TU WIEN Informatics

# ArcheTypE

## Archetypische Gebäudestrukturen als Grundlage für datengestützte Energieanalysen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Software Engineering/Internet Computing**

eingereicht von

**Lukas Rast, BSc**
Matrikelnummer 01227137

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Henderik Proper, PhD
Mitwirkung: Dipl.-Ing.in Dr.in techn. Julia Forster

Wien, 24. August 2026

| Lukas Rast | Henderik Proper |
|---|---|



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

# ArcheTypE

## Archetypal building structures as a basis for data-driven energy analyses

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Software Engineering/Internet Computing**

by

**Lukas Rast, BSc**
Registration Number 01227137

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Henderik Proper, PhD
Assistance: Dipl.-Ing.in Dr.in techn. Julia Forster

Vienna, August 24, 2026

| Lukas Rast | Henderik Proper |
|---|---|

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Declaration of Authorship

Lukas Rast, BSc

I hereby declare that I have written this thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work – including tables, maps and figures – which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.
I further declare that I have used generative AI tools only as an aid, and that my own intellectual and creative efforts predominate in this work. In the appendix "Overview of Generative AI Tools Used" I have listed all generative AI tools that were used in the creation of this work, and indicated where in the work they were used. If whole passages of text were used without substantial changes, I have indicated the input (prompts) I formulated and the IT application used with its product name and version number/date.

Vienna, August 24, 2026

Lukas Rast



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Danksagung

Zunächst möchte ich mich bei meinem Betreuer Henderik Proper bedanken, dass er mir die Möglichkeit gegeben hat, an diesem praxisrelevanten und interdisziplinären Thema zwischen Informatik und Raumplanung zu arbeiten. Sein wertvoller Input und seine Unterstützung bei der Ausarbeitung meiner Ideen haben wesentlich zum Entstehungsprozess der Arbeit beigetragen. Ebenfalls möchte ich mich bei Julia Forster für ihre umfassende fachliche Beratung hinsichtlich des raumplanerischen Kontexts sowie der Entwicklung der Grundidee für diese Arbeit bedanken. Darüber hinaus möchte ich mich auch bei Stefan Bindreiter für den fachlichen Austausch und die Unterstützung bei der praktischen Umsetzung raumplanerischer Aspekte bedanken.

Bei meinen Freunden und Kollegen, die mich durch meine Studienzeit begleitet haben und mit denen ich viele unvergessliche Momente teile, möchte ich mich ebenfalls bedanken. Danke, dass ich diesen Weg mit euch beschreiten durfte!

Ein besonderer Dank gilt meiner Familie, die mir das Studium ermöglicht und mich in jeder Phase unterstützt hat. Danke, dass ihr immer hinter mir gestanden und mich in meinen Entscheidungen stets bekräftigt habt!

Abschließend möchte ich mich bei meiner Partnerin Regina Schweiger für die Unterstützung und Geduld während der intensiven Zeit der Entstehung dieser Arbeit bedanken. Dein Rückhalt hat wesentlich dazu beigetragen, diesen letzten Abschnitt meines Studiums abzuschließen!



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Acknowledgements

Firstly, I would like to thank my advisor, Henderik Proper, for providing me with the opportunity to work on this practically relevant and interdisciplinary topic in informatics and spatial planning. His valuable input and support regarding my ideas contributed significantly to the development process of this work. Also, I would like to thank Julia Forster for her comprehensive expert guidance on the spatial planning context and her contribution to the initial idea of this thesis. Furthermore, I would like to thank Stefan Bindreiter for the professional exchange and the support with the practical application of spatial planning aspects.

Additionally, I would like to thank my friends and colleagues who accompanied me through my years of study and with whom I have shared many unforgettable moments. Thank you for walking this path with me!

I would also like to extend my special gratitude to my family for enabling my studies and supporting me through every phase. Thank you for always having my back and supporting me in my decisions!

Finally, I would like to thank my partner Regina Schweiger for her encouragement and patience during this intense period of developing this thesis. Your support has significantly contributed to me finishing this last chapter of my course of study.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Kurzfassung

Die kommunale Energie- und Raumplanung benötigt räumliche Informationen, um gezielte Maßnahmen für die Energiewende und nachhaltige Quartiersentwicklung planen zu können. Wesentliche Energiekennzahlen sind häufig weder auf Gebäudeebene noch flächendeckend öffentlich verfügbar; gleichzeitig stellen Behörden umfangreiche Geodaten zur Verfügung, deren Potenzial für Energieanalysen nur begrenzt ausgeschöpft wird. Diese Arbeit untersucht am Beispiel der Gemeinde Bruck an der Leitha, inwiefern sich diese Datenlücke mithilfe dieser Geodaten schließen lässt. Dafür werden simulationsbasierte Energiekennwerte als Zielgrößen verwendet. Der methodische Ansatz kombiniert die Aufbereitung morphologischer, funktionaler und räumlich-struktureller Merkmale mit einem zweistufigen hierarchischen Clustering auf Basis von partitionierten Gower-Distanzen. Die so erzeugten Gruppen werden zu fachlich interpretierbaren Gebäude-Archetypen zusammengeführt und durch Expertenannotationen sowie funktionalen Plausibilitätscheck und Stabilitätsanalysen validiert. Anschließend werden CatBoost-Regressionsmodelle für die gebäudebezogene Prognose von Stromverbrauch, Heizenergie, Primärenergie und $CO_2$-Emissionen trainiert. Anhand eines Vergleichs zwischen Modellvarianten mit und ohne Archetypeninformationen wird deren prädikativer Nutzen unter räumlicher und zufälliger Cross-Validation bestimmt. Im Zuge des Clusterings wurden insgesamt 22 Cluster identifiziert, die in einen Katalog aus 14 Gebäude-Archetypen überführt wurden. Die Zuordnungen erreichten eine Übereinstimmung von $84,6 - 87,2\%$ mit den Expertenannotationen und wurden zu $99,5\%$ als funktional plausibel eingestuft. Die Stabilitätsprüfungen bestätigten mit einer Zuordnungsübereinstimmung von über 90% eine robuste Struktur. Die Ergänzung der Modell-Eingangsmerkmale um die Archetypen senkte den mittleren absoluten Fehler der Prognosen der vier Energiekennwerte um $1,66 - 5,75\%$. Sie führten für beide Validierungsverfahren in 17 von 20 Folds zu einer Verbesserung, die sich breit über die Gemeinde verteilte. Abschließend wurden die gebäudespezifischen Vorhersagen auf einem 250-Meter-Raster zu Indikatoren für Energieintensität, Lastdichte, Sanierungsspotenzial und Gebäudestruktur verdichtet. Die Arbeit liefert damit eine räumlich explizite Informationsgrundlage zur Unterstützung kommunaler Planungsprozesse.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

The communal energy and spatial planning require spatial information to plan targeted measures for the energy transition and sustainable neighborhood and quarter development. Significant energy metrics are often neither publicly available at the building level nor with comprehensive spatial coverage; at the same time, authorities make extensive geodata available, whose potential for energy analyses is only limitedly used. This thesis examines, with the example of the municipality Bruck an der Leitha, how this data gap can be closed with the help of this geodata. As target values, simulation-based energy parameters are used. The methodical approach combines the processing of morphological, functional and spatial-structural features with a two-step hierarchical clustering based on partitioned Gower distances. The so-created groups are combined into building archetypes, which are interpretable from a technical perspective and validated through expert annotations as well as functional plausibility checks and stability analyses. Subsequently, CatBoost-regression models are trained for the building-level prognosis of electricity consumption, heating energy, primary energy and $CO_2$ emissions. Based on a comparison between model variants with and without archetype information, their predictive use under spatial and random cross-validation is assessed. In the course of the cluster, a total of 22 clusters were identified, which were transformed into a catalog of 14 building archetypes. The assignments reached a consistency of $84.6 - 87.3\%$ with the expert annotations and were classified as functionally plausible to a degree of 99.5%. The stability checks confirm a robust structure with an over 90% agreement rate. The addition of the model-input features of the archetypes reduced the mean absolute error of the predictions of the four energy indicators by $1.66 - 5.75\%$. For both validation processes, they improved in 17 out of 20 folds, which stretch far across the municipality. Finally, the building-specific predictions were aggregated on a 250-meter grid into indicators for energy density, load density, renovation potential and building structure. This work thus provides a spatially explicit information basis to support communal planning processes.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Contents

**Kurzfassung** **xi**

**Abstract** **xiii**

**Contents** **xv**

**1 Introduction** **1**
1.1 Motivation and Problem Statement . . . . . 1
1.2 Research Questions . . . . . 2
1.3 Anticipated Contribution . . . . . 3
1.4 Research Design . . . . . 3
1.5 Structure of this Thesis . . . . . 5

**2 Background** **7**
2.1 Spatial Planning in Rural Municipalities . . . . . 7
2.2 Building Archetypes and Clustering . . . . . 9
2.3 Machine Learning for Energy Indicator Prediction . . . . . 13

**3 Related Work** **17**

**4 Data Sources and Feature Engineering** **19**
4.1 Datasets . . . . . 19
4.2 Feature Engineering . . . . . 21

**5 Derivation of Building Archetypes** **29**
5.1 Aim and Conceptual Approach . . . . . 29
5.2 Feature Selection and Partitioning . . . . . 31
5.3 Clustering Workflow . . . . . 33
5.4 From Cluster Labels to Building Archetypes . . . . . 57
5.5 Validation Process of Archetypes . . . . . 62

**6 Energy Prediction Model** **69**
6.1 Training of Prediction Model . . . . . 69
6.2 Evaluation of the Prediction Model . . . . . 75



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

6.3 Derivation of Spatial Indices . . . . . 81

**7 Conclusion** **83**
7.1 Synthesis and Discussion of RQ1 . . . . . 83
7.2 Synthesis and Discussion of RQ2 . . . . . 85
7.3 Planning implications . . . . . 86
7.4 Limitations . . . . . 87
7.5 Future Work . . . . . 87

**Overview of Generative AI Tools Used** **89**

**List of Figures** **91**

**List of Tables** **93**

**Glossary** **95**

**Acronyms** **97**

**Bibliography** **99**

**Appendix** **105**
A Annotation Guidelines . . . . . 105
B Usage-Archetype Compatibility Catalog . . . . . 107
C Archetype Annotation Validation . . . . . 109
D Archetype Stability Validation . . . . . 110
E Out-of-Fold Prediction Diagnostics . . . . . 110
F Spatial Out-of-Fold Error Diagnostics . . . . . 115
G Spatial Indices . . . . . 116

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

# Introduction

Municipalities require spatially detailed evidence to plan energy transition measures in the building stock, yet many energy-relevant attributes (e.g. energy consumption and heating systems) are not publicly available at the building level. At the same time, municipalities and public authorities increasingly provide rich geospatial data through geographic information system (GIS), which digitally locates buildings and parcels and describes their geometry and spatial relationships. This data provides information on the physical form and spatial configuration of buildings as well as on functional characteristics. This thesis investigates how such morphological and functional descriptors can be translated into interpretable building archetypes and used as inputs for data-driven prediction of energy indicators. For an examined training municipality, simulation-based energy variables are used to produce building-level forecasts; subsequently, these are used for aggregated spatial indices along with uncertainty information.

## 1.1 Motivation and Problem Statement

The sustainable development of municipalities requires a deep understanding of how settlement structures and the energy behavior of buildings interact with each other. In this thesis, morphological features describe the measurable physical form and spatial configuration of buildings and their surroundings. In the building context, these include characteristics such as building footprint, volume, compactness, building density and development intensity [SW24]. Functional attributes, for example construction period, building use or technological equipment (e.g. heating system), reflect the use and energy standard of the building stock.

Morphological and functional features as well as spatial context together influence the energy requirements and energy efficiency of buildings and thus also entire settlement structures [ABT22]. High building density, for example, can lead to more efficient energy

supply, while different uses and building age classes have a significant impact on energy consumption due to user behavior and energy standards [NSM+23].

For municipalities, such relationships are highly relevant, as planning measures in the context of the energy transition must be spatially differentiated and adapted to local building structures. However, comprehensive energy indicators such as annual energy consumption or heating type are usually available only to a limited extent at the building level, as they are generally not publicly accessible. In contrast, extensive GIS data, for example on the geometry of buildings and parcels, are often freely available but have so far hardly been used to derive energy-related indicators.

The resulting data gap between freely available geospatial data and energy indicators with restricted access forms the starting point for this work. The methodological approach of this thesis bridges this gap by estimating the energy characteristics for buildings of rural municipalities based on morphological and functional features [WWG+25]. In addition, planning-relevant parameters are derived in order to analyze and interpret the results in their spatial context.

## 1.2 Research Questions

The objective of this work is to develop a method that analyzes the morphological and functional characteristics of municipalities and derives building archetypes. These identified archetypes are then used to predict energy indicators and relevant urban-planning parameters.

The methodological approach is based on a machine learning process developed within a training municipality, which serves as the ground truth, using simulated energy indicators from the SmartQ+ research project [BSR+25]. The first step is to identify archetypal building classes based on morphological and functional characteristics. An energy model is then trained using these archetypes, as well as further building data to predict energy indicators such as energy consumption [GTTB21]. Model performance is evaluated within the same municipality using a train–test split of the available data to assess prediction accuracy. The use of archetypes is intended to improve the robustness and generalizability of the model predictions.

Since archetypes act as a bridge between geospatial data and energy indicators, they must be interpretable as well as stable and provide measurable prediction improvements compared to models without archetypes. This leads to the following research questions:

- **RQ1**: How can building archetypes be derived from morphological and functional features, and how stable and interpretable are these archetypes across the municipality?
- **RQ2**: How does the inclusion of archetype labels influence the accuracy, robustness and uncertainty of the energy prediction models?



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 1.3 Anticipated Contribution

This thesis develops and evaluates a workflow to infer building-level energy indicators from widely available geospatial data by combining unsupervised archetype formation with supervised machine learning. The contribution of the work lies in developing a workflow for the clustering of building groups from various municipal data sources, introducing an interpretable archetype representation of the building stock and assessing whether this representation improves predictive performance and robustness compared to models without archetype information. In addition to building-level predictions, the workflow supports municipal-scale analysis by producing spatially explicit indicators that can be aggregated to and mapped on a regular grid.

The objectives of this thesis are:

- *O1*: Set of archetypal building classes representing combinations of morphological and functional attributes of buildings.
- *O2*: Trained energy prediction model (regression) using simulation-based energy indicators.
- *O3*: Assessment of how archetype labels affect the energy models' accuracy, robustness and uncertainty.
- *O4*: Derived spatial indicators and aggregated energy metrics at grid level.
- *O5*: Visualizations to display derived indicators and prediction model estimates.

## 1.4 Research Design

As illustrated in Figure 1.1, the methodological steps are built on each other and are designed to contribute to specific objectives (O1-O5) and to address the research questions (RQ1, RQ2), as outlined in the following subsections.

