are relevant at city level.

The results are presented in figure 5.10. The best results concerning sensitivity of classification range from 17% to 21%, and they are obtained when excluding block features from the model.

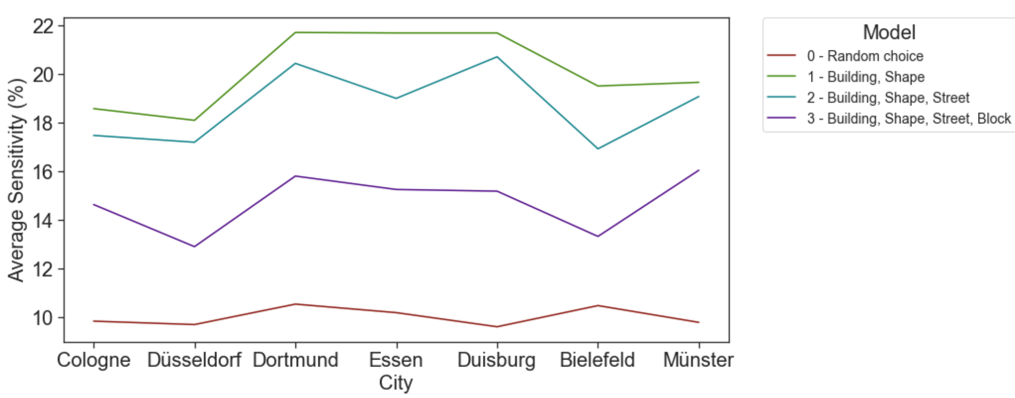


Figure 5.10: Average sensitivity scores for building age classification per region.

When examining the power of classification per age class (figure 5.8), the only conclusion we can draw is that class *"1949-1978"* is accurately identified, followed distantly by class *"1979-1986"* while the power of identification for all other classes is minimal. Inspection of the confusion matrices (e.g. for Cologne as test site, figure 5.11) reveals that while the strongest tendency is to label buildings as class *"1949-1978"* misclassification occurs among all age classes.

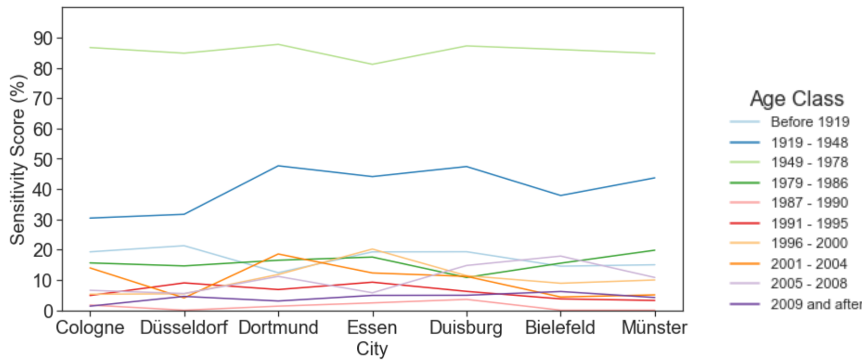


Figure 5.11: Sensitivity scores for each age class for building age classification per region.

5.3. AGGREGATED CLASSIFICATION

The segregation of construction years into building age classes has meaning only when it respects the evolution of the built environment in the national historical context and when it is relevant for a particular modelling application. Age classes that are relevant for building energy efficiency modeling differ from country to country. The constraints imposed by the means of survey, profile of individual research studies and privacy issues also lead to the existence of datasets with different aggregations of building age classes for the same country or region. If a supervised learning model detects actual common trends or dissimilarities across buildings for a particular age categorization, it is expected that the accuracy will significantly change when the model is trained for a different age categorization.

The level of detail of the available data from Census prevents a refinement of the ten age classes into smaller categories and then a possible re-combination of subgroups. It is however possible to aggregate the existing age classes into broader categories, as shown in figure 5.12. Two split points have been chosen to separate three broad classes of buildings: 1949 as a proxy for the beginning of a prolific period of residential building constructions in post-World War II Germany and 1979, the year when energy efficiency performance indicators have been introduced for the first time in the building standard codes.

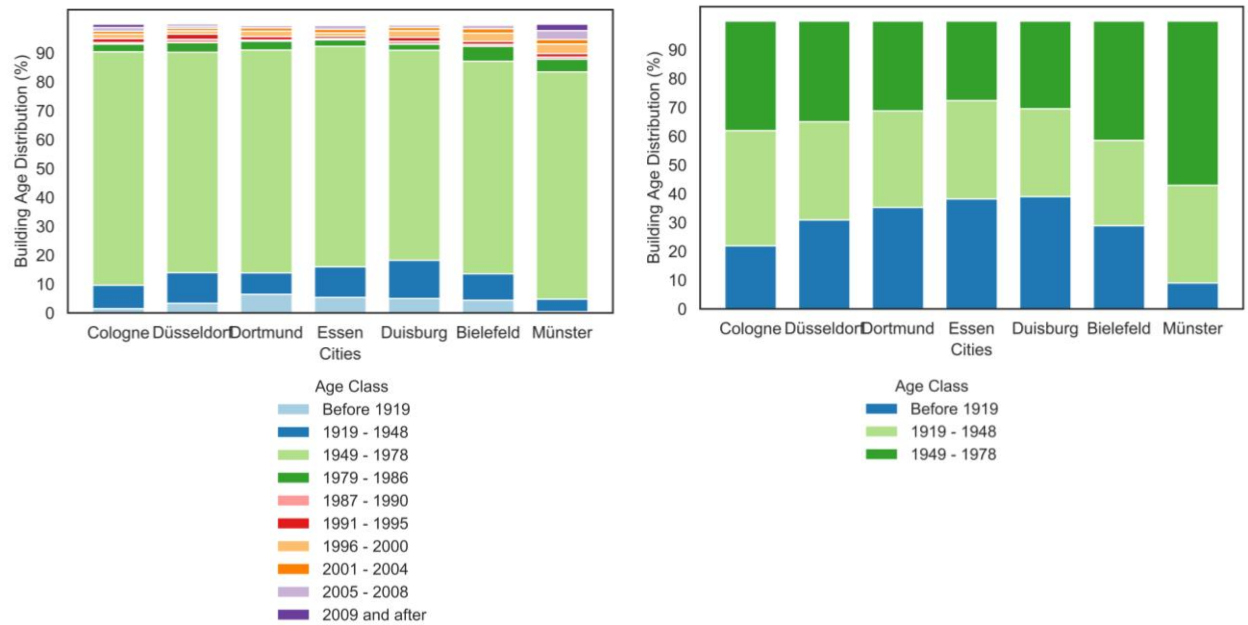


Figure 5.12: Distribution of age classes in the training dataset per city: on the left, distribution for the original ten-class categorization; on the right, the distribution for three-class categorization.

The power of prediction for the new age class has been tested for both city (figure 5.13) and region models (figure 5.15). Results at city level indicate an increase in accuracy of about 15% compared with results for ten-age classification, when normal sampling technique is used. For spatial sampling, the accuracy is doubled compared with ten-age classification, reaching an average 60% independent on the type of classification model used.

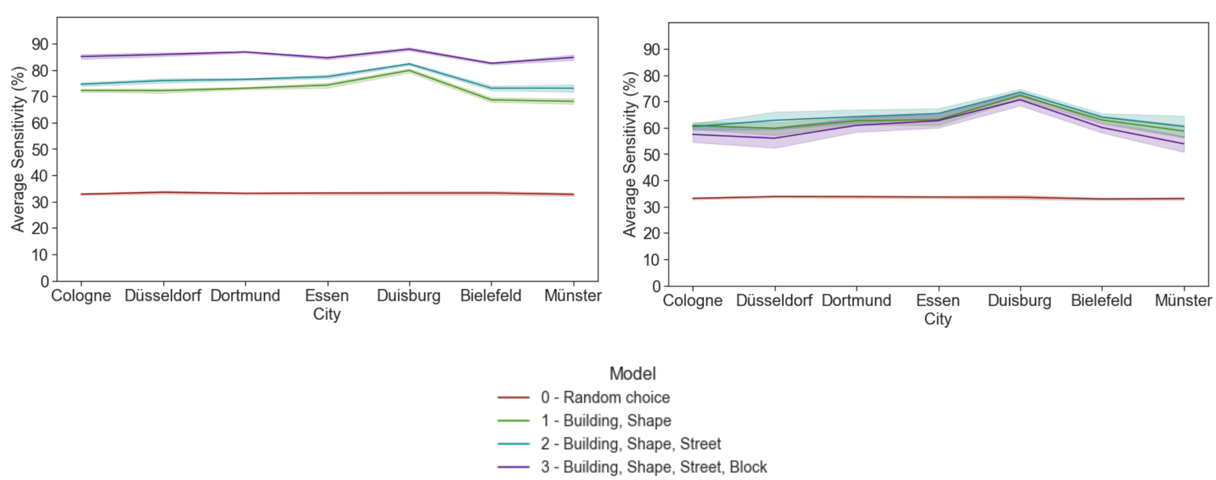


Figure 5.13: Average sensitivity scores for building age classification for the three-age category per city. On the left, normal stratified sampling is used for selecting training data. On the right, spatial sampling. Confidence intervals are computed based on 10 test results.

A closer inspection of the confusion between actual and predicted age classes (figure 5.14) leads to three observations. Firstly, the *"1949-1978"* class is consistently predicted with over 90% accuracy. Old buildings *"before 1948"* are seldomly labeled as new buildings *"1979 and after"*, while new buildings are more likely to be labeled as old buildings, for example, with a misclassification rate of 10% in several cities.