1. Data collection & feature engineering
2. Archetype formation & validation
3. Training & evaluation of the energy prediction model
4. Derivation of planning-relevant indices



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

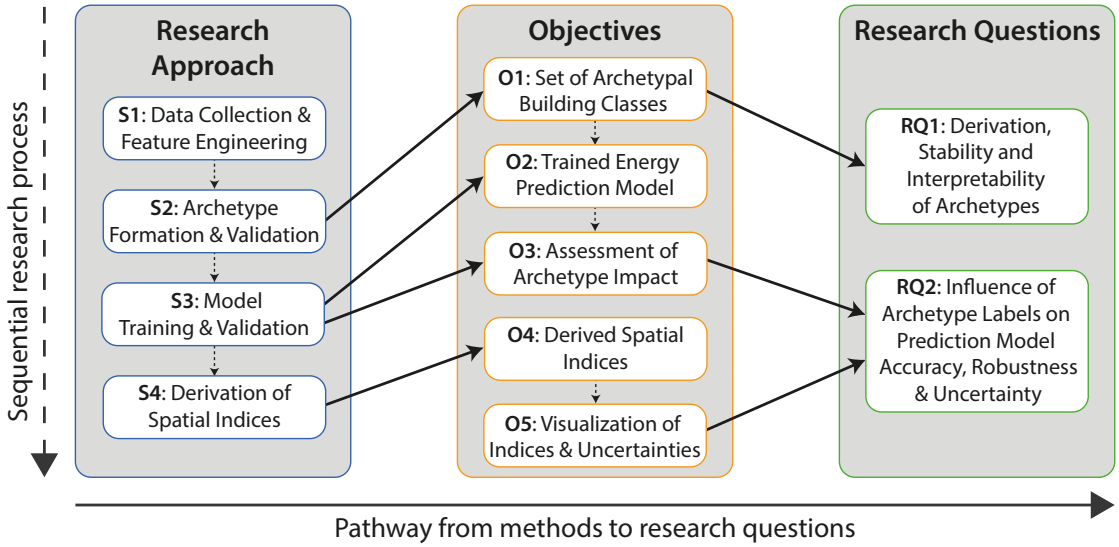


Figure 1.1: Research framework linking research approach, objectives and research questions.

### 1.4.1 Data Collection & Feature Engineering

The aim of this step is to create a consistent geospatial dataset for the municipality of Bruck an der Leitha for the formation of archetypes. The existing building and simulated energy data from the SmartQ+ research project is to be expanded to include GIS data on buildings and parcels, neighborhood structures and derived morphological features to form a consistent building-level dataset [BSR+25]. Further feature engineering will be applied to harmonize and standardize attributes and handle missing values. This step creates a feature set for the derivation process of the building archetypes and thus addresses RQ1. Furthermore, the resulting dataset will also be partially used as input data for the training of the energy prediction model.

### 1.4.2 Archetype Formation & Archetype Validation

In this step, building archetypes are to be derived from morphological, functional and spatial-structural characteristics at the building level, resulting in a set of archetypal building classes [SW24]. These archetypes form a reduced and interpretable typology of the building stock for subsequent energy analyses [DJLS20]. Although they are defined in general terms, their derivation and validation in this thesis are restricted to the municipality of Bruck an der Leitha, which serves as the case study area.

The derived building groups are assigned descriptive archetype labels and thus are mapped into an archetype catalog. The resulting archetype assignments are evaluated to assess whether they represent a meaningful typology of the local building stock. The validation results directly address RQ1 by assessing the archetypes' stability and interpretability



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

across the municipality.

### 1.4.3 Training & Evaluation of the Energy Model

Based on the preprocessed training dataset and the validated archetype labels, an energy prediction model is trained to predict annual energy indicators on building level [OAAS+22]. The model learns the relationships between available GIS data and simulation-based energy information [WWG+25]. Particular attention is given to the archetype labels, providing additional and compressed information of the building stock for the energy indicator predictions [SW24].

To evaluate the significance of archetypes for model performance, ablation studies are conducted in which the model is trained once with archetype labels and once without them. The resulting predictions of both models are evaluated and compared in regard to their accuracy and robustness. The assessment of the archetype labels' impact on model performance addresses RQ2.

### 1.4.4 Derivation of planning-relevant Indices

This step aims to aggregate the building level predictions to capture spatial patterns across the municipality. By combining these predictions with archetypes and building characteristics, additional spatial indices can be derived. By visualizing these indices, model outputs are translated into interpretable representations for planning applications, which highlights the interdisciplinary nature of the work. At the same time, spatial representations support the practical evaluation of model behavior regarding RQ2.

## 1.5 Structure of this Thesis

This thesis is divided into seven chapters which reflect the development process. Following this introduction, Chapter 2 provides the necessary background for a better understanding of the interdisciplinary approach between the domains of computer science and spatial planning. Here, the foundations for spatial planning in rural municipalities and the role of building archetypes are described as well as their derivation through clustering processes and, lastly, approaches for machine learning procedures for the prediction of energy metrics. Chapter 3 subsequently gives an overview over related work to describe the current state regarding the defined problem statement.

Before the actual implementation, in Chapter 4 the creation and processing of the dataset is described. This serves the purpose of introducing the reader further to the problem statement and to familiarize them with feature groups and building features. Additionally, the respective measures, which have been chosen for possible weaknesses or missing data, are described.

In Chapter 5, the conceptual approach for the clustering is described; this should help the reader understand the subsequent decisions better because the desired goal of the



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

approach is described. This, for example, includes the partitioning of the building stock, the chosen input features and the structure of the clustering workflow. Then, the process of clustering is explained in further detail and the interim results of the individual steps to the final result are interpreted. Finally, the translation into an archetype catalog as well as the validation of the building archetypes are described.

Chapter 6 describes the training and the evaluation of the energy prediction model. There, among other things, the decision process for the final target variables and the fine-tuning of the model is described in further detail. Additionally, the evaluation of the predictions, which is used to assess the quality of the model, is found here. These metrics serve the final comparison of the different model configurations, which are used for the evaluation of how the inclusion of archetype labels affects the prediction accuracy, robustness and inaccuracy. Following their evaluation, the model predictions are used to derive spatial indices, which are subsequently visualized in order to provide an information basis at the municipal level.

In the final Chapter 7, the results of this thesis are described in more detail and are discussed regarding the research questions. Furthermore, the implications for spatial planning are discussed. Finally, the limitations of the approach as well as possible future work in the area are discussed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

# Background

This chapter provides the conceptual and methodological background for the thesis. First, the role of spatial planning in rural municipalities is introduced and it is explained why spatially explicit information on settlement structures and the building stock is relevant for planning decisions. Furthermore, building archetypes as an approach for reducing the complexity of building stocks is discussed and the concept of clustering for archetype derivation is introduced. Finally, the chapter outlines the machine-learning-based prediction of energy-relevant building indicators.

## 2.1 Spatial Planning in Rural Municipalities

Spatial planning provides the institutional and strategic framework for steering the spatial development of settlements and land use. In Austria, it is organized as a multi-level public task involving the federal level (Bund), the state level (Länder) and the municipal level (Gemeinden) [Ös18]. In the context of spatial local planning, municipalities are in a central position, since it falls within their own sphere of responsibility. The Austrian Spatial Development Concept (ÖREK 2030) frames spatial development as climate-friendly and sustainable and calls for spatial resources to be used carefully; accordingly, settlement development must be considered in relation to climate mitigation, resource efficiency and land consumption [Ös21].

Against this background, building stock information serves as an evidential basis for local spatial planning; planning-relevant analyses require not only insight into structural composition and spatial organization, but also energy-relevant characteristics of the building stock. It supports the assessment of existing settlement patterns and development potentials, consistent with "Energieraumplanung", which integrates spatial dimensions of energy consumption and supply into spatial planning [SEKN17]. One important indicator is the morphology of the settlement structure, describing the form and arrangement of

buildings and built-up areas, including features such as footprint size, compactness and spacing between buildings [ZGP23].

Rural municipalities require their own assessment for planning decisions, since these settlement structures differ from urban contexts. In Austria, lower-density, dispersed development is associated with higher land use per capita, especially where detached single-family houses dominate [Ös23], higher car dependence and ongoing land take [Ös21]. Identifying compact and dispersed areas within a municipality helps to reveal inner development potential [Ös17]; detailed building stock information can also capture differences in energy-relevant characteristics [HML+25].

This internal heterogeneity can also be illustrated morphologically. As shown in Figure 2.1, the urban taxonomy classification of Bruck an der Leitha reveals a mix of low-density small-scale fabrics (green), denser built fabrics (yellow) and large-scale heterogeneous fabrics typical of industrial and service areas within the same municipality, rather than a single uniform "rural" fabric.

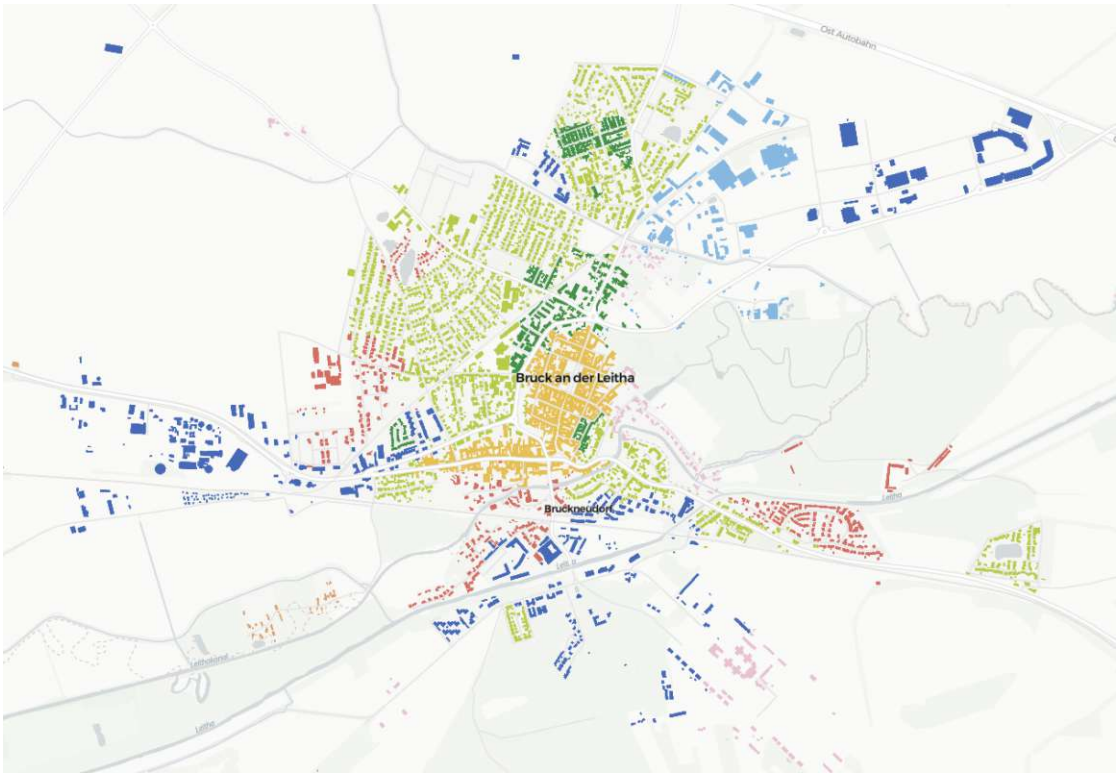


Figure 2.1: Urban Taxonomy map of Bruck an der Leitha, illustrating the morphological heterogeneity within a rural municipality.

*Source:* Screenshot from `https://urbantaxonomy.org` (last accessed: 2026-08-07).

Planning in rural areas, therefore, benefits from spatially explicit and interpretable evidence, since decisions concern specific buildings and neighborhood structures that



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

municipal-level aggregates alone cannot represent adequately. Regarding energy-relevant characteristics, urban building energy modeling (UBEM) can provide building-level insights and thereby support identifying spatial patterns of energy demand and retrofit potential [HML+25]. Methods that derive structured and location-specific evidence from available geodata can therefore provide valuable support for planning practice [TPP25].

## 2.2 Building Archetypes and Clustering

This section introduces building archetypes as a way to represent heterogeneous building stocks through interpretable groups and discusses clustering as a method for deriving such groups from building data.

### 2.2.1 Building Archetypes in Energy and Planning Contexts

Building archetypes represent building types with similar morphological, functional or energy-related characteristics. Therefore, they provide a simplified representation of the building stock by grouping buildings with comparable properties. In UBEM, archetypes can support the analysis of energy demand, retrofit potential and further planning-relevant indicators without requiring each building to be simulated individually [DJLS20, SW24].

The definition of building archetypes depends on the purpose of the analysis and on the available data. They are commonly defined using criteria such as building use, construction period, size class or building systems. This allows the representation of larger building stocks through a limited number of building typologies, which can be used for energy calculations and scenario analyses [LHBR23].

An established example in Austria is the TABULA/EPISCOPE building typology. It classifies residential buildings primarily by construction period and building type, distinguishes between single-family houses, row houses and multi-family houses and apartment blocks and uses representative model buildings to describe the energy-related characteristics of these typology classes [Aus15]. Heidenthaler et al. applied a related approach using energy certificate data from Salzburg, Austria, for deriving aggregated archetypes. They showed that these archetypes can provide a useful basis for energy assessments [HLR+22]. However, the discussed typologies are less suited for capturing detailed spatial differences within a municipality, since communal planning requires a more spatially explicit view of the local settlement fabric. This creates a need for archetypes that can reflect the local building stock characteristics.

### 2.2.2 Archetype Formation by Clustering

Data-driven archetype formation uses building data to derive a representation of the building stock rather than relying on predefined typology schemes. Domain knowledge can help to refine the selection and structuring of relevant building features, allowing local characteristics to be considered [PWK19]. Against this background, clustering provides a common approach for deriving such groups from available building data.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

As an unsupervised learning method, it identifies structure without predefined labels by grouping observations according to their similarity; the resulting building groups are therefore inferred from patterns in the input feature space [JMF99]. If they are technically meaningful and plausible, these groups can be interpreted as archetypes that reflect characteristics of the building stock from which they were derived [DJLS20]. In this work, this refers to the examined building stock of the municipality of Bruck an der Leitha.

### 2.2.3 Distance Measures for Mixed-Type Building Data

Clustering results depend largely on how the similarity between observations is measured. The choice of distance measure should therefore match the selected input features, since it determines which differences are emphasized in the resulting groups [Jai10]. For numerical data, Euclidean distance is commonly used; it measures the straight-line distance between two observations in a given feature space. Manhattan distance instead sums the absolute differences across all features. Both methods are sensitive to feature scaling, as larger numerical ranges can heavily influence the resulting distances [eve11].

Building-stock data, however, not only contain variables with different scales but also often combine varying types of information, such as continuous measures, binary indicators and categorical attributes. For such mixed-type data, Gower dissimilarity provides a general approach for calculating pairwise dissimilarities by combining variable-specific comparisons in a common distance measure [Gow71]. For two observations $i$ and $j$, Gower dissimilarity can be written as

$$d_{ij} = \frac{\sum_{k=1}^{p} \delta_{ijk} w_k d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk} w_k} \tag{2.1}$$

whereby $d_{ijk}$ denotes the dissimilarity for variable $k$, $w_k$ is an optional variable weight and $\delta_{ijk}$ indicates whether a valid comparison is available for that variable. Numerical differences are normalized by the variable range, while categorical and binary values distinguish only between match and mismatch [KR09]. This allows heterogeneous building data to be considered within the same dissimilarity measure without one-hot encoding.

### 2.2.4 Clustering Algorithms for Archetype Formation

Clustering algorithms differ in how they determine and form clusters. Common methods include partitioning, hierarchical and density-based approaches [JMF99].

Partitioning methods such as k-means and k-prototypes assign observations to a predefined number $k$ of clusters. K-means is efficient and widely used, but primarily designed for use with numerical values in a Euclidean feature space [Jai10]. K-prototypes addresses this by allowing mixed-type data by representing clusters through numerical means and categorical modes; their contributions to the clustering are controlled by a weighting parameter [Hua98]. Hierarchical clustering organizes observations into a nested structure,



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

represented by a dendrogram, where the leaves represent individual observations and branches show how clusters are merged. By cutting the dendrogram at different levels, this approach allows for a more detailed comparison of subgroups. Thus, an appropriate level of detail for the resulting clusters can be chosen [KR09, FSB+26]. Density-based methods, such as DBSCAN identify dense areas separated by lower density areas and can detect noise points. A limitation is that its results depend on parameters such as the neighborhood radius and the minimum number of points [EKS+96].

For archetype formation, the resulting groups should be statistically meaningful and interpretable to describe building and settlement structures. This makes the relationship between feature selection, distance measure, clustering algorithm and subsequent interpretation particularly important.

Table 2.1: Selected clustering approaches for archetype formation.

| Algorithm family | Principle | Relevance |
|---|---|---|
| $k$-means | Groups observations around numerical cluster centroids | Efficient, but limited to numerical feature space |
| $k$-prototypes | Extends k-means to mixed-type features | Can handle categorical features, but requires weighting between type categories and a predefined $k$ |
| Hierarchical clustering | Nested sequence of clusters | Allows different levels of aggregation to be examined |
| DBSCAN | Identifies dense areas separated by sparse areas | Can detect noise, but depends strongly on density parameters |

### 2.2.5 Cluster Interpretation and Validation

Because unsupervised clustering normally lacks ground-truth labels, the resulting groups cannot be validated through classification accuracy. Instead, the validity of the solution depends on a combination of internal quality measures and substantive interpretation [Jai10]. A common internal validation measure is the silhouette coefficient, which compares an observation's dissimilarity within the assigned cluster with that to the nearest alternative cluster [Rou87]. For an observation $i$, the silhouette value is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.2}$$

where $a(i)$ is the mean dissimilarity between observation $i$ and all other observations in the same cluster, and $b(i)$ is the mean dissimilarity between $i$ and the nearest alternative cluster. Higher silhouette values show that an observation fits well within its own cluster,



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

indicating a clear separation from neighboring clusters. Low or even negative values point to ambiguous assignments.

To better assess cluster quality, the aforementioned silhouette metric can be supplemented by the mean within-cluster distance, which captures the pairwise dissimilarity of all observations within a cluster. For a cluster $C_c$, it is defined as

$$\bar{d}_{\text{within},c} = \frac{\sum_{i<j,\, i,j \in C_c} d_{ij}}{\binom{|C_c|}{2}}, \qquad |C_c| \geq 2 \tag{2.3}$$

where $|C_c|$ is the number of observations in that cluster and $d_{ij}$ is the dissimilarity between the observations $i$ and $j$ [BP98]. A lower value indicates a more compact cluster. These values can also be aggregated to reflect the mean within-cluster distance of a cluster solution and thereby assess its homogeneity [Hen19]. In addition, values such as the minimum cluster size and the share of the largest cluster of a clustering solution can help to identify very small clusters that can fragment the solution or dominant clusters that can lead to a coarse solution [Was24].

The reliability of the assignment can be assessed for individual buildings by comparing their mean distance to the assigned cluster with that of the nearest alternative [Rou87]. A large margin shows a clear assignment, while small values suggest a boundary case; many such cases in a cluster indicate a weak separation from other clusters. Following a dissimilarity-based approach to cluster membership [LG19], this work uses a normalized distance-based affinity:

$$A_{ic} = \frac{m_{ic}^{-1}}{\sum_{h=1}^{K} m_{ih}^{-1}}, \qquad m_{ih} > 0 \tag{2.4}$$

where $m_{ic}$ is the mean dissimilarity between a building $i$ and cluster $c$, while $K$ denotes the overall number of clusters. Higher affinity values indicate a smaller distance to cluster $c$ compared to the other clusters. In this thesis, the difference between the highest and second-highest affinities is referred to as the affinity gap; a large value indicates a strong assignment, while smaller gaps indicate ambiguous cluster assignments.

Accordingly, internal metrics can indicate whether a cluster solution is compact and shows a good separation, but they cannot assess if the solution represents a meaningful building typology; this requires domain-specific knowledge of the building stock, spatial context and the intended application of the archetypes. Good quality metrics usually enhance interpretability, but there is no guarantee. In contrast, a solution with moderate compactness and separation may still be useful if it is fundamentally plausible and represents interpretable archetypes [PWK19, Hen19].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 2.3 Machine Learning for Energy Indicator Prediction

For the prediction of energy metrics, machine learning models can be used. There, known building features are used to teach the model relationships between these features and available energy metrics. This way, predictions can be made without creating a complex physical simulation for every building. In the course of this thesis, supervised regression is applied to predict energy metrics based on the morphological and functional features as well as the spatial-structural context. For this reason, this section introduces the underlying methods, relevant regression approaches as well as the metrics used to evaluate the prediction quality.

### 2.3.1 Supervised Regression and Surrogate Modeling

As opposed to the unsupervised clustering described in Section 2.2, for supervised learning known target values are present along with input features. There, for every observation $i$, input variables are present as feature vector $\mathbf{x}_i$ and a continuous target value $y_i$. Based on these, the model learns a function $\hat{f}$, which maps the input features to predicted target values:

$$\hat{y}_i = \hat{f}(\mathbf{x}_i) \tag{2.5}$$

$\hat{y}_i$ refers to the predicted value of the model for the observation $i$. In this thesis, an observation refers exactly to one building. Because the target values are continuous numerical values, this task is referred to as regression [JWH+21]. A distinction is made between aggregated and time-resolved forecasts. The former estimates a single aggregated target value, while time-resolved models represent the temporal development of the target value.

The target variables, which are used for the training of such a model, are either derived from real-world measurements or from physical simulations. In both cases, the model only learns the statistical relations between the input and target variables; thus, no physical or energetic process is simulated. A model, which is trained based on results of a more complex simulation model, is called a surrogate model. Based on the learned input-output relationship, the surrogate model approximates the initial simulation for new buildings without requiring computationally intensive calculations [FSK08]. However, it is notable that no empirically measured energy consumption can be predicted, only simulation-based key figures. Consequently, in this case, the quality of the model reflects the approximation to the original simulation, not how exactly the real energy consumption of buildings was predicted.

### 2.3.2 Regression Approaches

The choice of the model class affects prediction quality and determines which relations between the input feature space and target values can be mapped. Thus, the choice



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of a suitable model depends on the available input data and the specific objective of the analysis task. For the use case of the prognosis of energy metrics, relevant model approaches are summarized in Table 2.2.

Table 2.2: Overview of selected regression approaches for the prediction of energy indicators.

| **Approach** | **Principle** | **Relevance** |
|---|---|---|
| Linear Regression | Models linear relationship between input features and target variable | Transparent, but limited to nonlinear relationships |
| Random Forest | Averages multiple independently trained decision trees | Robust representation of nonlinear relationships |
| Gradient Boosting | Sequential development of trees that reduce the remaining prediction error | Methodological basis of XGBoost and CatBoost |
| XGBoost | Regularized and optimized Gradient Boosting approach | Established approach for structured tabular data |
| CatBoost | Gradient Boosting approach with dedicated processing of categorical features | Good for heterogeneous building data with categorical values |
| Support Vector Regression | Kernel-based regression for nonlinear relationships | Requires feature scaling and numerical encoding |
| Neural Networks | Multilayer approximation of complex relationships | Requires higher effort for tuning and regularization |

The linear regression assumes relatively simple additive relations. More complex relationships can, for example, only be taken into account through explicitly defined transformations, which is why they are more transparent and easier to understand [JWH+21]. In contrast, more complex approaches, such as tree-based ensemble methods, can also map non-linear relations between features and the target variable. Random forests combine multiple decision trees, trained on different subsets of features and observations. By averaging the predictions of these trees, the variance caused by individual unstable trees is reduced [Bre01].

Gradient boosting also creates an ensemble of decision trees, whereby these are not formed independently from one another. In this method, the model is expanded step by step because every newly added tree maps the structures from the observations that have not yet been covered sufficiently by the tree. The created model can be represented as the sum of the sequence of submodels:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x}) \tag{2.6}$$

where $F_{m-1}(\mathbf{x})$ represents the model before step $m$ and $h_m$ represents the newly added decision tree in this step. $\eta$ denotes the learning rate of the model, which controls the contribution of the added tree to the model [BH07]. New decision trees are adapted



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

to the negative gradient of the loss function, thereby reducing the remaining error. The sequential procedure allows that non-linear relations of the feature space can be mapped [Fri01]. XGBoost is an implementation of gradient boosting designed for computational efficiency. Thanks to the regularized learning and the use of an optimized approach for the construction of the trees, this approach is especially useful for structured tabular data [CG16].

CatBoost is another variant of gradient boosting, which was developed specifically processing categorical features. Instead of representing these through one-hot encoding, these are transformed into numerical values using target-based statistics. The observations are viewed in random order, whereby respectively only the target values of the earlier observations are used. Through this, it should be prevented that the own target value enters their feature representation. Additionally, ordered boosting is used, through which the remaining error of an observation is determined using a model trained only on preceding observations. Distortions in sequential modeling should be reduced through this approach [PGV+18].

Additional model approaches include support vector regression (SVR) and neural networks (NNs). However, SVR requires numerical coding and careful feature scaling [SS04], while NNs require a higher effort of the model configuration and tuning [GBCB16]. For the present tabular dataset which includes both numerical as well as categorical features, the CatBoost approach was favored.

### 2.3.3 Regression Model Evaluation

To determine the quality of the predictions, the model must be evaluated on observations that have not been used for training. The approach of the cross-validation (CV) splits the dataset into training and test folds. Subsequently, for each observation, an out-of-fold prediction is made from a model trained without that observation. These out-of-fold predictions form the foundation for calculating the following evaluation metrics [JWH+21].

The mean absolute error (MAE) describes the average absolute deviation between the actual target values, $y_i$ and the predictions of the model $\hat{y}_i$:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.7}$$

The MAE uses the same unit as the respective target variable, through which the mean error of the prediction can be interpreted directly. Additional metrics are the root mean squared error (RMSE), whereby large deviations are weighted more strongly than with MAE through the squaring, and the coefficient of determination $R^2$, which sets the prediction errors of the model in relation to the variance of the observed target values [CWJ21].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

# Related Work

In recent years, many approaches have been established using various combinations of building archetype-based modeling, machine-learning approaches as well as GIS-based methods for analyzing and predicting building energy performance indicators. Many of them operate at district or city scale, placing them within UBEM as reviewed by Ali et al. [ASH+21]. Since this thesis aims to predict energy indicators using archetypes derived from GIS data, this chapter focuses on studies combining these elements before describing methods for archetype derivation and energy forecasting using CatBoost.

Wang et al. [WWG+25] demonstrated that GIS data can be used to train an NN to forecast energy demands, such as heating or electricity. Their results show that such data, when sufficiently detailed, has the potential to capture nonlinear relationships between urban morphology and energy demand. Gholami et al. [GTTB21], on the other hand, classified the city of Bologna into eleven predefined archetypes and used a surrogate model for the approximation of energy intensity simulations. Kristen et al. [KHP20] used predefined archetypes, primarily from the TABULA/EPISCOPE typology, to develop physics-based archetype models based on hourly heat measurements from 927 houses in Aarhus. Both works show the ability of building archetypes as a bridge between building stock and energy modeling.

While the archetypes in the previously mentioned research works were predefined, Shen and Wang [SW24] examined the differences between standard-based, data-driven and hybrid approaches for the derivation of archetypes. They state that the selection of the approach depends heavily on the objective, data quality, and required accuracy. Deng et al. [DCYC22] developed such a hybrid approach for approximately 69,000 building footprints in Changsha. For that reason, they derived building types and construction periods from satellite imagery and real estate data and grouped them into 66 rule-based archetypes.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

De Jaeger et al. [DJLS20] followed a data-driven approach by comparing k-means, hierarchical clustering and random grouping for 1,230 single-family houses. The resulting archetypes were evaluated using dynamic energy simulations, with k-means using min–max scaling yielding the best results. They also emphasized that the selection of the feature set depends on the energy indicator under examination; the same applies to the clustering approach, which must be aligned with the intended application of the archetypes. Meshkinkiya and Hashemi [MH26] examined the derivation of archetypes using mixed data of numerical and categorical building features. Therefore, they compared hierarchical clustering based on Gower distance and k-prototypes. Whereas the latter produced more compact clusters in regard to numerical similarities, the Gower distance formed more homogeneous clusters with better separation. This highlights the need to align the choice of distance measure and clustering approach with regard to the intended interpretation of the archetypes.

For the prediction of building energy indicators, Pan et al [PZ20] trained a CatBoost model based on twelve features from Seattle's energy benchmarking data. They compared the results with Random Forest and the conventional gradient boosting model using cross-validation; both were outperformed by the CatBoost model. Qu and Liu [QL26] combined CatBoost with various metaheuristic optimization in their work to forecast energy demands for residential buildings from architectural features. Both studies showed that CatBoost's capability to show nonlinear relationships in heterogeneous building data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

# Data Sources and Feature Engineering

In this chapter, the used data sources along with the feature engineering pipeline based on the aforementioned are described. The feature engineering pipeline was developed for the creation of a consistent dataset for archetype clustering and for the training of the ML energy model.