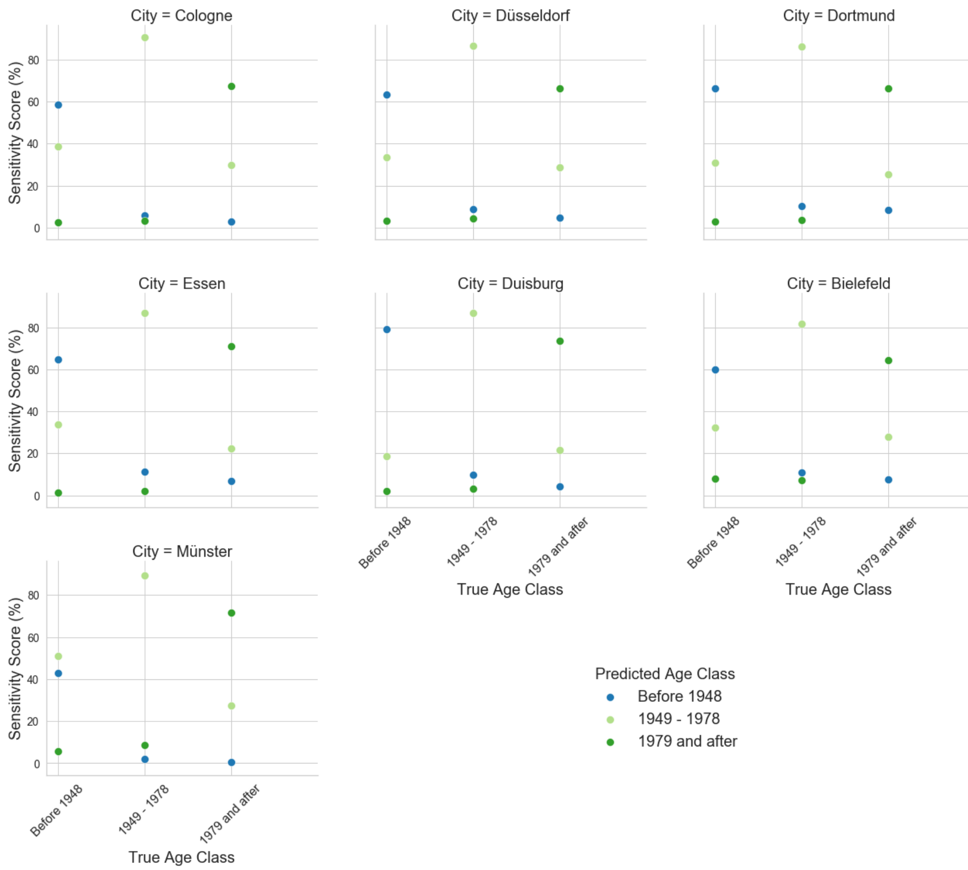


Figure 5.14: Distribution of predicted classification labels for buildings in a three-age category per city. Normal stratified sampling is used for selecting training data for classification with *Model 1*.

As for the city models, for region models results for three-age classification improve significantly compared with the former classification. The sensitivity increases from an average of 25% to 60 %. While the differences between models are reduced, the model that includes block features still produces the weakest results.

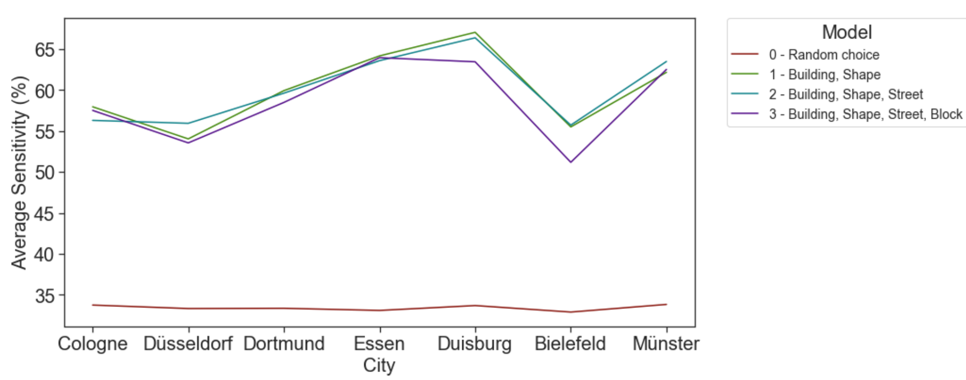


Figure 5.15: Average sensitivity scores for building age classification for three-age categorization per region.

It is interesting to note that compared with city models, the misclassification of both old buildings *"before 1948"* and new buildings *"1979 and after"* as belonging to the *"1949-1978"* age period is more likely to occur for region models. This trend is illustrated in figure 5.16. This is an indication that while the overall accuracy of classification across spatial scales increases for a three-age categorization, the applicability of model results for energy demand estimation is not necessarily improved.

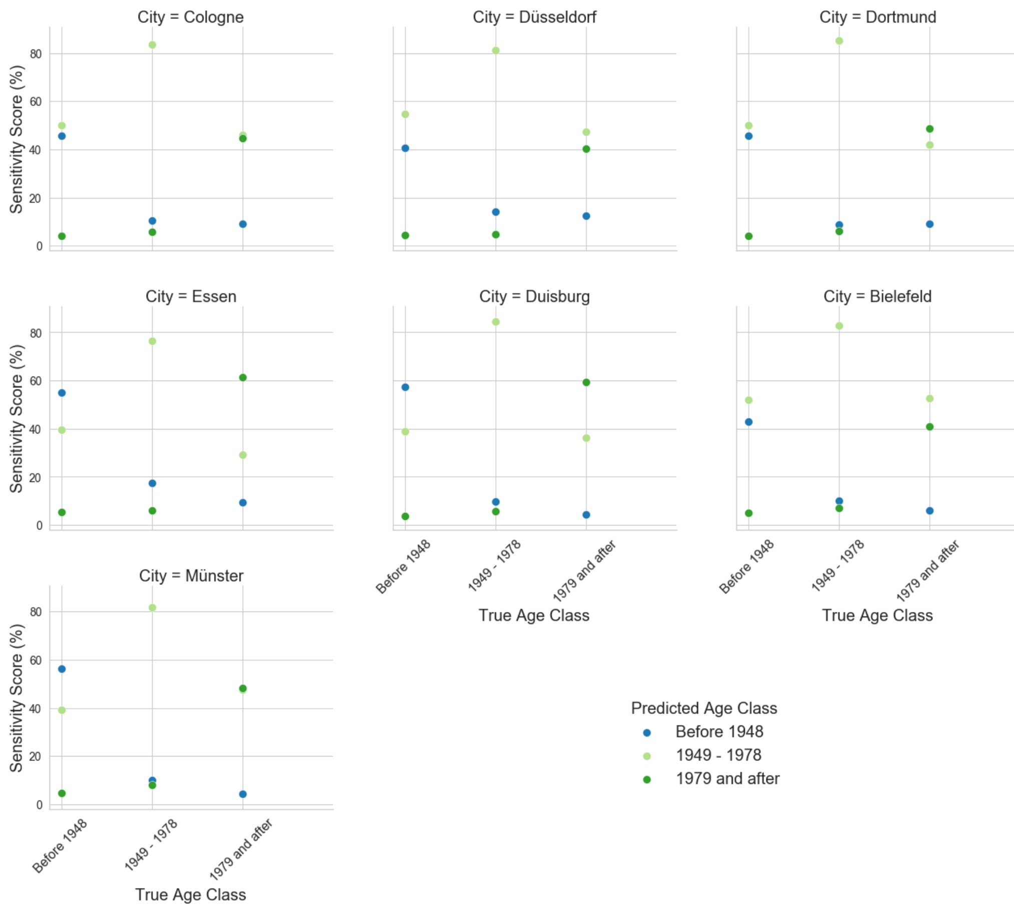


Figure 5.16: Distribution of predicted classification labels for buildings in every age category per region. Classification with *Model 1*.

6. DISCUSSION

The current chapter discusses the results obtained and their implications for model-based decision making and concludes with an overview of limitations and future directions of research. The structure of the first two sections follows the direction imposed by the research sub-questions we aimed to answer in this study.

6.1. BUILDING AGE PREDICTION

What is the influence of construction year on building energy efficiency?

The relationship between building age and energy efficiency has been discussed in detail in the chapter of literature review. To sum up, the age of the building is an indicator of the level of thermal insulation of the building and of the various heating and ventilation systems installed. These building characteristics, coupled with climate and inhabitant behavioral patterns define the energy consumption for building use and are important levees in the design of measure to improve building energy efficiency.

What model parametrization is most suitable for the automatic classification of building age?

Results indicate that model parameters are not as important as sampling design for the specifics of the modeling technique and of the classification problem investigated. A ten-class learning problem is a challenging classification task and it enhances the risk of class imbalance and distorted classification results. In order to curb the algorithm's tendency to over identify the majority class, a method of synthetically multiplying the observations from minority classes has been implemented. The result has been an improvement of 10% in sensitivity scores.

Another important aspect of sampling design highlighted by the research has been the importance of the spatial relationship between training and test samples. Accuracy is significantly improved when buildings for training and testing are in close vicinity. The reason behind this result is the fact the model takes advantage in this manner of the underlying spatial autocorrelation of features of buildings close to each other. When buildings for training and test are farther apart spatially, the autocorrelation component is reduced and the performance of the model deteriorates with more than 30% for the average accuracy over all classes.

What features are relevant for the classification of building age and what is the prediction success for different groups of features?

The most important type of features in terms of accuracy of classification within one city is building features that are spatially auto correlated. Firstly, block features rank the highest since all buildings from the same block share the same block attributes. Secondly, buildings alongside streets and in the corners of intersections will share the same street attributes, which makes these features the second most successful for classification. Lastly, both shape and general building features have been proven to have a spatial autocorrelation component although not as strong as block and street attributes. For the two first types of features, the accuracy can be further on improved when performing feature selection. It has been shown, for example that a small model of 6 features, which includes 4 street features related to connectivity and 2 complex shape features obtains best classification scores for all cities under analysis.

The conclusion stated above holds however only for a specific sampling design, which allows for buildings for training and testing being in close vicinity. Whenever this condition is not met, as for example in spatial sampling within a single city or in region-wide classification, the accuracy of classification decreases and the significance of groups of features is reversed. Shape and general building features outrank in performance the other types of features, especially block features.

In conclusion, the relevance of building attributes for age classification must be judged in a specific context related to the spatial scale of the classification model. General building features related to building geometry are the classical attributes for building age prediction used in literature. Building height and roof steepness have already been identified in the literature as important features for age prediction for selected buildings in UK (Rosser et al., 2019) and Canada (Tooke et al., 2014). Our results for the selected German buildings confirm these findings. Another roof related attribute has been judged important in our study: roof height, along with general attributes such as wall area and volume. Incorporating into our research attributes derived from building type classification (Droin, 2019; Wurm et al., 2016) has allowed us to highlight other relevant features: attributes related to the complexity of the shape of the building footprint and attributes related to the area occupied by a building and its neighbors.

These building features are judged the most relevant across different types of sampling designs and spatial scales. Nevertheless the drop in accuracy precision from 56% at city level to 20% at region level is an indication that that the importance of these attributes is most likely to stem from the fact that buildings within neighborhoods tend to resemble each other as shape due to regulations in building codes and spatial planning, for example constraints in height or lot size. Conclusions on the height, roof style or shape resemblance between ten age classes of buildings over an entire city or region cannot be firmly drawn.

Street features have not been explored in previous studies, with the exception of the research of Alexander et al. (2009) where distance to road has been used. The street metrics presented in this report have not been previously used for building age prediction, but have been employed for identifying urban structural typologies. We consider the further investigation of street attributes a promising venue of research since these features have been judged useful for the spatial autocorrelation component and have also the potential to be significant for region wide classification.

The advantage of using block features lies first and foremost with a method that allows the spatial component in training and test, at city level. Unlike street features, these features have proven the least significant when classifying regions and they appear to create confusion in the classification process. A tendency to improve classification results has been observed in the case of buildings constructed before 1919 but results did not replicate for the other age classes. This leads to the observation that while buildings definitely differ between cities, urban blocks differ even more.

How accurate is the classification and what is the model's power of generalization across different spatial scales?

The general conclusion that can be drawn is that accuracy depends on the model and sampling method employed. Results for city classification with normal sampling and region classification can be consulted in table 6.1 and table 6.2 respectively.

Table 6.1: Evaluation of city classification models per building age class. Results are averaged over all tests for 7 cities, a total of 70 tests. Normal stratified sampling is used for selecting training data.

Building Age	**Model 1**		**Model 2**		**Model 3**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Average for all ages	56,3	7,6	64,8	6,1	83,4	3,2
Before 1919	35,5	15,7	43,3	13,9	74,3	11
1919 - 1948	59,1	9,7	62,9	9,5	82,1	5,8
1949 - 1978	90	2,9	94,7	2,5	97,4	1,1
1979 - 1986	54,4	6,1	61,1	5,9	81,2	5,5
1987 - 1990	45,8	17,8	59,2	15,3	81,1	11,6
1991 - 1995	47,8	13,2	55,3	14,8	77,5	9,6
1996 - 2000	58,8	12,5	65,8	10	82,7	6,6
2001 - 2004	63,5	14,8	71,2	12,6	86,1	7,3
2005 - 2008	57,4	15,8	67,6	12,7	86,1	7,3
2009 and after	50,7	20,8	67	15,7	85,3	11,2

The highest accuracy of prediction was obtained for buildings built between 1949 and 1978. This represents the largest age category in the building stock and has been identified with accuracy over 80% across all sampling designs and across all spatial scales. The age class that follows consistently behind in evaluation scores is the set of buildings built before 1919 and 1948. However, the accuracy suffers when classification is performed at region level. Results for the oldest buildings, constructed before 1919 are not as consistent: at a city level, they output the least relevant accuracy results, while at region level the results are better than for the other age classes (buildings built after 1978).

With a favorable sampling design classification sensitivity for cities ranges on average between 56% and 83%. The results are in line with other results of classification by age that estimate an accuracy of prediction of approximately 70% with highest accuracies for older buildings. The conclusions regarding the oldest buildings (35% accuracy when using general building features) differ from some studies. For example, Alexander et al. (2009) report a classification accuracy of 90% for buildings built before 1919 and a 77% average accuracy over all classes. Rosser et al. (2018) report a classification accuracy of 88% for buildings built before 1915. Dissimilarities in appearance between the German and British building stock built before World War 1 could account for these differences. The sampling method we employed could be another reason. Choosing Census grid cells with uniform ages could lead to a sample of buildings that resemble each other, irrespective of their age, which leads to greater confusion in classification between classes.

For reasons addressed in the previous sections, we consider the referenced results to be optimistic since they ignore the spatial autocorrelation component that influences accuracy of results. This conclusion is based on spatial sampling tests that have shown that classification scores drop to approximately 45% when the examples used for learning obey to spatial constraints. Optimistic results are likely to be obtained in practice but only when the training pool contains buildings in the close vicinity of the target building for age estimation. If learning examples are sampled from regions farer away, it should be expected that accuracy of classification to suffer significantly to the point of the whole process becoming unusable for energy modeling purposes.

Furthermore, it has been observed that the results obtained on individual cities do not scale well to a region level. The region in this context is the combined area of all cities in the analysis, not the actual geographical area of North Rhine-Westphalia. The high level of heterogeneity among building features makes it difficult to identify correctly the building age. This occurs despite the fact that all cities belong to the same region and we would expect a common trend in construction standards. We can extrapolate the results to conclude that applying the classification for regions in different states for example, or between East and West Germany will lead to equally inconclusive results. This result reinforces the intuition gained through the tests performed with spatial sampling.

Table 6.2: Evaluation of region classification models per building age class. Results are averaged over all tests for 7 cities, a total of 7 tests.

Building Age	**Model 1**		**Model 2**		**Model 3**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Average for all ages	20,1	1,6	18,7	1,5	14,7	1,2
Before 1919	17,3	3,3	14,7	4,4	12,7	6,3
1919 - 1948	40,4	7,2	38	8,5	34,3	7,1
1949 - 1978	85,5	2,2	89,6	2	95,4	1,2
1979 - 1986	15,8	2,8	11,2	2,2	0,7	1
1987 - 1990	1,3	1,4	0,3	0,4	0	0
1991 - 1995	6,1	2,4	4,8	3,7	0,2	0,3
1996 - 2000	10,4	5	8,8	3,8	0,9	1,2
2001 - 2004	9,9	5,6	8,3	6,6	1,1	1,3
2005 - 2008	10,3	4,8	9,2	3,4	1	0,9
2009 and after	4,2	1,6	1,9	3,5	0,9	1,2

6.2. MODEL-BASED DECISION MAKING

What is the effect of misclassification of building age on energy demand estimation?

The purpose of this study is to design a method that helps national and local policy makers reach optimal decisions for improving the energy efficiency of the building stock.

The implications for policy making are only relevant for classification predictions that are acceptably accurate. After having discussed the power of prediction of the chosen model, we focus solely on the context where models achieve good classification results. This scenario most favorable for achieving a high degree of accuracy of prediction is when learning and prediction are made with data from the same city and when training data is available throughout the city.

We first observe that buildings from before 1978 are very rarely mislabeled as being built after this year. This ensures that energy consumption and retrofit requirements are not under-estimated. The case when old buildings would be labelled as new buildings are more damaging for policy making since it pictures an overly-optimistic version of reality where decision making is blocked. In this alternative scenario the problem would not be identified. There is however a tendency to over-estimate energy consumption due to the misclassification of buildings built after 1978 as being built between 1949 and 1978. A finer division of this age class that spans three decades could pinpoint the exact source of the misclassification error.

In order to broadly test the impact of age prediction on estimates of building energy demand, we will perform a simple modeling exercise, based on reference values for building energy demands per age. Through a complex computation method based on building typologies, climate data and building geometry, the energy demand for space and water heating for each building typology is estimated in the Tabula project (Loga et al., 2012). The data concerning the primary energy demand both before and after a standard renovation process are considered the reference values in our scenario (figure 6.1).

Using the classification results obtained for all cities, our goal is to derive new energy estimates for each age class and compare them with the reference values. Before this, however, we must unify Tabula age classes and building types and the Census age classes. As it can be observed in figure 6.1, building energy demands varies per building type. This aspect is however outside the scope of our research, and our target is to provide an overall estimate per age class irrespective of the building type. For this, the values for the four building types will be averaged into a single value per age class. In what concerns the discrepancies between age categorization, the timespans are aligned by either averaging over several ages or assigning repeatedly the same value to Census age classes. The results of this process can be observed in table 6.3 which lists for each building age a reference value for primary energy demand for space and water heating, expressed in kilowatt-hours per square meter per annum.

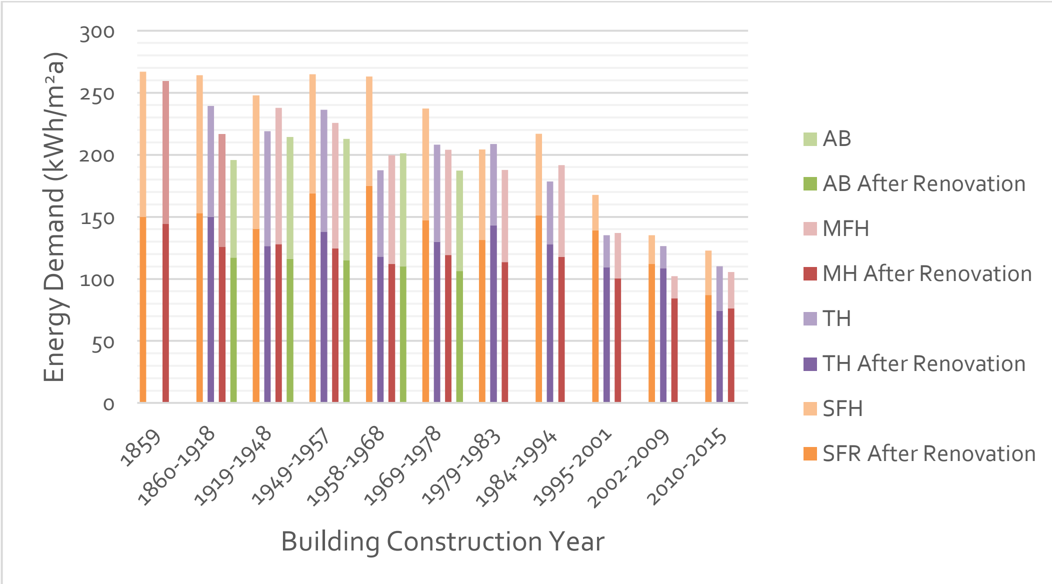


Figure 6.1: Total primary energy demand for heating and domestic hot water per building type (single family house; terraced house; multifamily house; apartment block) and building construction year. kWh/m2 a = kilowatt-hours per square meter per annum. Data retrieved from TABULA WebTool, http://webtool.building-typology.eu/#bm. Copyright Institut Wohnen und Umwelt GmbH 2012-2016 (Institut Wohnen und Umwelt , 2020).

Table 6.3: Reference values for energy demand per Census building age class.