## 4.1 Datasets

This project is entirely based on preexisting datasets for which the availability and connectivity of the relevant datasets were reviewed beforehand. Creating a common data basis that combines functional, morphological and energy-related information at building level was the aim for this step. All building-related data was obtained from the SmartQ+ project [BSR+25], while the cadastral parcel data was obtained from the NOE Geoshop [Lan25]. For this project, the following datasets were used:

- **Building attributes & geometries:** Functional and morphological building features and multipolygon building footprints (2023)
- **Building energy information:** simulated aggregated energy information on building level (2023)
- **Building energy profiles:** simulated hourly energy profile on building level (2023)
- **Parcel geometries (DKM parcels):** Cadastral parcel polygons (2025)

In the context of this thesis, the datasets play different roles. The building geometries make up the spatial basis of this analysis and the SmartQ-building data mainly delivers



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

functional and morphological attributes of the buildings (see Table 4.1). The energy-related data provides descriptive and target variables. The property geometries enable the derivation of additional parcel-related parameters.

Table 4.1: SmartQ+ building attributes

| Feature | Description | Type |
|---|---|---|
| id | Project-internal ID of the building | numerical |
| additional_information | Textual information about the building | text |
| monument_protection | Flag if the building is under monument protection | boolean |
| category_tu | Project-internal numerical code for building usage | numerical |
| grouped_category | Grouped categories after main building usages | text |
| count_floors | Number of floors of the building | numerical |
| height | Highest point of the building in meters | numerical |
| construction_period | Project-internal numerical code for building's construction period | numerical |
| construction_description | Textual description of the period span | text |
| heating_flag | Flag if the building is heated | boolean |
| heating_description | Textual description of the heating system | text |
| main_building | Flag if the building is a main or side building | boolean |
| surface | Surface of the building's footprint | numeric |

All datasets were imported into a PostgreSQL/PostGIS-database to ensure uniform processing. Through that, geospatial operations, for example area calculations, intersection formations and neighborhood analyses, could be carried out reproducibly and consistently. The merging of the building geometries with the building attributes and the simulation-based energy indications of the SmartQ+ project was easily possible using the project-internal ID of the buildings. In this thesis, the building was the central analysis and data object, so all features were either merged at building level or derived from this level. Functional and energy-related information was directly available at building level; further attributes were added by linking them to the building geometry, the belonging properties and the immediate neighborhood.

Special attention was paid to the challenges of data quality, which are derived from the geometric inconsistencies between the datasets. These show themselves, for example, through an overlap in certain places or through not entirely being within a single cadastral parcel. While overlaps are mainly relevant for the determination of attachment-attributes, parcel exceedance/parcel boundary crossing are mostly relevant in area-based intersections and overlay analyses. Both challenges were explicitly addressed in the further course of data preparation. The geometry data of the buildings and cadastral parcels were harmonized into a uniform coordinate reference system and checked for geometric validity. As a common `CRS EPSG:31256 (MGI/Austria GK East)` was



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

used to enable geospatial operations in metric units.

The SmartQ-energy profiles of the buildings originally were available in hourly resolution at building level and consisted of about 36 million entries. To reduce the amount of data and to enable a more efficient processing in the database, the time series was converted into a daily representation for which every day saved the 24-hour values in the form of an array. Through that, the number of lines could be reduced to about 1.6 million while the hourly load structure for the further steps of analysis was preserved. In addition, this more compact visualization eases further processing in machine-learning pipelines because daily load profiles can be treated as inputs with a fixed length instead of having to be reconstructed from single hourly datasets.

## 4.2 Feature Engineering

In this section, the steps of the feature engineering pipeline are described; these have been used to prepare the initial data for the archetype-clustering and the training for the ML-energy model. The pipeline was implemented as a sequence of reproducible SQL scripts in PostgreSQL/PostGIS. Interim results were validated with tabular plausibility checks in the database in addition to visual validations in QGIS.

The starting point of the pipeline was the merging of the building geometries with the building data and simulation-based energy data, followed by data harmonization. Through this mapping the building objects were assigned with energy data in addition to relevant features such as usage category, construction period as well as initial morphological features such as building height and number of floors. Based on this, further building-related, parcel-related and neighborhood-related derivations were made.

In the first step of the pipeline, the initial dataset was reduced to the buildings that were determined to be energy-relevant for this analysis. In the research project SmartQ+ 4,254 buildings were simulated, including ancillary buildings. Because this thesis focuses on the derivation of energy indicators based on functional, morphological and spatial-structural characteristics and because the majority of ancillary buildings do not have space-conditioning, the following selection rule for energy-relevant buildings was applied: A building was included in further analysis if it was either tagged as a main building or had space conditioning. Through that, unheated ancillary buildings, for example, garages, were excluded, since their energy consumption is mostly process-related and can not be explained through morphological and spatial-structural building features. This filtering step improves the interpretability of results and aligns the dataset with the intended thesis scope. This reduced the number of buildings in the dataset to 3,032.

The filtered building set was extended with further morphological attributes in the next step of the pipeline; these attributes were derived from building footprints and the number of floors (see Table 4.2). These features provide a compact description of the structural form, size and usable space of the buildings, thus forming a basis for their morphological characterization. Among other things, they include floor area, perimeter, aspect ratio and



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

compactness degree, which describes the efficiency of the building shape. In addition, the gross floor area (GFA) was calculated, which equals the building's floor area multiplied by the number of floors. This was designated as a proxy, because here it was assumed that the area on the upper floors is equal to the area of the building footprint. The same approximation logic was used for calculating volume; here, the floor area was multiplied by the maximum height of the building. In the further course the variable `b_area` was used for the floor area of a building, because it was directly calculated from the building footprints, which all further geometric operations are based on. Because of that all morphological features are based on a consistent spatial data foundation. The variable `surface`, which is included in the simulation data, was not adopted as a primary variable, as the comparison of both variables showed minor deviations. Since the maximum relative deviation was only at 0.003%, no further measures were taken because no content distortions were expected.

Table 4.2: Morphological building features

| Feature | Description | Unit |
|---|---|---|
| b_area | Area of the building footprint | $m^2$ |
| b_perimeter | Perimeter of the building footprint | $m$ |
| b_obb | Oriented bounding box for robust aspect ratio | % |
| b_compactness | Compactness index of the building (Measure of form efficiency $(4\pi A)/P^2$ | |
| b_aspect_ratio | Ratio of building's length and width (Based on b_obb since more robust) | % |
| b_perim_area_ratio | Perimeter per area (exposure proxy) | % |
| gfa_proxy | Gross floor area (building area x number of floors) | $m^2$ |
| volume_proxy | Building area x building height | $m^3$ |

In addition to form-related morphological features, attachment features derived from the building footprints were used to distinguish between detached and attached buildings and to quantitatively assess the degree of structural connection. For this purpose, possible neighbor pairs for every building were formed and the length of shared wall sections was determined. In general, this was done using the intersection length of the touching building boundaries. Because the building geometry was sometimes inaccurate, special cases had to be considered. For example, there were overlaps in geometries or through slight shifts actually adjacent buildings were modeled as separate. In the case of overlaps, the shared wall length was not directly derived from the overlapping area, as shown in Figure 4.1, because this would create artificial inner edges and overestimate the degree of attachment. Instead, each building's edge length within the other building's polygon was calculated; the smaller value was then used as an estimate of the shared wall length. In addition, a tolerance of $0.20m$ was introduced for taking positional shifts of nearly touching building footprints into account. Very short intersection areas of under $0.05m$ were interpreted as geometric noise and not determined to be a real shared wall.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

(a) Selected building has an overlap with its neighboring building to the left.

(b) The highlighted red area indicates the intersection between the two building footprints.

Figure 4.1: Example of overlapping building geometries.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

The attachment attributes were derived from the determined shared wall sections per building (see Table 4.3). These features supplement the building parameters with spatial-topological information and describe if and to what extent a building is structurally connected to neighboring buildings.

Table 4.3: Building attachment features

| Feature | Description | Type |
|---|---|---|
| is_attached | Flag if the building is attached or detached | boolean |
| touching_buildings_count | Number of buildings that are attached | numeric |
| shared_wall_length | Length of shared walls of that building (in $m$) | numeric |
| attachment_ratio | Ratio of the building's walls that are attached to another building | numeric |

The spatial distribution of the `attachment_ratio`, which measures the degree of structural integration on a continuous scale, is shown in Figure 4.2. Buildings without shared wall sections are shown in blue; increasing proportions of shared wall length are coded in light to dark shades of red. The figure illustrates the spatial differentiation between detached buildings in sparsely developed areas and more densely developed areas.

Table 4.4 shows the features that show a relation between buildings and parcels. Parcel-specific features are used in this thesis for two reasons: Firstly, cadastral parcels are legally and functionally relevant spatial units used in municipal planning and build a



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

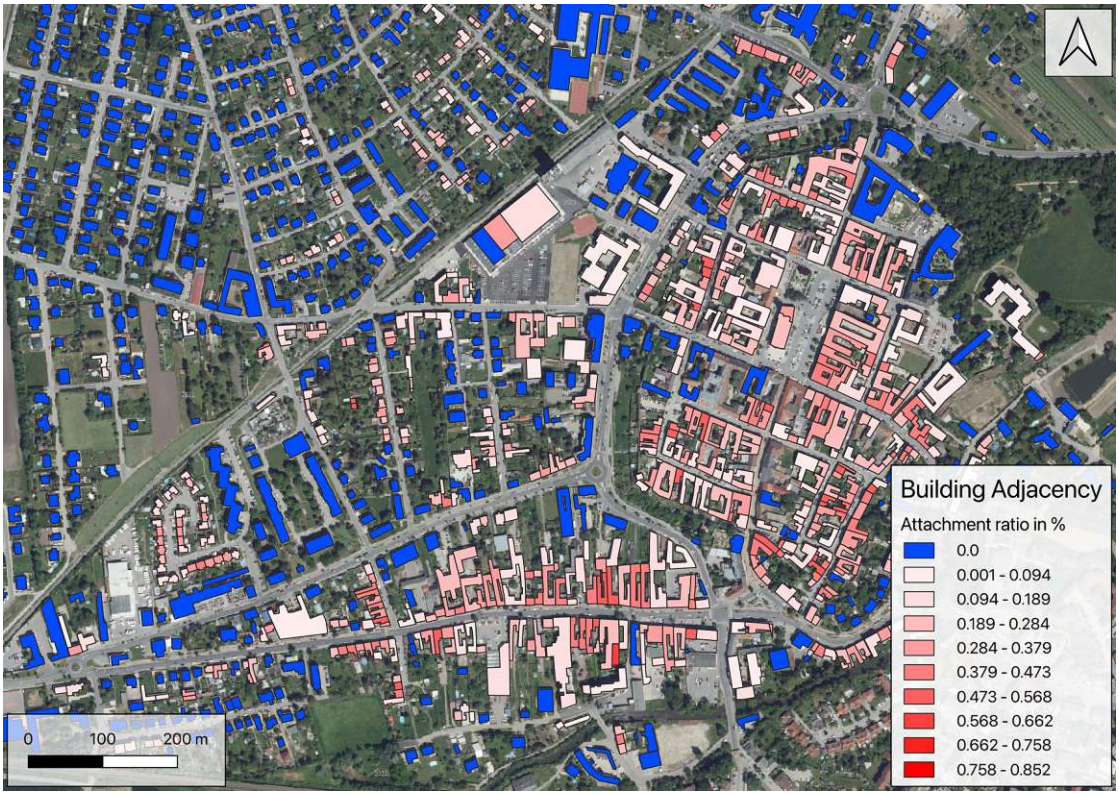


Figure 4.2: Spatial distribution of the `attachment_ratio`. Blue indicates detached buildings; red tones indicate increasing shares of shared wall length.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

context that goes beyond the building area itself. Secondly, density measurements derived from parcels measure the settlement intensity and serve as informative predictors for building typology and energy-related characteristics.

For the assignment between buildings and parcels, the intersections were first calculated. The main challenge here is that the building is not always entirely located within the parcel. The reason for this could be a geometric inaccuracy of the polygons caused by discrepancies between the datasets, through which a building on the boundary of the parcel overlaps with a different parcel. To make the link between building and parcel reproducible, the intersections between each building geometry and all overlapping parcel geometries were calculated. To filter out sliver intersections caused by data discrepancies, all intersections with an area under $5m^2$ were discarded. For the remaining pairs (building-parcel) the `coverage_ratio` was calculated which was used for the determination of the best pairing based on the highest coverage ratio.

The distribution of the resulting coverage ratio values is shown in Figure 4.3. As most



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.4: Parcel related features

| Feature | Description | Unit |
|---|---|---|
| parcel_area | Area of the parcel footprint | $m^2$ |
| intersect_area | Area of the building-parcel intersection | $m^2$ |
| coverage_ratio | Ratio of the building's area and the intersection area | % |
| bcr | Building coverage ratio (building area / parcel area) | % |
| far | Floor area ratio (building gfa / parcel area) | % |
| is_single_building_parcel_high | Flag for showing if only high conf. building on parcel | boolean |
| parcel_bcr_all | Aggregated BCR of all building intersections on that parcel | % |
| parcel_far_all | Aggregated FAR of all building intersections on that parcel | % |

values are strongly concentrated around 1.0, the parcel assignment quality can be regarded as overall high. To make the quality of parcel association explicit and to distinguish between robust, plausible and uncertain building–parcel matches, a categorical confidence indicator `confidence_flag` was assigned based on coverage ratio:

- **high**: coverage ratio $\geq 0.80$
- **medium**: $0.65 \leq$ coverage ratio $< 0.80$
- **low**: coverage ratio $< 0.65$

To not distort the values of indicators at parcel-level such as the floor area ratio (FAR) and building coverage area (BCR), a robust assignment between building and the most relevant parcel is necessary. Both values are common urban-planning indicators that show the development intensity of a parcel. This is why they are suitable to determine whether a building is located in a detached or a compact context. To retain the most realistic building density, the following procedure was used to calculate FAR and BCR: for each parcel, all building intersections that have not yet been discarded as slivers were aggregated. This prevents only the ratio between the best pairing from being calculated, which would result in the loss of intersections larger than $5m^2$ and a reduction in density. For additional parcel context, the flag `is_single_building_parcel_high` was added to indicate whether a parcel only contains one building with high confidence. This should improve interpretability, since parcels with only one building FAR/BCR mainly reflect the affected building rather than the broader parcel context.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

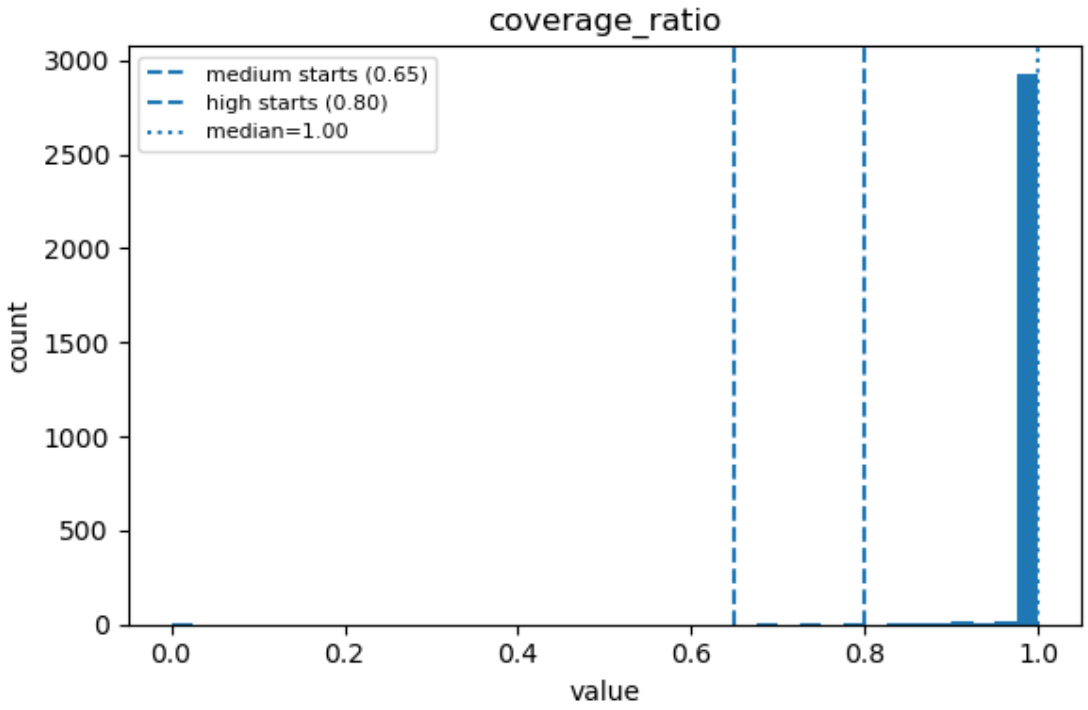


Figure 4.3: Distribution of coverage ratios for building–parcel intersections and corresponding confidence thresholds.