Tabula Age Classification	*Census Age Classification*	*Building energy demand (kWh/m²a)*
1859 1860-1918	Before 1919	**246**
1919-1948	1919-1948	**229,8**
1949-1957 1958-1968 1969-1978	1949-1978	**219**
1979-1983	1979-1986	**200,3**
1984-1994	1987-1990	**195,7**
	1991-1995	**195,7**
1995-2001	1996-2000	**146,7**
2002-2009	2001-2004	**121,3**
	2005-2008	**121,3**
2010-2015	2009 and after	**112,9**

Based on the age-aligned reference values (table 6.3), the energy demand estimates per building age obtained after using the classification model are computed with the following formula:

$$E_{new}(age_k) = \sum_i E_{ref}(age_i) * p_{age_i}(age_k)$$

where $p_{age_i}(age_k)$ is the probability that age_k is labeled as age_i and $E_{ref}(age_i)$ is the reference energy demand for class age_i.

The results for energy demand before and after renovation can be consulted in figure 6.2 and figure 6.3 respectively. The confidence intervals are computed over all available test for all cities. The results for the weakest and strongest performing models are reported: *Model 1* , with 56,3 % accuracy of classification, and *Model 3*, with 83,4 %.

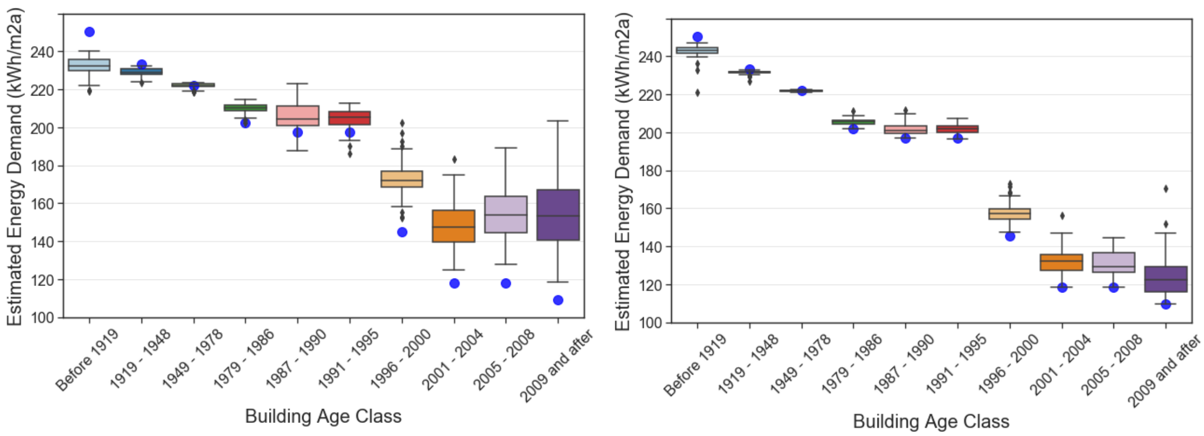


Figure 6.2: Total primary energy demand for heating and domestic hot water according to the age of building stock as estimated by the classification model. Blue dots represent the actual energy demand, as identified in the Tabula project. kWh/m^2 a = kilowatt-hours per square meter per annum.

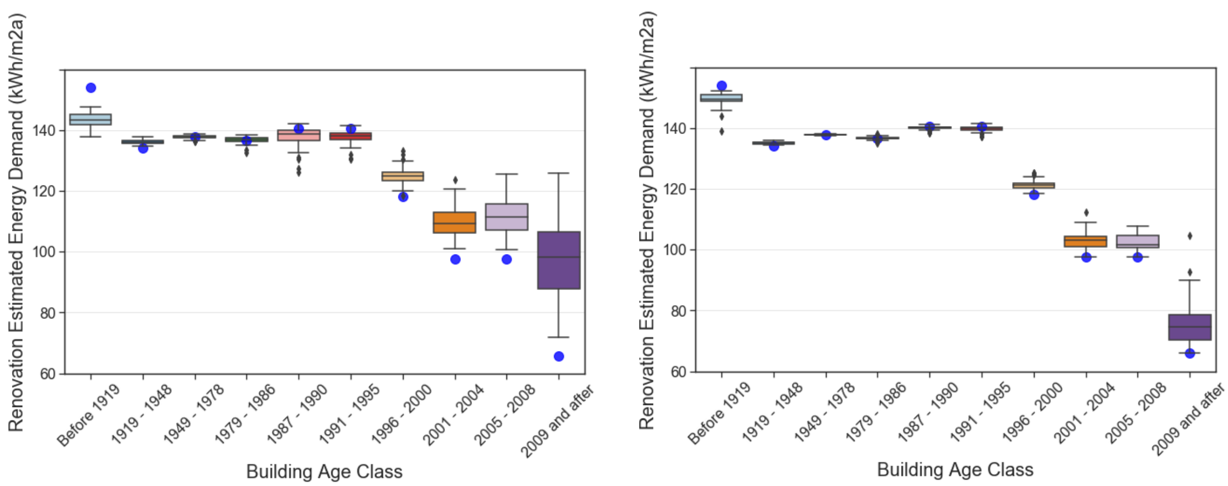


Figure 6.3: Total primary energy demand for heating and domestic hot water after building renovation according to the age of building stock as estimated by the classification model. Blue dots represent the actual energy demand after renovation, as identified in the Tabula project. kWh/m^2 a = kilowatt-hours per square meter per annum.

The largest variations in energy estimates occurs for the newest buildings, especially built after 2009. For estimates after renovation, both models show a marked improvement. This leads to the conclusion that if an individual decision is to be made concerning the impact of renovation for a building whose age has been classified with either one of the proposed classification models, the prognosis will be accurate for buildings between 1919 and 1995.

We extend the modeling exercise from a standard building to city wide residential building stock. The energy demand for a city is computed using the formula:

$$\sum_i E(age_i) * Total\ Floor\ Area(age_i)$$

where $E(age_i)$ is building energy demand (E_{ref} for reference values and E_{new} for estimated values) and $Total\ Floor\ Area(age_i)$ is the sum of floor areas over all buildings of age age_i.

The results in figure 6.4 demonstrate how the variations in energy prediction at individual building level are smoothed out when increasing the size of the energy model. The difference between reference and model estimated energy demand in this scenario is less than 1 % of the reference value. As illustrated in the individual building case, this difference will be even lower for after renovation energy estimates. The reason behind these positive results is the age distribution of the building stock and the energy demands associated with age classes. Firstly, for all cities approximately 50% of the stock (figure 4.9) is represented by buildings of an age class that is identified with over 90% accuracy *("1949-1978").* Secondly, old and new buildings are not generally confused by the classification model. Buildings built before 1978 are always represented as such. Since these buildings are considered to have worst energy performance than newer buildings, the model leads to acceptable accurate energy estimates, especially at larger spatial scales. In conclusion, the method performs very well on the identification of old buildings in the German cities analyzed, which give an indication to policy makers where to focus renovation efforts.

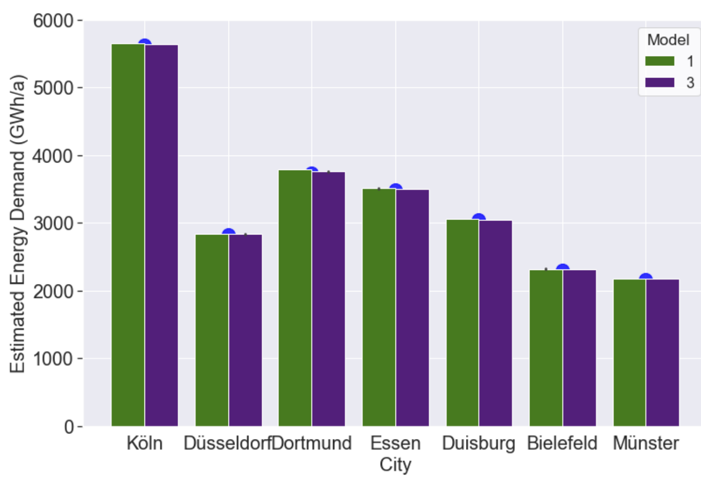


Figure 6.4: Total primary energy demand for heating and domestic hot water per city. Blue dots represent the actual energy demand. GWh/ a = gigawatt-hours per annum.

It is expected that the variations in energy demand estimates to be greater for models at a higher resolution, for example individual neighborhoods. The option in this scenario would be to use a model with high accuracy, for example a model that takes advantage of the spatial component like *Model 3*.

The need for individual building age information is dependent on the resolution of the energy simulation. Zirak et al. (2020) have used the Census information to conclude that heating demand for an entire city is acceptable, yielding a difference of 4.5% compared with municipality reference data. Zirak et al. (2020) also conclude that an important factor in energy modelling is the size of the heated space. The advantage of our proposed method over a technique that uses aggregates of age over neighborhoods, as for example the Census survey that provides 100 x 100 m grid cell building data, is that with a classification model, for a specific building we can associate the age with building geometry, and especially building footprints, and the extent of heated area is an essential parameter in estimated building energy demand. This level of detail allows for more precise energy estimates than using general building typologies or aggregated age data.

Another advantage in using this classification model is its power to identify the class of buildings in need of renovation. The aim of sustainable renovations measures is to improve the energy efficiency of the existing building stock by enhancing the performance of the thermal building envelope while at the same time reducing greenhouse gas emissions through the use of renewable energy sources. Renovations have different impact depending on the building age. Refurbishment of buildings built before 1919 was shown to have little impact on energy consumption and this is due to the structure of these buildings and the difficulties encountered into their renovation (Energy Efficiency For EU Historic Districts Sustainability, 2013). Renovations of buildings built between 1950 and 1990 have a significant impact on energy consumption as shown by the work of Michelsen and Müller-Michelsen (2010), cited by the EUFFESE report.

The usability of an age classification model for policy decisions concerning energy efficiency is only as good as the accuracy of the energy models build upon it. The discussion in the introductory chapters of the thesis has shown the general trend of the relationship between building age and energy consumption estimates. This relationship is not always straightforward. In their study Bigalke et al. (2012) as cited by EUFFESE report (Energy Efficiency For EU Historic Districts Sustainability, 2013), have shown that the measured energy consumption for buildings is on average 30% lower than calculated estimates, due to differences in user behavior and uncertainty in U-value estimates. This is truer for buildings built before 1949. This findings in supported by other studies and has been defined as the *pre-bound effect*, the gap between performance and actual energy consumption (Minna Sunikka-Blank & Ray Galvin, 2012). While an age prediction model cannot alleviate this gap in modeling results, it can be a useful tool for performing uncertainty analysis that quantifies the exact lag between estimates and actual consumption.

6.3. LIMITATIONS AND FUTURE WORK

There are two types of limitations that are worth mentioning. The first type consists in technical limitations and the second type in limitations in the usability of the model for energy-modeling purposes.

6.3.1. TECHNICAL IMPROVEMENTS

From the first category, we consider the major limitation of the present study to be the data acquisition method. Due to the scarcity of publicly available data on construction year a sampling method based on information extraction from spatially aggregated data has been designed. The effect of the data acquisition method on classification accuracy has not been analyzed. It is possible that by selecting from the Census data grid cells of 100 m x 100 m that only contain buildings of the same age we have introduced a sampling bias. It would be noteworthy to extend this analysis to the entire building stock of the cities under analysis and verify the reproducibility of results.

Random stratified sampling has been proven to lead to more optimistic classification results than spatial sampling. This aspect has not been discussed in most of the literature concerning building age prediction and should be further on explored. Research in this direction could include three aspects. Firstly, the extent of spatial autocorrelation per feature can be estimated by means of spatial indices which will lead to identifying the building features most likely to correlate spatially. Secondly, to counteract this tendency to auto-correlate, different spatial sampling methods could be designed. We have suggested a simple split of buildings by blocks, but other strategies that take into account the morphology of the urban region should be considered. Lastly, a different learning method could be employed, for example an algorithm that accounts for spatial autocorrelation.

The chosen classification model, Random Forest, is a powerful, robust and widely-used supervised learning method. It has however its drawbacks and judging on the particularities of the learning problem at hand – imbalanced multiclass classification of spatial data – further developments should include the analysis of performance of other learning models, for example Deep Learning. Another improvement that could be brought to the classification model is the feature selection procedure. Backwards feature elimination has been shown to produce local and not global optima for classification accuracy.

Another direction of research concerns building features most likely to predict building age. The association between buildings and the properties of nearby streets properties has been given little weight in previous studies on building age prediction. Our results show that these features deserve a closer investigation, along with some complex shape features. Building attributes extracted from the properties of the block that surrounds them have proven to be weak predictors of building age at larger spatial scales. The reason would be that labelling multiple age classes with the same attribute creates confusion in the classification process. One avenue of research would be to test a classification of construction year per block where the majority age of buildings in the block would be the block class label.

6.3.2. IMPROVEMENTS FOR ENERGY MODELING

The first limitation for policy analysis support is derived from the data acquisition method and concerns the fact that the available age classification in Census is not the standard energy-based typology of buildings for Germany. The typology, as defined in the Tabula project, comprises of 11 classes. The major difference consists in the segmentation of the Census age class 1949 – 1978 into several age classes in the Tabula project. For further developments a different age split could be considered, provided that the data is available or can be accurately deduced. We consider this to be of lesser importance, since there are ways to reclassify the available data. The important question concerns the majority class and the power of the model to identify between the specifics of each decade of construction comprised in this class: 1949 – 1978.

Last but not least, estimation of the impact of the resulting age classification has been done through literature study by analyzing the associations found in literature between building age and energy consumption and through a simple energy modeling exercise. Further work should include a more complex energy modeling component where energy estimates are compared with reference data. This step is essential in quantifying the uncertainty that the model introduces in the decision-making process. Lomas and Wright (2010) have shown the influence of input parameters on model results varies according to different building typologies (age and size characterization). In their study on CO_2 emissions, the influence of model inputs was significantly higher for detached and old buildings, than the rest of the building stock (Firth et al., 2016).

In the absence of concrete data on building refurbishment, energy estimates can be misleading. The age information on itself is not sufficient for estimating energy consumption. Moreover, the EFFESUS report remarks that visual inspection of a building is also not a sufficient indicator for estimating the age or energy consumption. The laws that protect the façade appearances of historical buildings do not extend to the entire building, and energy efficiency renovations are likely to have been performed even without visible external indications (Energy Efficiency For EU Historic Districts Sustainability, 2013). Our proposed method does not cover in any way the renovation status of a building.

Last but not least, the questions of data privacy and ethical considerations concerning machine learning use arise. The danger of such a modeling technique is that an open source model is proposed to precisely identify information that has been judged as private by national laws. While the discussion extends the scope of our research, we consider it a worth-while endeavor for a deeper analysis in the context of elaborating policy regulations for the responsible-use of machine learning models and predictions.

7. CONCLUSION

The investigation pursued in this research project has shown that using supervised classification for predicting building age has a positive impact on the accuracy and level of detail of energy efficiency modeling. This potential is best explored in situations where there is available data for training from buildings in the same spatial region, and more specifically the same city, as the buildings whose ages are to be predicted. Accuracy of classification is optimal when the age of buildings in the vicinity of the target dataset is known. These results do not reproduce when training a model on a dataset from a different city or group of cities. This is a standard example when more training data does not necessarily mean better prediction results. These results must however be interpreted in the context in which they could be a result of the historical evolution of building construction methods and building standards in Germany. It can be expected that the heterogeneity in shape and location of the building stock per age period to vary from country to country. The segmentation of construction years into different age classes is in itself a country-specific process.

This lack of power of generalization does not however nullify the potential of using this method as a decision-making aide. The class of models where building age is an essential input parameter is bottom-up models and these are by their nature high-resolution localized models. Building an energy model for a neighborhood or a city is a common procedure in this practice, with results being scaled up at regional or national level. This neighborhood-by-neighborhood or city-by-city approach can benefit from the predictions of a classification model that uses only local data for training and prediction.

In this context of localized training, the accuracy of the method reaches up to 82% with best results obtained in identifying buildings more likely to be energy-inefficient, i.e. built before 1979. The ability to differentiate between old and new buildings is essential for accurate estimation of energy consumption and retrofit needs. The model has a tendency to over-estimate the number of old buildings and leads to slightly pessimistic results concerning energy consumption and retrofit requirements for new buildings. Nevertheless, for energy models at the scale of a city, it has been proved that differences compared with reference values are minimal due to the high proportion of older buildings in the building stock.

The study has gone further than other studies by applying the analysis to a larger scale than one city or a set of neighborhoods in a city. The method employed can be easily replicated for the analysis of the entire state, and also scaled for national analysis, provided the challenges of data acquisition and data compatibility are overcome.

Concerning the relevance of building features for age classification, an in-depth analysis has been done to illustrate the differences in feature importance across spatial scales. Classical building features related to building geometry have shown to be relevant across all spatial scales, and their relevance has been decomposed on two directions: general similarity between buildings over regions and local similarity between buildings in close vicinity one of another. In the later scenario, building features are significantly more conclusive for classification. The study has also brought forward new types of features in the context of building age prediction: building footprint features, street and block features. Out of them, selected shape and street features have been deemed important and deserving of further analysis.

In conclusion, the scientific significance of this research work is three-fold. Firstly, the study is a first attempt to analyze the potential of predicting building age from known data both within the same city and from different cities. Secondly, it illustrates the effect of spatial autocorrelation on producing optimistic classification results. Lastly, it highlights new categories of features that have proven to be relevant for prediction and discusses their relevance at different spatial scales. Furthermore, the model was proven to produce energy estimates for energy models at city scale that are within 1% of the reference energy demand, making it a useful tool for supporting energy efficiency modeling.

References

Aksoezen, M., Daniel, M., Hassler, U., & Kohler, N. (2014). Building age as an indicator for energy consumption. *Energy and Buildings*, *87*, 74-86.doi: 10.1016/j.enbuild.2014.10.074

Alexander, D., Lannon, S., & Linovski, O. (2009). The Identification of analysis of regional building stock characteristics using map based data. *Building Simulation.* Glasgow: Eleventh International IBPSA Conference. Retrieved from http://www.ibpsa.org/proceedings/BS2009/BS09_1421_1428.pdf

Alhamwi, A., Medjroubi, W., Vogt, T., & Agert, C. (2017). OpenStreetMap data in modelling the urban energy infrastructure: A first assessment and analysis. *Energy Procedia*, *142*, 1968-1976. doi:10.1016/j.egypro.2017.12.397

Angel, S., Parent, J., & Civco, D. L. (2010). Ten compactness properties of circles: measuring shape in geography. *The Canadian Geographer*, *54*(4), 441-461. doi:10.1111/j.1541-0064.2009.00304.x

Armesto Gonzalez, J., Docampoa, M. L., & Canas Guerrero, I. (2006). The application of new technologies in construction: Inventory and characterisation of rural constructions using the Ikonos satellite image. *Building and Environment*, *41*, 174-183.doi: 10.1016/j.buildenv.2005.01.017

Belgiu, M., & Dragut, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31. doi:10.1016/j.isprsjprs.2016.01.011

Berghauser Pont, M., Stavroulaki, G., Bobkova, E., Gil, J., Marcus, L., Olsson, J., . . . Legeby, A. (2019). The spatial distribution and frequency of street, plot and building types across five European cities. *Environment and Planning B: Urban Analytics and City Science*, *46*(7). doi:10.1177/2399808319857450

Bigalke, U., Discher, H., Lukas, H., Zeng, Y., Bensmann, K., & Stolte, C. (2012). *Der dena-Gebäudereport. Statistiken und Analysen zur Energieeffizienz im Gebäudebestand.* Ed. Deutsche Energie-Agentur.

Biljecki, F., & Sindram, M. (2017). Estimating building age with 3D GIS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4.* Melbourne. doi:10.5194/isprs-annals-IV-4-W5-17-2017

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *MC Bioinformatics*.

BMVBS / BBSR. (2009). *Benchmarks für die Energieeffizienz von Nichtwohn-gebäuden.* Federal Ministry of Transport, Building and Urban Development. BBSR-Online-Publikation.