A small number of buildings, precisely nine objects, showed no parcel overlap and were thus not assigned. These cases were manually reviewed in QGIS and were found to be located near the municipal boundary. They remain in the building dataset but require special treatment in the calculation and later application of parcel-based metrics.

To depict the context of the neighborhood structure, the spatial-topological structure of buildings within a 100-meter radius was considered (see Table 4.5). First, a point-based representation of the geometry of every building was calculated. If this point is located within the boundaries of the circle, which is spanned by a different building, then the first building is considered to be in the neighborhood of the latter building. The circle-approach was preferred to a possible grid-approach, since the chosen boundary of $100m$ has the same distance in every direction and the corners are not the furthest from the center, as it would be the case in a square. The construction period of a building is a categorical value, but each of the categories has an associated numerical value in the range of 1 to 9. This value was used to represent the median of the construction period.

These variables represent the urban structure of the wider surroundings, adding morphology and spatial structure at building and parcel level. Among other things, it records whether a building is embedded in a dense residential structure, a heterogeneous settlement structure or a mixed-use context.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 4.5: Neighborhood structure ($100m$ radius)

| Feature | Description | Unit |
|---|---|---|
| nb_count_100m | Number of buildings in the 100m neighborhood | count |
| nb_median_b_area_100m | Median footprint area of neighboring buildings | $m^2$ |
| nb_median_gfa_100m | Median GFA proxy of neighboring buildings | $m^2$ |
| nb_median_construction_period_100m | Median construction period of neighboring buildings | |
| nb_dom_share_construction_period_100m | Dominant share of construction period in the neighborhood | % |
| nb_share_residential_100m | Share of residential buildings in the neighborhood | % |
| nearest_building_distance_100m | Distance to the nearest neighboring building | $m$ |
| median_distance_to_3_nearest_100m | Median distance to the 3 nearest neighboring buildings | $m$ |
| median_distance_to_5_nearest_100m | Median distance to the 5 nearest neighboring buildings | $m$ |
| nb_share_attached_100m | Share of attached buildings in the neighborhood | % |
| nb_median_attachment_ratio_100m | Median attachment ratio of neighboring buildings | % |

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

# Derivation of Building Archetypes

This chapter describes how the archetypal building structures were derived from the feature space constructed in Chapter 4. The aim of this step was to convert the heterogeneity of the building stock into a reduced typology.

The data-driven and domain-guided derivation of the archetypes aims to map the structural and functional diversity of buildings in a form that can be used as a robust input for the training of the energy model (see Chapter 6). Thus, the archetypes form a bridge between the derived building features and the ML-based energy analyses. At the same time, the archetypes should lead to structurally understandable building groups and provide a basis for the technical validation through experts.

The chapter shows the complete process for the archetype derivation. First, the concept and aim of the archetype approach are introduced. Next, the selection of variables and partitioning of the buildings are described. Then the clustering workflow, starting with the baseline clustering to the final two-stage weighted clustering, is documented. The resulting clusters are interpreted, mapped to archetype labels and lastly validated with expert annotations and stability checks.

## 5.1 Aim and Conceptual Approach

In Section 2.2, the concept of building archetypes was already introduced as a possibility to group buildings with a similar structure and function. Building on that, the approach is realized in this chapter for the test community of Bruck an der Leitha. The aim is not only to view the heterogeneous building stock as a collection of individual building characteristics, but to transform it into a reduced and, at the same time, interpretable building typology. This typology, which consists of identifiable archetypes, should help to reveal recurring patterns and combinations of characteristics within the building stock [PWK19]. The thought behind it is that buildings do not only differ based on



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

single attributes such as height or use case, but rather through the interplay of multiple factors such as the functional, morphological and spatial characteristics [ZGP23]. For example, small detached residential buildings such as single-family homes, or large-scale halls are identified based on these characteristics. The reduction of complexity is not due to the reduced number of buildings considered, but in the grouping of similar building structures into interpretable groups.

The created archetypes thus form a layer between the individual building characteristics and the subsequent energy-related analysis [SW24]. Individual buildings remain the observed units, while multiple attributes are considered jointly to group similar configurations at the newly introduced level of typical building structures [TKO+18]. In the context of the subsequent ML analysis, this is relevant because structural combinations of features do not need to be learned implicitly by the model, but are already described beforehand in the building typology in explicit form.

The archetypes used in this thesis are not to be understood as a fully predefined typology but they should rather describe locally derived groups of buildings, which, in this context, exhibit similar functional, morphological and spatial-structural characteristics [PWK19]. They differ from pure usage categories because, apart from usage, the characteristics such as building structure, architectural integration and neighborhood context are used.

Existing building typologies like TABULA/EPISCOPE serve as an important reference for the description of typical residential buildings [LSD16, Aus15]. The aforementioned could not directly be applied. On the one hand, this is because these typologies only cover residential buildings, but in this thesis, commercial and industrial buildings as well as social infrastructure are also considered. On the other hand, for the application in a rural community, a finer, locally adapted spatial-morphological differentiation is necessary than is provided by the coarse residential building typology. Such a local adaptation is essential because non-representative archetypes of the examined building stock may result in modeling errors [DJLS20].

For creating a local typology for the municipality Bruck an der Leitha, it is essential that the building stock is meaningfully and comprehensibly represented in the archetype catalog. To achieve this goal, the derivation of the building archetypes was done as a data-drive and domain-guided clustering process involving functional interpretation. In the clustering process, structures were first identified in the feature space based on the existing data. The selection and naming of the final archetypes were carried out considering technical plausibility and with regard to their intended use.

The category of residential buildings, which makes up the majority of the building stock, is especially diverse. The relatively large number of buildings and the different building designs give way to a more detailed classification. Examples include detached single-family homes, row houses and apartment blocks. Due to the limited number of buildings in the other usage groups, however, the distinction between large-scale and small-scale building structures was the primary focus.

The selection of the final clustering solution was not only based on the best clustering



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

metrics but the decisive factor was that the resulting clusters could be meaningfully mapped into the archetype catalog [Hen15]. At the same time, the clusters had to be robust enough so that the derived archetypes could serve as a basis for ML-based analysis in the further course of this thesis; this led to the requirement of both domain-specifically interpretable and statistically plausible clustering results. Based on the aforementioned, the question for the next step was which features of the feature space are best suited to represent such building structures.

## 5.2 Feature Selection and Partitioning

The starting point for archetype formation was not the complete set of features derived in Section 4.2, but rather a specifically selected subset of variables which were expected to contribute directly to the differentiation of typical building structures. The selection focused on features that describe the buildings' form-related characteristics, their functional context, their relationship to adjacent buildings, and their immediate spatial surroundings. Attention was paid to keeping the feature set as meaningful as possible by avoiding semantic overlap between features [ATL18]. In unsupervised clustering, redundant features can hide the intended grouping structure and reduce the interpretability of the resulting clusters [DB04]. This selection process reflects the assumption that building archetypes cannot be described by individual attributes alone, but emerge as compact representations of recurring patterns within the selected feature space.

To systematically select features relevant to archetype clustering, the variables were assigned to higher-level feature dimensions (see Table 5.1). This step made it easier to assess which attributes within each grouping could be expected to make an independent contribution to the differentiation of building structures, while also making the overall selection process more transparent. The dimensions essentially correspond to the labels of the feature tables derived in Section 4.2, with an additional dimension for the functional context.

The selected feature set includes functional, morphological, attachment-based, parcel-based and neighborhood-based features. Morphological building characteristics such as footprint area and height describe the size and shape of the building's body. Attachment-based variables describe whether and how strongly a building is physically connected to adjacent buildings. Parcel-based features provide information on the spatial relation between buildings and their parcels, while neighborhood variables describe the surrounding settlement structure within a local context.

The attribute `category_tu`, which contains more detailed usage information of the buildings (see Table 4.1), was purposefully excluded from the clustering input. Instead, it was retained for subsequent cluster interpretation and plausibility checks. This decision was made to avoid letting detailed, project-specific usage assignments predetermine the clustering results; this aligns with the thesis' aim to derive archetypes primarily from freely accessible data.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.1: Archetype clustering features.

| Feature | Type | Feature dimension |
|---|---|---|
| grouped_category | categorical | partitioning |
| construction_period | categorical | functional |
| b_area | numerical | morphology |
| height | numerical | morphology |
| gfa_proxy | numerical | morphology |
| b_compactness | numerical | morphology |
| b_aspect_ratio | numerical | morphology |
| b_perim_area_ratio | numerical | morphology |
| shared_wall_length | numerical | attachment |
| touching_building_count | numerical | attachment |
| attachment_ratio | numerical | attachment |
| dom_parcel_area | numerical | parcel |
| parcel_bcr_all | numerical | parcel |
| parcel_far_all | numerical | parcel |
| is_single_building_parcel_high | binary | parcel |
| nearest_building_distance_100m | numerical | neighborhood |
| median_distance_to_5_nearest_100m | numerical | neighborhood |
| nb_count_100m | numerical | neighborhood |
| nb_share_attached_100m | numerical | neighborhood |
| nb_median_attachment_ratio_100m | numerical | neighborhood |
| nb_share_residential_100m | numerical | neighborhood |

At the same time, coarse usage information was still required to avoid comparing buildings belonging to fundamentally different functional contexts. For this reason, the clustering was partitioned by `grouped_category` to identify more specific structural differences. This procedure follows the hybrid-clustering approach of Borges et al., who divided heterogeneous building stocks by usage into broader segments before a more detailed clustering is applied to each subset [BTBPR22].

This partitioning step was particularly important because the building stock is strongly dominated by residential buildings. As shown in Table 5.2, the `Residential` category contains 2,625 of the 3,032 buildings and thus is 86.58% of the analyzed building stock. Because segmentation choices can affect the representativeness of building archetypes [AM24], clustering all buildings together in one step would most likely have resulted in a structure shaped mainly by the large residential group and smaller usage groups such as social infrastructure, commercial buildings or industrial buildings would have been underrepresented in the clustering solution.

The partitioning strategy therefore was done for two different purposes. Firstly, it reduced the risk that dominant usage groups would obscure smaller but relevant building types. Secondly, it allowed the clustering process to focus on different structural distinctions



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.2: Distribution of buildings across the functional partitions used for clustering.

| Partition | Functional group | Buildings | Share |
|---|---|---|---|
| P01 | Commercial | 175 | 5.77% |
| P02 | Industry | 155 | 5.11% |
| P03 | Other | 24 | 0.79% |
| P04 | Social infrastructure | 53 | 1.75% |
| P05 | Residential | 2,625 | 86.58% |

Table 5.3: Caption

within each functional context. For example, the relevant differences within the residential partition are mainly related to attachment, building size, parcel structure and settlement density. In contrast, the distinction between small-scale commercial buildings and large halls is more strongly related to building footprint size, building volume and operational surroundings.

Thus, the feature selection and partitioning strategy created the basis for a subsequent clustering workflow. Instead of searching for a global cluster structure across the entire building stock, the process first separated the data into broad functional contexts and then identified structurally meaningful groups within these partitions. This makes the resulting clusters easier to interpret and provides a more stable basis for mapping them to the final archetypes.

## 5.3 Clustering Workflow

After a final feature selection in Section 5.2, the next step was in the conversion of partitioned buildings and their associated variables into meaningful building clusters. The clustering process thus serves the central methodological step of transforming the prepared dataset, as described in Chapter 4, into a semantically interpretable set of building clusters, which then can be transferred into the archetype catalog created in this thesis.

For the first step, a simple approach was chosen to assess the applicability of the chosen feature set for the use case and to serve as a reference for future results. Building on that, refinements in the approach were made step by step in order to yield results which are applicable to planning issues, energy prognoses as well as the simultaneously created archetype catalog. The following sections describe the preprocessing and the distance calculations, the implementation of the baseline approach and lastly the two-stage clustering approach derived from the aforementioned.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

### 5.3.1 Feature Preprocessing and Distance Calculation

The clustering does not directly work with raw data for the grouping of buildings, but compares buildings based on pairs of differences between single buildings. Before the actual formation of the clusters the chosen features had to be prepared in a way so the differences are incorporated consistently and meaningfully context-wise [Gow71]. This step was necessary because the used variables describe different aspects of buildings and differ regarding value range and distribution. Using this approach, the distance measurement is not characterized mainly by distortions from different scales, individual outliers as well as skewed distributions, but relevant differences between buildings are shown [Hen15].

#### Selection of Distance Measurement

To measure distances the Gower dissimilarity was chosen, which was introduced in Section 2.2. Initially, this approach was chosen because the candidate feature set included numerical, categorical and binary features. The final feature set was predominantly numerical, except for one binary flag and the construction period, which was retained as a categorical variable, because it was considered relevant for the distinction of building typologies. Furthermore, the numerical features describe different structural dimensions, such as size, distances, density as well as form factors and proportions. The Suitability of using Gower dissimilarity with hierarchical clustering is supported by Meshkinkiya, who observed that this combination leads to better-separated building clusters than k-prototypes for mixed building data [MH26].