BMVBS. (2011). *Typologie und Bestand beheizter Nichtwohngebäude in Deutschland.* Federal Ministry of Transport, Building and Urban Development. BMVBS-Online-Publikation.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126-139. doi:10.1016/j.compenvurbsys.2017.05.004

Braun, I. (December 2015). *The Architectural Legacy of the 1960s*. Retrieved on 16. October 2019 from Goethe Institut: https://www.goethe.de/en/kul/arc/20668421.html

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.doi: 10.1613/jair.953

Chen, C., Liaw, A., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. Retrieved from: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

CityGML. (2019). Retrieved from Open Geo Spatial: https://www.opengeospatial.org/standards/citygml

Costanzo, A., Montuori, A., Silva, J. P., Silvestri, M., Musacchio, M., Doumaz, F., . . . Buongiorno, M. F. (2016). The Combined use of airborne remote sensing techniques within a GIS environment for the seismic vulnerability assessment of urban areas: an operational application. *Remote Sensing, 8*(146). doi:10.3390/rs8020146

Creutzig, F., Baiocchi, G., Bierkandt, R., Pichler, P. -P., & Seto, K. C. (2015). Global typology of urban energy use and potentials for an urbanization mitigation wedge. *Proceedings of the National Academy of Science, 112*, S. 6283–6288. doi:10.1073/pnas.1315545112

Das Plattenportal. (2019). Retrieved on 16. October 2019 from Plattenbau der DDR-Architektur: https://www.jeder-qm-du.de/ueber-die-platte/plattenbau-typen/

Data tool. (2019). Retrived on 18. February 2020 from ENTRANZE - Policies to ENforce the TRAnsition to Nearly Zero Energy buildings in EU-27: https://www.entranze.eu/

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., ..., Hartig, F. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40(*8*)*.913-929. doi: 10.1111/ecog.02881

Decision tree learning. (2020). Retrieved on 21. February 2020 from Wikipedia: https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., . . . Ohlemüller, R. (October 2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography, 30*(5), 609-628. doi: 10.1111/j.2007.0906-7590.05171.x

Droin, A. (2019). *Semantic labelling of building types: A comparison of two approaches using Random Forest and Deep Learning.* Graz: German Aerospace Center.

Economidou, M. (March 2012). Energy performance requirements for buildings in Europe . *REHVA*, 16-21. Retrieved from https://www.researchgate.net/publication/296505832_Energy_performance_requirements_for_buildings_in_Europe

Economidou, M., Atanasiu, B., Staniaszek, D., Maio, J., Nolte, I., Rapf, O., . . . Zinetti, S. (2011). *Europe's buildings under the microscope. A country-by-country review of the energy performance of buildings.* Buildings Performance Institute Europe (BPIE).

Energy Efficiency For EU Historic Districts Sustainability. (2013). *European building and urban stock data collection.* From https://www.effesus.eu/wp-content/uploads/2016/01/D-1.1_European-building-and-urban-stock-data-collection.pdf abgerufen

European Commission. (26. August 2019). *Energy performance of buildings*. Retrieved on 22. October 2019 from European Commission: https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings/overview

European Data Portal. (2017). *Open.NRW Hackathon*. Retrieved on 3. February 2020 von European Data Portal: https://www.europeandataportal.eu/en/highlights/opennrw-hackathon

Filippidou, F., & Jimenez Navarro, J. (2019). *Achieving the cost-effective energy transformation of Europe’s buildings.* Luxembourg: Publications Office of the European Union. doi:10.2760/278207

Firth, S. K., Lomas, K., & Wright, A. (2010). Targeting household energy efficiency measures using sensitivity analysis. *Building Research & Information*, *38*(1), 25-41. doi:10.1080/09613210903236706

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly.

Gil, J., Beirão, J. N., Montenegro, N., & Duarte, J. P. (2012). On the discovery of urban typologies: data mining the many dimensions of urban form. *Urban Morphology, 16*(1), 27-40. Retrieved from https://www.researchgate.net/publication/256895610_On_the_discovery_of_urban_typologies_Data_mining_the_many_dimensions_of_urban_form

Global Energy Assessment. (2012). *Global Energy Assessment - Toward a Sustainable Future.* Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria: Cambridge University Press. Retrieved from https://iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/Global_Energy_Assessment_FullReport.pdf

Grothe, S. (10. September 2010). *A Massive facelift for Eastern Germany*. Retrieved on 24. October 2019 from Spiegel: https://www.spiegel.de/international/germany/reunification-renovations-a-massive-facelift-for-eastern-germany-a-720326.html

Hermosilla, T., Palomar-Vázquez, J., Balaguer-Beser, Á., Balsa-Barreiro, J., & Ruiz, L. A. (March 2014). Using street based metrics to characterize urban typologies. *Computers, Environment and Urban Systems*, *44*, 68-79. doi:10.1016/j.compenvurbsys.2013.12.002

Institut Wohnen und Umwelt . (23. 1 2020). *TABULA Web Tool*. Von Building Typology: http://webtool.building-typology.eu/#bm abgerufen

International Energy Agency. (2016). *International energy agency world energy outlook.* Retrieved from https://www.iea.org/reports/world-energy-outlook-2016

Jensen, J., & Cowen, C. (May 1999). Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing*, *65*(5), 611-622. doi: 10.1002/9780470979587.ch22

Jilge, M., Heiden, U., Habermeyer, M., Mende, A., & Juergens, C. (2017). Detecting Unknown Artificial Urban Surface Materials Based on Spectral Dissimilarity Analysis. *Sensors*. doi:10.3390/s17081826

Kadaster. (15. October 2019). Retrieved on 24. October 2019 from Open Data Kadaster: https://zakelijk.kadaster.nl/open-data

Karatasou, S., Laskari, M., & Santamouris, M. (1. July 2018). Determinants of high electricity use and high energy consumption for space and water heating in European social housing: Socio-demographic and building characteristics. *Energy and Buildings*, 107-114. doi:10.1016/j.enbuild.2018.04.019

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence. 2*, S. 1137–1143. ACM. doi:10.5555/1643031.1643047

Krehl, A., Siedentop, S., Taubenböck, H., & Wurm, M. (2016). A Comprehensive View on Urban Spatial Structure:Urban Density Patterns of German City Regions. *Internation Journal of Geo-Information, 5*. doi:10.3390/ijgi5060076

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1-5. Retrieved from http://jmlr.org/papers/v18/16-365.html

Lemmens, M. (2011). *Geo-information: Technologies, Applications and the Environment.* Springer.

Li, Y., Chen, Y., Rajabifard, A., Khoshelham, K., & Aleksandrov, M. (2018). Estimating building age from Google Street View images using Deep Learning. In S. Winter, A. Griffin, & M. Sester (Hrsg.), *10th International Conference on Geographic Information Science. 114*, S. 1-7. Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik. doi:https://10.4230/LIPIcs.GISCIENCE.2018.40

Loga, T., Diefenbach, N., Stein, B., & Born, R. (2012). *Tabula: Further development of the German residential building typology.* . Darmstadt: Institut Wohnen und Umwelt GmbH. Von https://www.building-typology.eu/downloads/public/docs/scientific/DE_TABULA_ScientificReport_IWU.pdf abgerufen

Lowry, J. H., & Lowry, M. B. (2014). Comparing spatial metrics that quantify urban form. *Computers, Environment and Urban Systems, 44*, 59-67. doi:10.1016/j.compenvurbsys.2013.11.005

Mangold, M., Österbring, M., Wallbaum, H., Thuvander, L., & Femenias, P. (2016). Socio-economic impact of renovation and energy retrofitting of the Gothenburg building stock. *Energy and Buildings, 123*, 41-49. doi:10.1016/j.enbuild.2016.04.033

Marcotullio, P. J., Sarzynski, A., Sperling, J., Chavez, A., Estiri, H., Pathak, M., & Zimmerman, R. (2018). Energy transformation in cities. In C. Rosenzweig, W. Solecki, P. Romero-Lankao, S. Mehrotra, S. Dhakal, & S. Ali Ibrahim (Hrsg.), *Climate Change and Cities: Second Assessment Report of the Urban Climate Change Research Network* (S. 443–490). New York: Cambridge University Press.

Marshall, S. H. (2005). Urban pattern specification. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.5605&rep=rep1&type=pdf

Mata, E., Kalagasidis, A. S., & Johnsson, F. (2014). Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK. *Building and Environment, 81*, 270-282. doi:10.1016/j.buildenv.2014.06.013

McGarigal, K., & Marks, B. J. (1995). *Fragstats: spatial pattern analysis program for quantifying landscape structure.* (F. S. Department of Agriculture, Hrsg.) Portland, OR, U.S: Gen. Tech. Rep. PNW-GTR-351.

Michelsen, C., & Müller-Michelsen, S. (2010). Energieeffizienz im Altbau: Werden die Sanierungspotenziale überschätzt? Ergebnisse auf Grundlage des ista-IWH-Energieeffizienzindex. *Wirtschaft im Wandel, 16*(9), 447-455. From https://www.iwh-halle.de/fileadmin/user_upload/publications/wirtschaft_im_wandel/9-10-5.pdf abgerufen

Minna Sunikka-Blank & Ray Galvin (2012): Introducing the prebound effect: the gap between

performance and actual energy consumption, Building Research & Information, 40:3, 260-273

Mueller, M., Segl, K., Heiden, U., & Kaufmann, H. (2006). Potential of High-Resolution Satellite Data in the Context of Vulnerability of Buildings. *Natural Hazards, 38*, 247-258. doi:10.1007/s11069-005-8637-x

Nahlik, M. J., Chester, M. V., Pincetl, S. S., Eisenman, D., Sivaraman, D., & English, P. (2017). Building Thermal Performance, Extreme Heat, and Climate Change. *Journal of Infrastructure Systems, 23*(3). doi:10.1061/(ASCE)IS.1943-555X.0000349

Nilsson, N. J. (1996). *Introduction to Machine Learning.* Stanford , CA. Retrieved from https://ai.stanford.edu/~nilsson/mlbook.html

Normalized difference vegetation index. (2019). Retrieved on 7. January 2020 from Wikipedia: https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index

Nosratabadi, S., Mosavi, A., Keivani, R., Ardabil, S., & Aram, F. (2019). State of the Art Survey of Deep Learning and Machine Learning Models for Smart Cities and Urban Sustainability. In V.-K. A. (Hrsg.), *Engineering for Sustainable Future* (S. 228-238). Springer.doi: 10.1007/978-3-030-36841-8_22

Oikonomou, M. (2014). The transformation of the urban block in the European City. *21st International Seminar on Urban Form "Our common future in urban morphology", 2.* Porto. Retrieved from https://www.researchgate.net/publication/309782868_The_transformation_of_the_urban_block_in_the_European_City