A Euclidean distance calculation could also have been applied after scaling the numerical variables and coding the binary feature accordingly. For this use case, Gower's distance was the suitable choice because it does not calculate differences as distances between numerical feature vectors, but computes partial distances on the feature level. Subsequently, these were merged for valid features as a weighted mean into a total distance [Gow71]. This has the advantage that binary and categorical values are used in their native form and missing values, for example the parcel data for certain buildings, can be excluded from the pairwise calculation. Therefore, no values had to be imputed and no artificial similarity had to be created between buildings. An additional criterion for the distance calculation was that optional weighting for features should be supported so that in later iterations of the cluster formation the significance of single features can be targeted.

#### Feature Preprocessing

To ensure a consistent and content-wise meaningful distance calculation of the pairwise dissimilarities, the used features were first processed in sequential steps. The most important preprocessing decisions are summarized in Table 5.4 and are explained in more detail in the following.

At the beginning, the chosen variables from Table 5.1 were categorized based on their data type and saved in separate arrays; the final feature set for the distance calculation



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.4: Preprocessing decisions before distance calculation.

| Issue | Treatment | Rationale |
|---|---|---|
| Mixed feature types | Type-specific encoding of numeric, binary and categorical variables | Preserves the intended feature interpretation |
| Invalid or inconsistent values | Type-specific validation | Excludes invalid values from distance calculation |
| Neighborhood missingness | Imputed to represent spatial isolation | Treats isolation as meaningful information |
| Missing parcel information | Retained as missing | Avoids artificial parcel context |
| Heavy-tailed numeric variables | Selective `log1p` transformation | Reduces pronounced upper tails |
| Numeric value ranges | Fixed bounds or robust 5–95% quantile bounds with winsorization | Provides feature-specific ranges and limits the influence of extreme values |

includes a binary, a categorical and 18 numerical variables. The binary feature was coded consistently with the numerical indicator variables `0` or `1`, and the numerical variables were uniformly converted into the floating-point format `Float32`. The codes of the variable `construction_period` were saved as `String` and not as metric values but as labels for different construction periods.

**Feature Type Definition and Validation** Following, all variables were checked regarding missing, non-interpretable or content-related invalid values. The binary variable `is_single_building_parcel_high`, for example, was checked whether the values were only `0` and `1`. In the strings of the categorical variable `construction_period`, first the preceding and following spaces were discarded and afterward checked, whether the existing values corresponded to one of the nine permitted categories. Other values than these were also treated as missing. This approach prevented spaces or invalid values from being considered as distinct construction periods in the distance calculation.

For the numerical values, it was first checked how many values per feature were missing or did not represent a numerically interpretable value; the latter was also treated as a missing value. Depending on the content definition of the individual features, plausibility checks were administered after. The features, which by definition can not be assigned a negative value, were checked for that. The values of variables, which represent a ratio or proportion with a specified range of values of `[0,1]`, was checked, whether they were really within this interval. In both steps invalid values were marked as missing.

**Handling Missing Values** In the course of the validation, for every feature the number of initial and the number of additional missing values from the plausibility checks were documented. This ensured the possibility to check whether the missing data already existed in the initial dataset or only resulted from the validation of invalid values. As



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

shown in Table 5.5, all missing values already existed before the validation. In the following, the usage of missing values is different depending on the reason why it is missing. On the one hand, some values of parcel data are missing since they could not be assigned because of missing information. On the other hand, the neighboring features were not defined for the buildings that had no buildings in the examined $100m$ radius.

Table 5.5: Identified missing values and subsequent handling.

| Feature group | Affected features | Missing values | Subsequent handling |
|---|---|---|---|
| Neighborhood context | `nearest_building_distance_100m`, `median_distance_to_5_nearest_100m`, `nb_share_attached_100m`, `nb_median_attachment_ratio_100m`, `nb_share_residential_100m` | 25 per feature | Encoded as an expression of spatial isolation |
| Parcel context | `dom_parcel_area`, `parcel_bcr_all`, `parcel_far_all` | 9 per feature | Retained as missing and excluded pairwise from the distance calculation |

In a total of 25 buildings, the feature `nb_count_100m` has the value `0`, which is why no values for the features, which describe the neighborhood context, are given. This means that in a 100-meter radius around the building, there is no other building. This information was thus not treated as an unknown or undefined value, but as the consequence of spatial isolation of the buildings. Accordingly, the values were imputed in a way so the spatial isolation is depicted accurately and so it could be considered for the distance calculation. For `median_distance_to_5_nearest_100m` and `nearest_building_distance_100m` the value was set the upper limit `100` of the search radius; a value of `0` would not be suitable content-wise, since it would mean a building in the immediate spatial proximity. The undefined proportional or relative variables `nb_share_attached_100m`, `nb_median_attachment_ratio_100m` and `nb_share_residential_100m` were assigned with the value `0`. In both cases the imputed values do not have the same meaning as the other regular values. While in an existing neighborhood context they can describe a far distance to the nearest building or an observed proportion of 0%, here in combination with `nb_count_100m = 0` a strongly isolated building is indicated.

Missing parcel data was handled differently because for a total of nine buildings no assignment to a parcel was possible. This means that the calculation for `dom_parcel_area`, `parcel_bcr_all` and `parcel_far_all` was not possible. These values could not reliably be derived from the other building attributes, such as building area or the number of floors, which is why they were retained as missing. For the calculation of Gower's distance, the missing values were excluded for the respective building pairs.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

**Distribution Analysis and Logarithmic Transformation** After the validation of the feature set and the treatment of missing values the distribution of the numerical features was examined. This step had the aim to identify the variable which exhibit heavily skewed or heavy-tailed distributions. These are relevant because major outliers could dominate the partial distances in Gower's distance calculation; differences in the typical value range, such as those around the median, are relatively reduced and lose significance [Hen15]. This phenomenon occurs especially in scale-type features such as building or parcel areas, since there are only very few large objects that extend the value range significantly.

Table 5.6: Distribution indicators and transformation decisions for selected features.

| Feature | Skewness | Median | $P_{99}$ | $P_{99}$/Median | Transform | Indication |
|---|---|---|---|---|---|---|
| `b_area` | 12.62 | 150.52 | 2,421.26 | 16.09 | Yes | Large median–$P_{99}$ gap |
| `gfa_proxy` | 17.72 | 428.95 | 9,619.89 | 22.43 | Yes | Pronounced upper tail |
| `dom_parcel_area` | 15.03 | 676.52 | 91,999.99 | 135.99 | Yes | Extreme median–$P_{99}$ gap |
| `shared_wall_length` | 2.60 | 0.00 | 55.45 | – | Yes | 56.66 % zeros; skewed positive tail |
| `parcel_far_all` | 3.92 | 0.73 | 4.00 | 5.46 | Yes | Unbounded and right-skewed |
| `b_aspect_ratio` | 4.06 | 1.15 | 3.61 | 3.15 | No | Limited upper range |
| `height` | 2.65 | 9.00 | 18.00 | 2.00 | No | Moderate median–$P_{99}$ gap |
| `b_perim_area_ratio` | 2.57 | 0.36 | 0.82 | 2.26 | No | Moderate upper-tail deviation |
| `nearest_building_distance_100m` | 2.32 | 17.44 | 68.23 | 3.91 | No | Range bounded (100 m radius) |
| `nb_count_100m` | 1.67 | 29.00 | 106.00 | 3.66 | No | Low skewness |

To evaluate whether such a distribution exists for a feature, the following values were calculated: median, mean, minimum and maximum, 90%-, 95%- and 99%-quantiles as well as the distribution skewness. Additionally, histograms were created to enable a visual inspection. The decision whether a feature should be considered for a logarithmic transformation was based on a combination of distribution form, quantile intervals and the value range of the features. For the value range it was important whether the values were presented in a fixed interval, for example $[0, 100]$, or in an open ended value range [Hen15].

As shown in Table 5.6, the features `b_area`, `gfa_proxy` and `dom_parcel_area` could be identified as strongly skewed to the right and heavy-tailed. This becomes evident



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

because high skewness values are present and the upper quantiles for all three variables are significantly well above the median. Additionally, the two features `parcel_far_all` and `shared_wall_length` were incorporated in the selection of variables to be transformed, because for both of these features, similar observations were made in a milder form. For `shared_wall_length`, because of the content meaning of the feature and the structure of the data, the median is at `0`. This is why the relation of median to the upper quantiles is not meaningful, but the values in the histogram (see lower left panel of Figure 5.1) show a clearly pronounced right-hand outlier in the distribution.

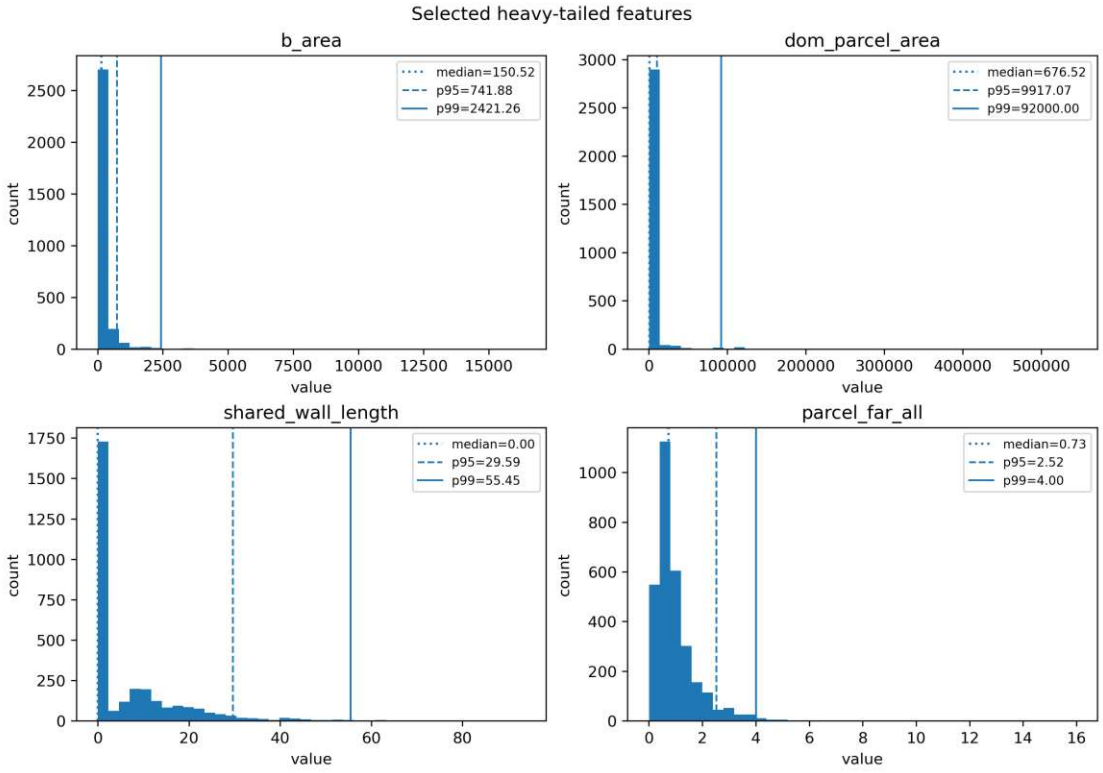


Figure 5.1: Distributions of four examined numerical features with pronounced upper tails, including the median and the 95th and 99th percentiles.

For the other skewed variables, for example `height` and `b_aspect_ratio`, a transformation was refrained from. This had various reasons: Partly the deviation of the upper quantiles from the median was not pronounced enough or the value range was limited through the structure of the features. The selection was done deliberately restrictive, so only strongly pronounced distribution outliers were damped and possibly relevant differences in the rest of the feature set were not unnecessarily compromised.

The five chosen features were transformed with *log1p* in the next step. For this, for every value $x$ the function $log(1 + x)$ was applied. Compared to the standard logarithm, $log(1 + x)$ is also defined for the value `0`, which was particularly relevant for the fea-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

ture `shared_wall_length` due to the high proportion of detached buildings. The transformation keeps the order of the values, while the distance in the upper values is more strongly compromised than the lower values. An example of the effects of the transformation is shown in Figure 5.2 for the feature `gfa_proxy`. As the figure shows, the transformation results in a substantially more compact distribution and reduces the pronounced right-skewness of the raw values.

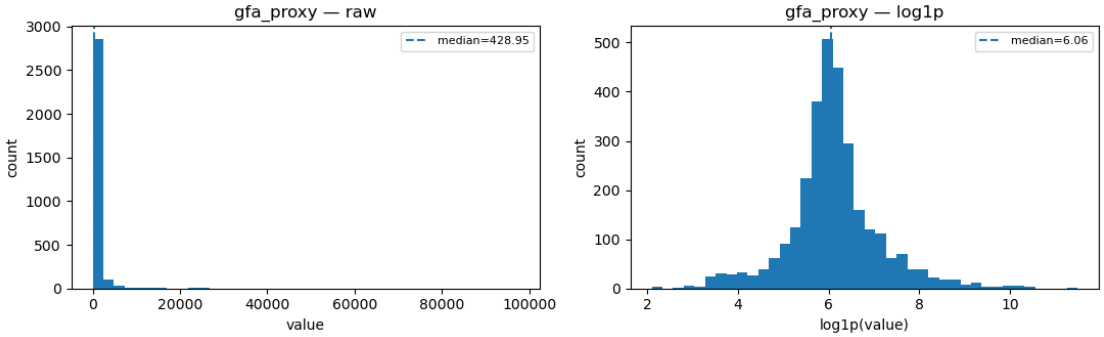


Figure 5.2: Distribution of `gfa_proxy` before and after $log1p$ transformation, illustrating the reduction of the variable's right-skewness.

**Numeric Value Ranges and Winsorization** After the transformation a valid value range was set for every numerical feature. These ranges served as preparation for scaling information for the later calculation of the numerical subdistances in the Gower dissimilarity. The definition was done feature-specifically, because the used variables showed different units, scales of magnitude and boundaries. For features with natural scales of magnitude, for example $[0, 1]$ for proportional values or $[0, 100]$ for the distance features in a $100m$ radius, the natural boundaries were retained. For the features without a fixed boundary, robust boundary values were determined with the 5-% and 95%-quantiles; for logarithmically transformed features, this was done on a transformed scale. Values outside the previously determined interval were then, in a winsorization sense, limited to the lower or upper interval boundary. Through these measures it was ensured that the subsequent distance calculation is based on comparable value ranges and extreme outliers do not dominate the distance calculation.

### Calculation of Gower's Distances

The actual normalization of the numerical features only took place during the Gower distance calculation, whereby the association subdistances were calculated on the basis of the previously processed values. For a building $i$ and a numerical feature $k$ the value $z_{ik}$ can be expressed as

$$z_{ik} = \begin{cases} \text{clip}\left(\log(1 + x_{ik}), l_k, u_k\right), & \text{if feature} k \text{was log-transformed} \\ \text{clip}\left(x_{ik}, l_k, u_k\right), & \text{if feature} k \text{was not log-transformed} \end{cases} \tag{5.1}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

where $l_k$ and $u_k$ denote the previously determined lower or upper boundary of the respective feature. Thereafter, the subdistance of the numerical features between the two buildings $i$ and $j$ was calculated as a relative difference:

$$d_{ijk} = \frac{|z_{ik} - z_{jk}|}{R_k}, \qquad R_k = u_k - l_k \tag{5.2}$$

Numerical differences were thus not assessed based on their absolute order of magnitude, but relatively to the value range of the respective feature. The resulting feature-specific subdistances were then aggregated according to the general Gower's distance formula, introduced in Equation 2.1. Categorical features received a partial distance of `0` if the features matched and `1` when they differed. Binary features were considered analog using the absolute difference of the `0/1`-coded values. Missing values were excluded pairwise from the numerator and denominator using the availability indicator $\delta_{ijk}$.

The results of the distance calculations were symmetrical Gower distance matrices, which showed the pairwise dissimilarities between the buildings shown in the Table 5.2 partitions. This means that the distance was not calculated between buildings of different partitioning. The diagonal of every matrix equals 0 because a building does not have a distance to itself. These matrices serve as a basis for the subsequent hierarchical clustering.

### 5.3.2 Partition-wise Computed Baseline Clustering

The baseline approach described in this section was purposefully implemented as a simple clustering procedure, since the approach should serve as a reference for further developments. Based on the results of the process, one of the things that should be evaluated is to what extent the chosen feature set and the Gower distances described in Section 5.3.1 are suitable for the formation of functionally appropriate building groups. For this, the baseline clustering was carried out as flat clustering with all features selected for the final feature set in a single run. The cluster calculation was applied to the previously calculated distance matrices in the single partitions; this means that, per partition, its own calculation was carried out. Additionally, the use of feature weights was intended, but deliberately left out for the baseline iteration for the sake of simplicity.

#### Hierarchical Clustering

The actual clustering process was implemented as hierarchical clustering. This approach was already described in Section 2.2 and is suitable for this use case because the previously calculated Gower distances can be used directly as input [MH26]. Therefore, the different feature types did not have to be mapped in a joint Euclidean feature space. Since distance matrices were available in quadratic form, they contained a diagonal with self-distances that consisted exclusively of 0 values, along with each pairwise distance stored twice. This is why they were first transformed with `squareform` into a one-dimensional distance vector which includes only the definite pairwise distances. Through this step



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the Gower distances were brought to a representation, which was stripped of redundant pair distances. This representation complies with the needed input format for the next steps of the cluster calculation.

For the clustering an agglomerative approach was used, in which first every building represents its own cluster. Within a partition, these single clusters were then merged step by step. A linkage criterion decided how the distance between the already formed clusters was derived from the pairwise building distances. The SciPy function `linkage` created a linkage matrix from the one-dimensional distance representation, whereas the linkage matrix describes the merging steps. Every line stands for the unification of two clusters and also provides information of the distance for the merge and how many buildings the newly formed cluster contains. In one partition with $n$ buildings, the linkage matrix thus contains $n - 1$ merging steps and thus maps the entire cluster hierarchy in that partition.

The cluster hierarchy, as described by the linkage matrix, is visualized as a dendrogram. In such a hierarchical tree, the leaves on the lowest level show the single buildings while the branches depict the mergers of buildings or the already formed clusters. The height of the merge on the y-axis depicts the linkage distance at which the respective building or cluster was grouped. This distance is determined on the basis of the previously calculated pairwise Gower distance and the respectively used linkage method.

Which clusters were merged in the single steps and how the resulting hierarchy is structured thus depends on the linkage method. As possible linkage criterion, the methods `Single`, `Complete` and `Average Linkage` were considered. Single-linkage is based on the smallest distance between two clusters, through which a kind of chaining effect can happen in which structurally different groups can be merged via individual similar buildings. On the contrary, complete linkage uses the largest distance and thus encourages very compact clusters. Single atypical buildings can delay mergers, which could encourage very fragmented clusters.

Average linking includes all pairwise distances between two clusters, which means that the same and different building pairs have less of an influence. Thus, the choice of average as a linkage method was a reasonable compromise since for the formation of building groups a certain structural variation could be allowed.

The comparison of the two dendrograms, which are depicted in Figure 5.3 and Figure 5.4, show the influence of the chosen linkage method on the form of the clustering hierarchy. For average linkage, the final merge of the largest subgroups occurs at a linkage distance of about 0.34. Additionally, in the lower branches it becomes evident that certain groups of buildings were already merged at a short distance. On the contrary, in the cluster hierarchy, which was created with complete linkage, it becomes evident that most groups remain separated up to a longer distance. This also reflects in the final merge, which occurs at approximately 0.74.

Because the residential partitions with 2,625 were too large for a complete and meaningfully interpretable depiction, the following figures only show the topmost levels of the respective clustering hierarchies. Not bracketed values on the x-axis show the ID of



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

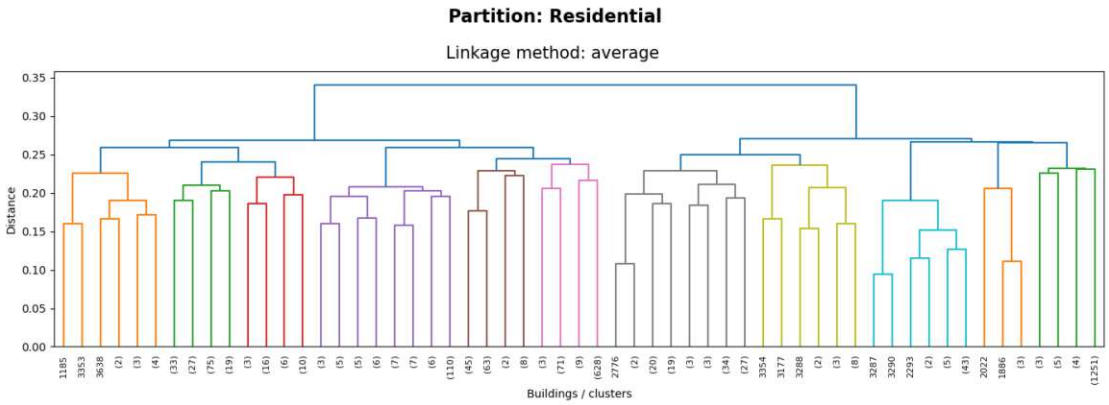


Figure 5.3: Truncated dendrogram of the hierarchical clustering for the residential partition, based on average linkage.

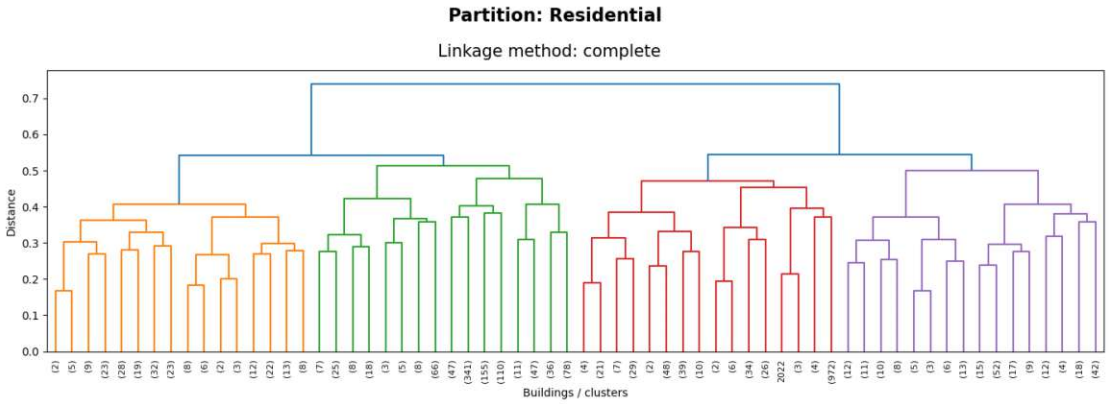


Figure 5.4: Truncated dendrogram of the hierarchical clustering for the residential partition, based on complete linkage.

individual buildings, bracketed values show the number of buildings in an already formed cluster for the respective branch.

In the scope of this work, the dendrograms were used for a visual comparison of the cluster hierarchies that have been created with the different linkage methods. This way, it was evaluable which method is best for the formation of interpretable building groups. The merging heights do not depict a comparable quality indicator because the distances between clusters are defined differently across approaches. This is why the assessment was mostly based on the structure of the resulting hierarchies, for example whether a too strong main cluster was created at an early stage or too compact partial groups occurred. The selection of a suitable candidate solution was primarily based on the quality and size criteria described in the next section.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

### Derivation and Evaluation of Candidate Solutions

Based on the created clustering hierarchies, whose upper levels are shown exemplarily for the residential partition in Figure 5.3, in the next step, multiple flat candidate solutions were derived. These emerge conceptually by a cut through the complete dendrogram. The higher this threshold, the less and thus coarser clusters emerge; similarly, a lower threshold leads to a more differentiated solution with more clusters. This results in the height of the cut determines how finely the buildings are divided into the clusters.

In the implementation, the derivation of the candidate solutions was carried out using the SciPy function `fcluster`. In detail, the criterion `maxclust` was used so the step was not operated via an explicit distance but via the defined maximum amount of clusters $k$. The upper limit of $k$ was calculated separately for each partition by taking the partition size, a maximum cluster amount of `max_k=12` and a desired minimum size of the clusters of `min_cluster_size_diag=6` into account. This was intended to prevent, especially in many small partitions, a too strong fragmentation into many small clusters. Subsequently, all permitted $k$ values for the respective partitions were tested and the resulting cluster labels were saved in a joint structure.

To assess the created candidates in the next step, for every combination of partition and $k$ several quality and diagnosis metrics were calculated. These describe different aspects of the candidates, such as cluster separation, compactness, cluster size structures and the unambiguity of building assignments. An overview of the used metrics is shown in Table 5.7.

Table 5.7: Quality cluster metrics used for evaluating baseline candidate solutions.

| Metric | Interpretation |
|---|---|
| `silhouette` | Mean separation of buildings from their assigned cluster and the nearest alternative cluster |
| `median_silhouette_sample` | Robust summary of all building-level silhouette values in a cluster solution |
| `neg_silhouette_share` | Share of buildings with negative silhouette values, indicating potentially problematic assignments |
| `mean_within_cluster_distance` | Mean pairwise distance within clusters, used as a compactness diagnostic |
| `min_cluster_size`, `median_cluster_size`, `max_cluster_size` | Size distribution of clusters within a candidate solution |
| `singleton_clusters`, `clusters_lt_6`, `clusters_lt_8` | Counts of very small clusters, indicating potential fragmentation |
| `largest_cluster_share` | Share of the partition assigned to the largest cluster, indicating a potentially coarse solution |
| `affinity_gap_12`, `low_gap_share` | Building-level assignment ambiguity and the share of buildings with a small affinity gap |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

For the assessment of cluster separation, the silhouette values of the possible solutions were examined first. For this, the silhouette coefficient was calculated, which was already introduced for Equation 2.2. It compares the average distance to the own cluster for every building, with the lowest to an alternative cluster. In general, it can be said that a high silhouette value is aimed for, because it indicates a clear distinction between the clusters [Rou87]. In reality, with real heterogeneous building data, only moderate values are expected. As a rough guide, the values of $> 0.5$ were interpreted as indicating a robust cluster structure, the range $0.25 - 0.5$ as indicating moderate robustness and values below as indicating low robustness [KR09]. Because this value depicts the discrimination globally, the median was calculated additionally, since it is more robust to especially well or poorly assigned buildings. Additionally, based on `neg_silhouette_share`, the share of the buildings with a negative silhouette value was assessed. This is the case when a building, on average, is closer to a different cluster than the assigned cluster [Rou87]. Such buildings can be interpreted as edge cases or as objects difficult to classify.

Furthermore, the compactness of the created cluster solution was examined. For this, Equation 2.3 was applied to calculate the mean pairwise distance of buildings within the same cluster. This metric provides a measure of the similarity of the assessed features within a cluster. This means that a low distance in the cluster, which means a homogeneous building group, is a good basis for the later mapping on building archetypes [Jai10].

In addition to the distance-based quality metrics, the clusters were assessed regarding their size dimensions. For this, the minimum and maximum in addition to the median cluster size was calculated. The number of singleton clusters and clusters with less that six or eight included elements was recorded. For further diagnosis, the share of the largest cluster of the respective partition was noted. These size-based metrics should help to assess whether the candidates are unsuitable for this thesis application despite their good distance-based characteristics. To be methodically usable, the clusters, as discussed in Section 2.2.5, have to enable an interpretation as building archetypes.

As a further diagnosis, the clarity of the individual building assignments was considered. For this the affinity-score, described in Equation 2.4, was calculated; this transforms the mean distances of a building to the clusters into normalized inverse-distance proximity values. The metric `affinity_gap_12` derived from the aforementioned should also serve for the silhouette coefficient because it does not assess overall cluster separation, but it shows how the highest affinity of a building is distinct from the second-highest affinity. A small distance here shows a potentially ambivalent assignment. To now examine candidate solutions regarding their affinity, the parts of the buildings, in which the affinity gap was below the threshold value of 0.05, were assessed.

#### Selection of partition-wise Cluster-Solutions

Based on the evaluation of the candidate solutions, subsequently for each partitioning, a final cluster solution could be chosen. The selection did not take place solely through a single metric like the best silhouette but was carried out as a rule-based evaluation



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of several criteria, as shown in Table 5.8. Through this, a controlled trade-off between statistical separation and interpretability could be achieved.

Table 5.8: Criteria used for selecting the baseline partition-wise solutions.

| Criterion | Purpose | Function Version |
|---|---|---|
| `min_cluster_size` | Avoid overly fragmented solutions with very small clusters | v1–v3 |
| `silhouette` | Main separation criterion; from v2 onward used with a tolerance range rather than maximized | v1–v3 |
| `k_requested` | Ensure a minimum level of granularity, especially for the residential partition | v2–v3 |
| `largest_cluster_share` | Identification of dominant clusters leading to coarse solutions | v2–v3 |
| `neg_silhouette_share` | Tie-breaker favoring fewer uncertain assignments | v3 |
| `mean_within_cluster_distance` | Tie-breaker favoring more compact and internally homogeneous clusters | v3 |
| `low_gap_share` | Tie-breaker favoring fewer ambiguous assignments | v3 |

In the course of the baseline-clustering process several iterations were carried out, in which the rule-based selection function was developed further step by step. In the first variation for every partition the solution was chosen that satisfied a rule regarding a minimum size for the clusters and thus showed the best silhouette score. Additionally, when there was a tie of silhouette value then the solution with the larger `min_cluster_size` and subsequently smaller `largest_cluster_share` was prioritized. This approach was suitable as a simple reference, but it was already evident that the strong metric optimization leads to a coarse separation of the clusters.