Open Data. (2019). Retrieved on 15. October 2019 from Bundesamt für Kartographie und Geodäsie: https://gdz.bkg.bund.de/index.php/default/open-data.html

Open Government Germany. (2019). *Second National Action Plan (NAP) 2019 – 2021 in the Framework of Germany's Participation in the Open Government Partnership (OGP).* Berlin: Federal Chancellery. Retrieved from https://www.opengovpartnership.org/wp-content/uploads/2019/09/Germany_Action-Plan_2019-2021_EN.pdf

Open NRW. (2019). Retrieved from https://open.nrw/

Ortlepp, R., Gruhler, K., & Schiller, G. (2018). Materials in Germany's domestic building stock: calculation model and uncertainties. *Building Research & Information, 46*(2), 164-178. doi:10.1080/09613218.2016.1264121

Pahl-Weber, E., & Henckel, D. (2008). The Planning system and planning terms in Germany: A Glossary. *Studies in Spatial Development*. Retrieved from https://www.econstor.eu/bitstream/10419/60979/1/719731003.pdf

Pedregos, F. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, 2825-2830. Retrieved from http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the Prediction Performance of Spatial Models via Spatial k-Fold Cross Validation. *International Journal of Geographical Information Science*, 1-18.doi: 10.1080/13658816.2017.1346255

Reinhart, C. F., & Cerezo Davila, C. (2016). Urban building energy modeling: A review of a nascent field. *Building and Environment*, *97*, 196-202. doi:10.1016/j.buildenv.2015.12.001

Reuben, J. (2017). *In Pre-WWII Berlin, the Shape of Your Roof Was a Highly Political Decision*. Retrieved on 14. 10 2019 from Atlas Obscura: https://www.atlasobscura.com/articles/the-heated-highly-political-roof-war-that-captivated-berlin-before-world-war-ii

Roehrkasten, S., & Steinbacher, K. (2016). *Germany: Promoting an Energiewende Domestically and Globally.*

Rosser, J. F., Long, G., Zakhary, S., Boyd, D. S., & Mao, Y. R. (2019). Modelling Urban Housing Stocks for Building Energy Simulation Using CityGML EnergyADE. *International Journal of Geo-Information, 8*(163). doi:10.3390/ijgi8040163

Rosser, J., Boyd, D., Long, G., Zakhary, S., Mao, Y., & Robinson, D. (2019). Predicting residential building age from map data. *Computers, Environment and Urban Systems, 73*, 56-67. doi:10.1016/j.compenvurbsys.2018.08.004

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches. *IEEE Computational Intelligence Magazine, 13*(4), 59-76. doi: 10.1109/MCI.2018.2866730

Schettler-Köhler, H.-P., & Ahlke, I. (2016). *Implementation of the EPBD in Germany.* Retrieved from https://epbd-ca.eu/ca-outcomes/outcomes-2015-2018/book-2018/countries/germany

Sentinel-2 Resolution and Swath. (2019). Retrieved on 24. October 2019 from European Space Agency: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/instrument-payload/resolution-and-swath

Seyedzadeh, S., Rahimia, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering, 6*(5). doi:10.1186/s40327-018-0064-7

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews, 13*, 1819–1835.doi: 10.1016/j.rser.2008.09.033

Tooke, T. R., Coops, N. C., & Webster, J. (2014). Predicting building ages from LiDAR data with random forests for building energy modeling. *Energy and Buldings, 68*, 603-610. doi:10.1016/j.enbuild.2013.10.004

UNECE. (2018). *Mapping of existing energy efficiency standards and technologies in buildings in the UNECE region.* Geneva: UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE). Retrieved on 10. February 2020 from https://www.unece.org/fileadmin/DAM/hlm/Meetings/2018/09_05-07_St._Petersburg/EE_Standards_in_Buildings_full_version.ENG.pdf

United Nations. (2019). *The Sustainable Development Goals Report.* New York. Retrieved from https://unstats.un.org/sdgs/report/2019/The-Sustainable-Development-Goals-Report-2019.pdf

United Nations Economic Commission for Europe . (2018). *Mapping of existing energy efficiency standards and technologies in buildings in the UNECE region.* Geneva: United Nations Economic Commission for Europe . Retrieved on 10. February 2020 from https://www.unece.org/fileadmin/DAM/hlm/Meetings/2018/09_05-07_St._Petersburg/EE_Standards_in_Buildings_full_version.ENG.pdf

U-values. (3. August 2019). Retrieved on 23. October 2019 from Designing Buildings: https://www.designingbuildings.co.uk/wiki/U-values

Witten, I. H., Eibe, F., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier.

Wurm, M., Schmitt, A., & Taubenböck, H. (May 2016). Building types' classification using shape-based features and linear discriminant functions. *IEEE JOURNAL OF SELECTED TOPICS INIEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(5), 1901-1912. doi:10.1109/JSTARS.2015.2465131

Zensus. (15. October 2011). Retrieved on 17. 10 2019 from Zensus: https://www.zensus2011.de/EN/Home/home_node.html

Zeppelzauer, M., Despotovic, M., Sakeena, M., & Koch, D. (2018). Automatic Prediction of Building Age from Photographs. *ICMR '18: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, (S. 126-134). doi:10.1145/3206025.3206060

Zirak, M., Weiler, V., Hein, M., & Eicker, U. (2020). Urban models enrichment for energy applications: Challenges in energy simulation using different data sources for building age information. *Energy, 190*. doi: 10.1016/j.energy.2019.116292

APPENDICES

APPENDIX A: MODELLING

APPENDIX A.1: CLASSIFICATION FEATURES

By reviewing the relevant literature for building and neighborhood age prediction, we have identified a set of 89 building attributes for age classification. These attributes of classification features belong to 4 categories: general building features (listed in table A.1), building shape features (listed in table A.2), street features (listed in table A.3) and block features (listed in table A.4). The information presented contains the name and description of the feature, accompanied by the mathematical formula used for computation if computation is not trivial, and the literature sources.

Table A.1.1: Building features.

Name	Description	Reference
Height	Height of building relative to the ground (the height is measured from the highest point of the roof).	(Biljecki & Sindram, 2017) (Rosser, et al., 2019) (Tooke, Coops, & Webster, 2014)
Wall height	Difference between building height and roof height (distance between highest and lowest point of the biggest roof surface).	
Number of storeys	Number of above ground stories: $\frac{height}{3} - 1$.	(Biljecki & Sindram, 2017) (Rosser, et al., 2019) (Tooke, Coops, & Webster, 2014)
Area	Area of building footprint.	
Perimeter	Length of building outline.	
Volume	Computed considering roof geometry (flat, prism, pyramid, etc).	
Volume (default)	Computed considering the building as a parallelepiped: $area * height$.	
Floor area	Sum of all floor areas: $area * (number\ of\ stories + 1)$	
Roof slope	Angle between the ground surface and one lateral surface of the roof (the biggest roof surface is considered the most representative for the slope).	Using raster data in (Rosser, et al., 2019) (Tooke, Coops, & Webster, 2014)
Roof area	Sum of all roof surfaces.	
Vertex number	Number of vertices that make up the footprint.	(Rosser, et al., 2019)
Building parts	Number of building parts that make up the building (main building and annexes).	

Wall area	Sum of all wall surfaces.	(Tooke, Coops, & Webster, 2014)
Surface area	Sum of all surfaces, roof and walls.	
Normalized perimeter index	Indicator of building footprint compactness: ratio of the perimeter of the equal area circle to perimeter $\frac{2\sqrt{a\pi}}{p}$.	(Angel, Parent, & Civco, 2010) (Rosser, et al., 2019)
Shape complexity	Ratio of perimeter to perimeter of circle of equal area. P**2 / 4*pi	
Compactness	Exposed building per unit volume: ratio of surface area to $V^{2/3}$.	(Tooke, Coops, & Webster, 2014)
Fractal	Indicator of shape complexity and fragmentation: ratio of perimeter to area $\frac{\ln\left(\frac{p}{4}\right)^2}{\ln a}$.	(McGarigal & Marks, 1995) (Wurm, Schmitt, & Taubenböck, 2016) (Tooke, Coops, & Webster, 2014)
2D Shape index	Indicator of shape smoothness: ratio of perimeter to perimeter of the equal area square $\frac{p}{4*\sqrt{a}}$.	(McGarigal & Marks, 1995)
3D Shape index	Indicator of 3D smoothness: ratio of perimeter to perimeter of the equal volume cube.	(Wurm, Schmitt, & Taubenböck, 2016)

Table A.1.2: Building shape features.

Name	**Description**	**Reference**
Building type	'Detached', 'semi-detached', 'terraced', 'end-terraced' or 'complex'	(Droin, 2019)
Height-area ratio	Ratio of height to area.	
Detour	Length of the convex hull of the polygon. The convex hull is the shortest connection of vertexes that fully contains the polygon.	(Angel, Parent, & Civco, 2010)
Detour index	Ratio of the perimeter of the equal area circle to detour length.	
Range	Distance between the furthest most vertex points of the building footprint.	
Range index	Normalized *Range* using the diameter of the EAC (equal area circle).	
Exchange	Area of the polygon within the EAC considering they have the same centroid.	
Exchange index	Normalized *Exchange* by dividing the interior area by the EAC area.	
Cohesion	Average Euclidean distance between 30 randomly selected interior points.	
Cohesion index	Normalised *Cohesion* using the radius of the EAC and a constant defined by Angel et al. (2010).	
Proximity	Average Euclidean distance from all interior points to the centroid.	
Proximity index	The average distance of a circle to its center, given as two thirds of its radius (de Smith, 2004)	
Spin	The average of the square of the Euclidean distances between all interior points and the centroid.	
Spin index	Normalized *Spin* using half of the squared radius of the EAC.	
Number of consecutive	Number of direct neighbors.	(Droin, 2019)

neighbors		
Area neighbors	The area of the enclosing polygon containing all direct neighbors.	
Perimeter neighbors	The perimeter of the enclosing polygon containing all direct neighbors.	
Relative area neighbors	The relative area of the building feature compared to the enclosing polygon containing all neighbors.	

Table A.1.3: Street features.

Name	**Description**	**Reference**
Betweenness centrality	Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v.	(Berghauser Pont, et al., 2019)
Load centrality	The load centrality of a node is the fraction of all shortest paths that pass through that node.	Available with osmnx (Boeing, 2017)
Closeness centrality	Closeness centrality of a node v is the reciprocal of the sum of the shortest path distances from v to all n-1 other nodes.	
Degree centrality	The degree centrality for a node v is the fraction of nodes it is connected to.	
Neighborhood degree	The average degree of the neighborhood of each node.	
Distance to road	Distance from the centroid of the building footprint to closest road.	(Alexander, Lannon, & Linovski, The Identification of analysis of regional building stock characteristics using map based data, 2009)
Street length		(Gil, Beirão, Montenegro, & Duarte, 2012)
Street width		
Street connections	The number of connecting street segment to a node v.	
Street connectivity	Ratio of street length to number of intersections.	(Berghauser Pont, et al., 2019)
Street intersections		

Table A.1.4: Block features.

Final	**Description**	**Reference**
Perimeter	Length of block footprint (p)	(Hermosilla, Palomar-Vázquez, Balaguer-Beser, Balsa-Barreiro, & Ruiz, 2014)
Area	Area of block footprint (a)	
Built-up area	A ratio of built-up area size to the area of the block.	(Gil, Beirão, Montenegro,

		& Duarte, 2012)
Gross floor area per block	Building footprint area * (number of storeys + 1) (sum over buildings)	(Berghauser Pont, et al., 2019)
Floor space index	Ratio between gross floor area and block area.	
Open space ratio	Ratio between non-built area and gross floor area.	
Function richness	Number of function classes of buildings in a block.	(Lowry & Lowry, 2014)
Simpson diversity index	Function of number of classes and buildings per class: $1-\frac{\sum_i n_i(n_i-1)}{N*(N-1)}$ where N is the total number of buildings and n_i the buildings of function *i*.	
NDVI average		
Vegetation area	Area covered with vegetation (as estimated using NDVI)	(Hermosilla, Palomar-Vázquez, Balaguer-Beser, Balsa-Barreiro, & Ruiz, 2014)
Vegetation ratio	Ratio of vegetation area to block area.	
Number of buildings	Total number of buildings in a block.	
Maximum building height	The maximum height of buildings in a block.	
All storeys per block	Sum of all storey for all buildings in a block.	
Maximum number of storeys per block	Maximum number of storeys for buildings in a block.	
Built-up volume	Sum of all building volumes in a block.	
Built-up volume mean	Mean building volume.	
Built-up volume normalized	Ratio of mean building volume to block area.	
Shape index	$\frac{\mathrm{p}}{4\sqrt{\mathrm{a}}}$	
Fractal dimension	$2*\frac{\log \mathrm{p}/4}{\log \mathrm{a}}$	
Compactness	The degree to which the shape is close to a circle $\frac{4\pi \mathrm{a}}{\mathrm{p}}$.	(Berghauser Pont, et al., 2019) (Hermosilla, Palomar-Vázquez, Balaguer-Beser, Balsa-Barreiro, & Ruiz, 2014)

APPENDIX A.2: HYPERPARAMETERS TEST

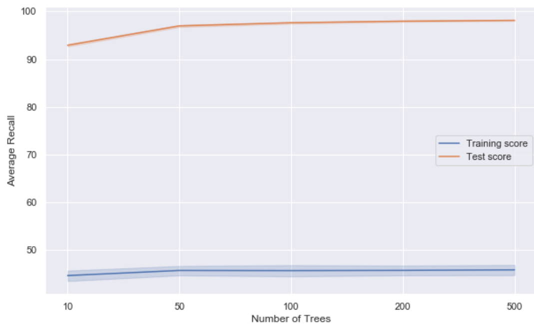


Figure A.1.1: Hyperparameter test for varying number of trees.

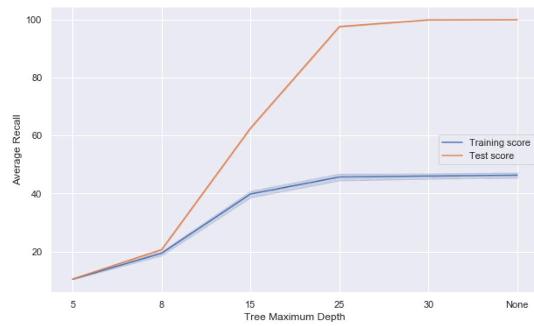


Figure A.1.2: Hyperparameter test for varying maximum tree depth.

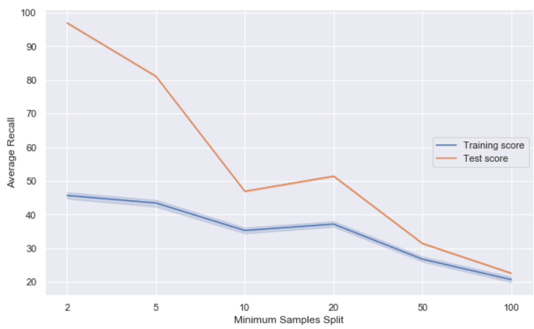


Figure A.1.3: Hyperparameter test for varying minimum samples split.

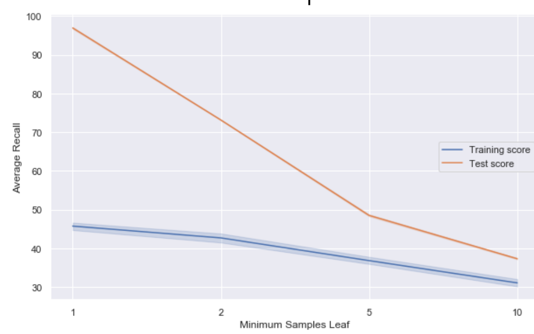


Figure A.1.4: Hyperparameter test for varying minimum samples leaf.

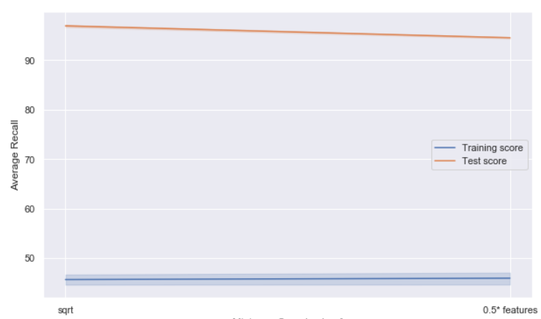


Figure A.1.5: Hyperparameter test for varying number of features.

APPENDIX B: RESULTS

APPENDIX B.1: FEATURE IMPORTANCE RANKS PER CITY

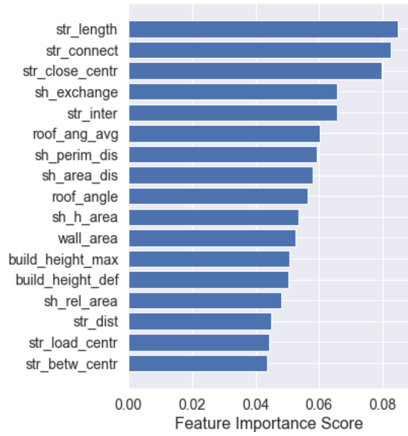


Figure B.1: Features ranked by importance in the *Model 2* classification for Cologne.

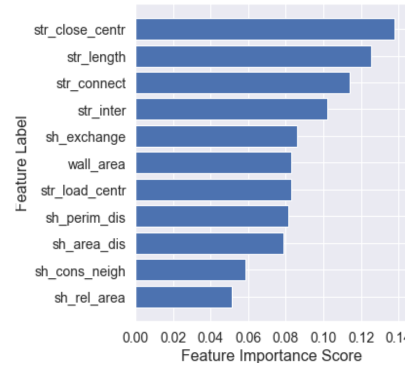


Figure B.2: Features ranked by importance in the *Model 2* classification for Düsseldorf.

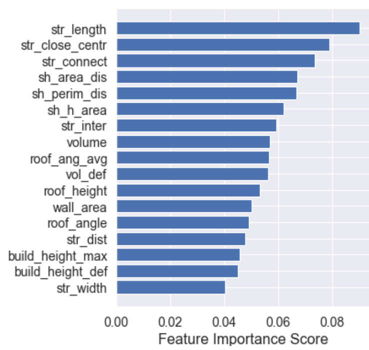


Figure B.3: Features ranked by importance in the *Model 2* classification for Dortmund.

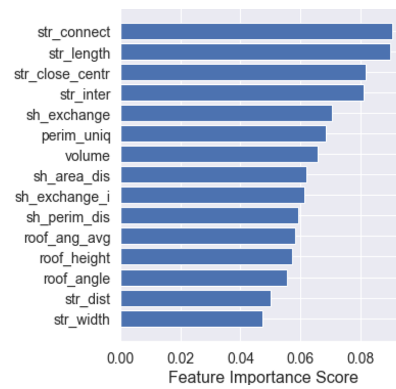


Figure B.4: Features ranked by importance in the *Model 2* classification for Duisburg

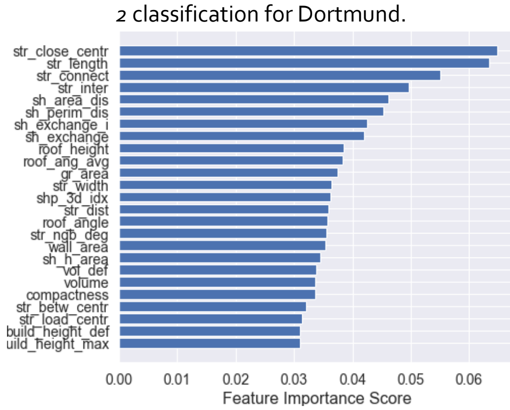


Figure B.5: Features ranked by importance in the *Model 2* classification for Essen.

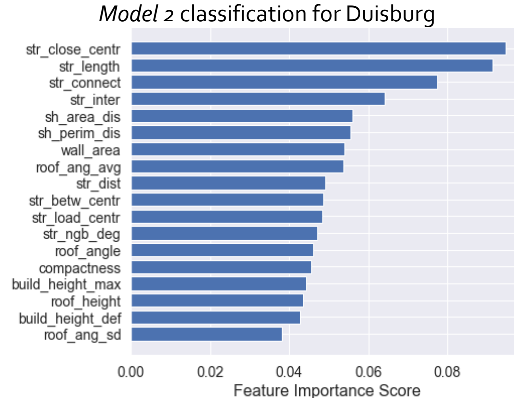


Figure B.6: Features ranked by importance in the *Model 2* classification for Bielefeld.