In the second version, the selection function was extended by a tolerance approach. Thus, the solution with the highest silhouette score was not primarily chosen, but rather any solution whose silhouette value was in the tolerance range of the best partition solution. In the first tolerance iteration, a tolerance of 20% and in further iterations, 10% was used. This way, candidates which show a slightly lower separation, but therefore a greater granularity could be considered; this way a better interpretability should be created. Additionally, in this step a partition-specific minimum value for the cluster number $k$ was introduced. By default, this was $k = 2$, but for the residential partition, because of the strong heterogeneity, the values $k = 3$ and $k = 4$ were experimented with as well. Lastly, also the part of the largest cluster was taken into account to evade solutions with very dominant clusters. First, `max_largest_cluster_share_default` was set at 60% and, in later iterations, at 70%. From the number of valid candidates, the solution with the highest $k$ was chosen.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The final variant of the selection function still used the tolerance approach for the silhouette, but different and additional decision criteria were chosen. While in the earlier stage only the solution within the tolerance range, which had the largest cluster number $k$ was chosen, the selection function in this step uses more robust criteria. As the first decision criterion, the share of negative silhouette values `neg_silhouette_share` was checked to reduce edge cases in the allocations. As a second criterion `low_gap_share`, the share of buildings with small affinity gaps between the best and the second-best cluster was considered to choose solutions with less ambivalent allocations. Additionally, the mean distance between clusters `mean_within_cluster_distance` was taken into account to prefer more compact solutions. Only then, criteria such as the exact silhouette score, the minimum cluster size and finally the cluster amount $k$ were taken into consideration for the selection. An additional threshold value for the maximum share of the largest cluster was introduced for solutions with $k = 2$, since the chance for a dominating cluster is higher in this case.

The stepwise development of the selection function aimed to achieve a balanced solution without relying on a single metric. While the first variant favored metrically stable but also too coarse solutions, the second step differentiated more strongly but produced qualitatively weaker candidates. The final function represented the balance compromise between cluster quality and interpretable granularity, which was aimed for. Because of this approach, the clustering was not seen as a simple optimization but as a process in which quality metrics and interpretability were carefully balanced.

### Baseline Solutions and Initial Assessment

Based on the selection function described in the previous section, for every partition, the best-suited candidate solution could be chosen. Together, these formed the solution of the hierarchical clustering. In the course of the baseline clustering, a total of 12 cluster solutions were found. In the iterations, the different developmental stages of the selection function were applied, and various combinations of the selection parameter were tested.

The solutions, which were created in the course of the baseline runs, are shown in Figure 5.5. Here it becomes evident that the results of the individual iterations, especially the steps of the selection function, differ greatly. Additionally, the baseline runs formed different compromises between granularity and robustness. These conflicts of objectives are especially evident in the runs `B0` and `B10`, which are the runs with the largest discrimination ($median silhouette > 0.4$), because they were at the same time coarse and every partition was only divided into two clusters. In contrast, the highly granular solutions had a large number of clusters but also showed lower silhouette values and assignment quality, with significant declines in some cases. For example, the iterations `B1` to `B3`, which entirely originated in the second step of the selection function, and `B5`, `B8` and `B9` from the final selection function belong to the aforementioned.

Lastly, the following solutions were identified, showing a balanced approach regarding cluster amount and robustness: `B4`, `B6`, `B7` and `B11`. In these cases the silhouette and



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Figure 5.5: Comparison of the baseline clustering runs by total and residential cluster count, median silhouette score and shares of potentially uncertain assignments.

allocation metrics were closer to the coarse solutions, but with a slightly higher number of clusters. The number of clusters was particularly relevant within the residential partition. This can be attributed to the partitioning, in regard to the number of buildings included, being the most dominant, which enables a relatively differentiated solution without having the cluster lose interpretability and meaning.

As expected, the coarser solutions favor better separation and lower uncertainty, but lower interpretability is present because of a lower number of clusters. This is why the challenge was to find a solution that enables content differentiation without losing stability regarding the cluster formation. The solution `B7` was identified as the most balanced compromise, since three clusters are formed in the residential partition, while for the three considered quality metrics regarding cluster separation and assignment it yielded one of the best results. The metrics were evenly distributed across the partitions, making the solution robust across all partitions. Compared to the coarser two-cluster solutions, `B7` provided more differentiation regarding the residential partition while maintaining a



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

strong separation. At first glance, B8 seemed to be an interesting alternative because of the high level of differentiation in the residential partition. Upon a more careful look, it becomes evident that the largest residential cluster with 1,440 buildings was even more dominant than the largest cluster of the B7 solution with 1,321 buildings. Additionally, the residential partition includes four very small clusters, which only make up 1.4%, and thus the partition shows a strong fragmentation.

Apart from the assessment of the cluster numbers and the quality metrics, the baseline cluster solutions were also checked spatially. For this, the existing cluster labels at the building level were used to symbolize the buildings cartographically regarding their cluster association. The spatial visualization was not used as a further evaluation criterion for selection, but only served as a basis for the qualitative review regarding plausibility and distribution of the clusters.

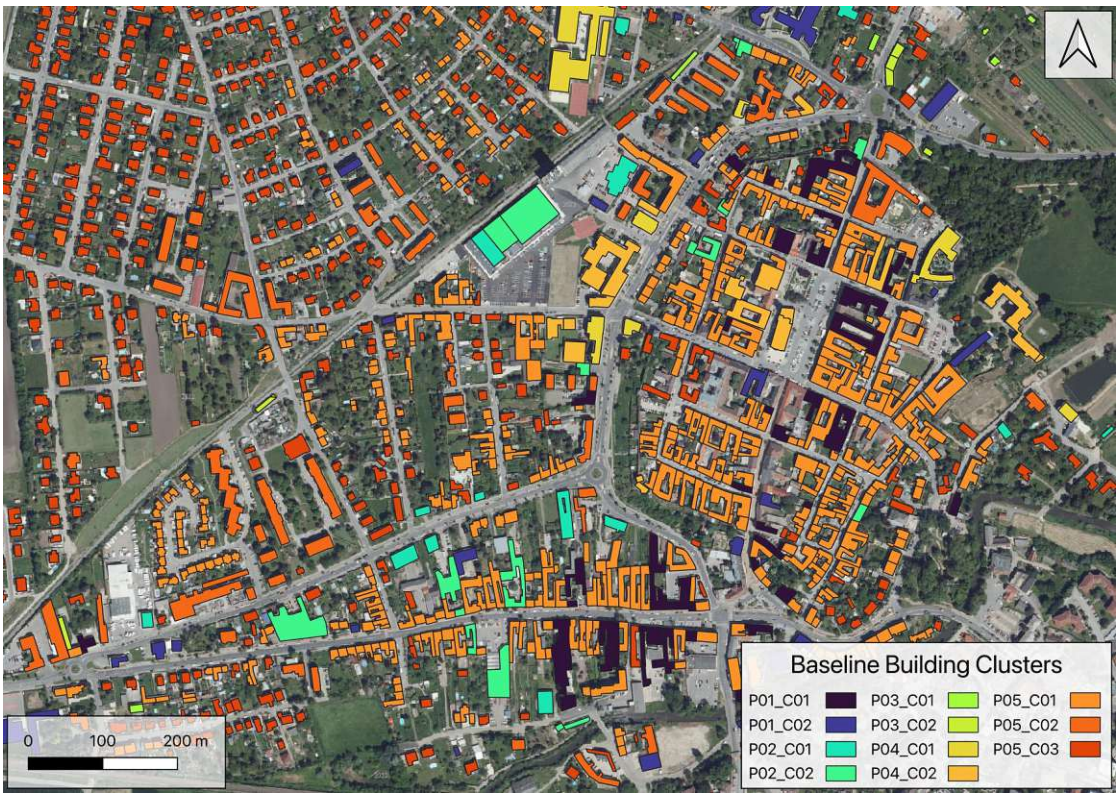


Figure 5.6: Spatial distribution of the baseline clustering result for run B7.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

Figure 5.6 shows the spatial distribution of the cluster assignments for B7. The clusters received a label structure from which the associated partition and the local cluster number could be derived. For example, P01_C02 describes the second cluster of the first partition. The individual cluster labels were assigned a color and the associated buildings



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

were colored accordingly. The colors solely serve the purpose of differentiating the clusters and neither evaluate nor indicate content similarity. However, the symbolization was assigned on the basis of a gradient, which is why consecutive cluster labels, for example those within one partition, have a similar color value. Because the cluster formation is not based on spatial coordinates but on a defined feature space, connected areas are not necessary, but can appear due to the nature of settlement structures. For the plausibility check, it was mainly relevant whether the formed groups were both spatially and structurally plausible.

While the spatial distribution was inconspicuous in general, the visual and attributive check showed that the baseline solutions are strongly characterized by basic structural features. Especially, the structural attachment of a building dominated the assignment of the clustering. This is why the partitions were divided into the two groups "detached building" and "attached building" based on their attachment status; in the residential partition, buildings had an additional intermediate form. This intermediate category is technically plausible because the structural attachment is a relevant feature for building archetypes. The differentiation of detached and attached buildings thus is relevant because they usually differ regarding proportions of exposed envelope area, building typology, and urban context. The structural attachment, therefore, may influence energy-related characteristics, making it relevant for the application of this work.

At the same time, the baseline solutions showed that the attachment-related features dominate the cluster formation too strongly; these included `touching_building_count`, `attachment_ratio` and `shared_wall_length`. This is why other relevant building data, for example construction period and size, were not included enough in the formation of the groups. This resulted in the loss of fine distinctions regarding other building information or in it only being limitedly visible. The resulting clusters formed a kind of "attachment-groups" instead of substantial building archetypes.

Since the separation after structural integration is sensible both from a technical standpoint and for the specific application, it should not be attenuated too much or even removed entirely from the feature space. Rather, it shows the boundaries of the flat approach for the present dataset: If the dominant attachment features of the structural integration are processed with all further features in a single step, this may obscure finer differences. A lesser weighing of these features could have been possible, but would have weakened meaningful separation. The baseline thus indicated that the selected feature space can map structured building groups, but it also showed the demand for a more refined approach.

This is why the attachment status was used as a starting point for a structured further development. The subsequent two-stage approach uses the observations by first mapping the differences in structural integration in an explicit first step. The resulting groups then undergo a second clustering step for further differentiation, based on the remaining features.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

### 5.3.3 Two-Stage Clustering Approach

Since the baseline approach yielded robust and somewhat structured results, two boundaries were still shown. Firstly, with more robust results, too few clusters were identified, which made a meaningful derivation into building clusters possible only to a limited extent. Secondly, it became apparent that the features of the structural integration dominated the distance calculation and thus overlaid finer differences in the rest of the feature space.

From these results, the motivation for a two-step clustering approach was formed. This approach should keep the desired strong separation of the structural integration and, at the same time, enable a distinction based on the features regarding morphology, usage and neighborhood context. In the first step, the clustering was solely based on the attachment features. The created groups were then used as a starting point for the second step, in which morphological, functional and contextual differences within the identified attachment types were examined. In the last step, the feature weighting for the Gower distance calculation was used to improve the results of the final clustering solution and to align them for the targeted representation of the building archetypes.

#### Stage 1: Attachment-based Clustering

The first step of the two-stage approach is mostly based on the methodological approach of the baseline clustering. The upstream steps, such as the partitioning after `grouped_category`, the preprocessing of the features and also the calculation of the paired Gower distance matrices were kept. The formation of the candidate solutions was also carried out as hierarchical clustering with a subsequent assessment based on the same quality metrics and final selection of the best solutions per partition with the selection function.

The key difference was in the feature set used for the calculation of the distance matrix. While the baseline approach used the entire feature space in a flat run, in this step only the attachment-features, which are listed in Table 5.9, are used. The distance metric itself stayed the same, with the exception of the reduced feature space.

Table 5.9: Feature set used for the attachment-based stage 1 clustering.

| Feature | Type | Interpretation |
|---|---|---|
| `touching_building_count` | numeric | Number of directly touching buildings |
| `attachment_ratio` | numeric | Share of perimeter attached to neighboring buildings |
| `shared_wall_length` | numeric | Length of shared walls |

Because the aim of the clustering was the formation of coarse connection types, the maximum cluster number per partition was limited based on the parameter `max_k_default` during the formation of the candidate solutions. This value was defined as $k = 4$ to take po-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

tential transitional forms or intensities of the structural co into account. For the residential partition, a minimum amount of clusters was given with `min_k_by_partition=3` because at least one intermediate form of building with a loose structural connection was considered relevant here for building typologies such as semi-detached houses. For the other partitions this value remained at $k = 2$. Additionally, with `preferred_k_default=3`, a favored cluster amount was defined so candidates near this coarse structure would be prioritized among solutions of equal quality. Candidate solutions with very small clusters were revised already during the formation process. For this, a minimum size of four buildings per cluster was defined. Clusters that fell below this threshold were merged with the nearest cluster. As opposed to the baseline clustering, methodologically unwanted solutions were not filtered out afterward but were already evaded before the assessment. So, the focus in the first step was on a few attachment patterns, but therefore they were easier to interpret from a technical standpoint.

The formation of the candidate and the selection function followed the base logic of the baseline approach, but were supplemented with restrictions for the new aim. For example, the logic of the silhouette-tolerance range, the wanted minimum cluster size and the maximum share of clusters in the partitions were kept. Because several iterations were tested in stage 1, slight adaptations regarding the maximum share of the clusters during the respective iterations were made.

Table 5.10 lists the iterations of the attachment-based clustering. The run `S1_4` was identified as the best solution because it showed the highest silhouette coefficient and, at the same time, the lowest amount of negative silhouette values and the lowest share of low assignment confidence. With a total of 11 identified clusters, the solution was also very compact and matched the aim of deriving coarse and interpretable connection types. After `S1_4`, the optimization for step 1 was finished, because the already very high quality metrics and the compact structure met the objective and further parameter tuning was not expected to add substantial methodological value.

Table 5.10: Comparison of attachment-based stage 1 clustering solutions.

| Run | Clusters | Median silhouette | Neg. silhouette share | Low-gap share |
|---|---|---|---|---|
| `S1_0` | 11 | 0.667 | 0.045 | 0.054 |
| `S1_1` | 12 | 0.768 | 0.026 | 0.043 |
| `S1_2` | 13 | 0.732 | 0.018 | 0.034 |
| `S1_3` | 12 | 0.999 | 0.014 | 0.021 |
| **`S1_4`** | **11** | **0.999** | **0.012** | **0.020** |

As the spatial representation of the median silhouette values in Figure 5.7 shows, detached buildings as well as building structures with little or uniform structural connection, for example, semi-detached houses, showed a clear classification. Lower or isolated negative values showed themselves at transitional forms, for example, buildings with a punctual or several short contact surfaces to neighboring buildings. The spatial review thus confirmed



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

that attachment-based clustering can structurally capture clear connection patterns, while transitional forms were shown to be less delineated cases.

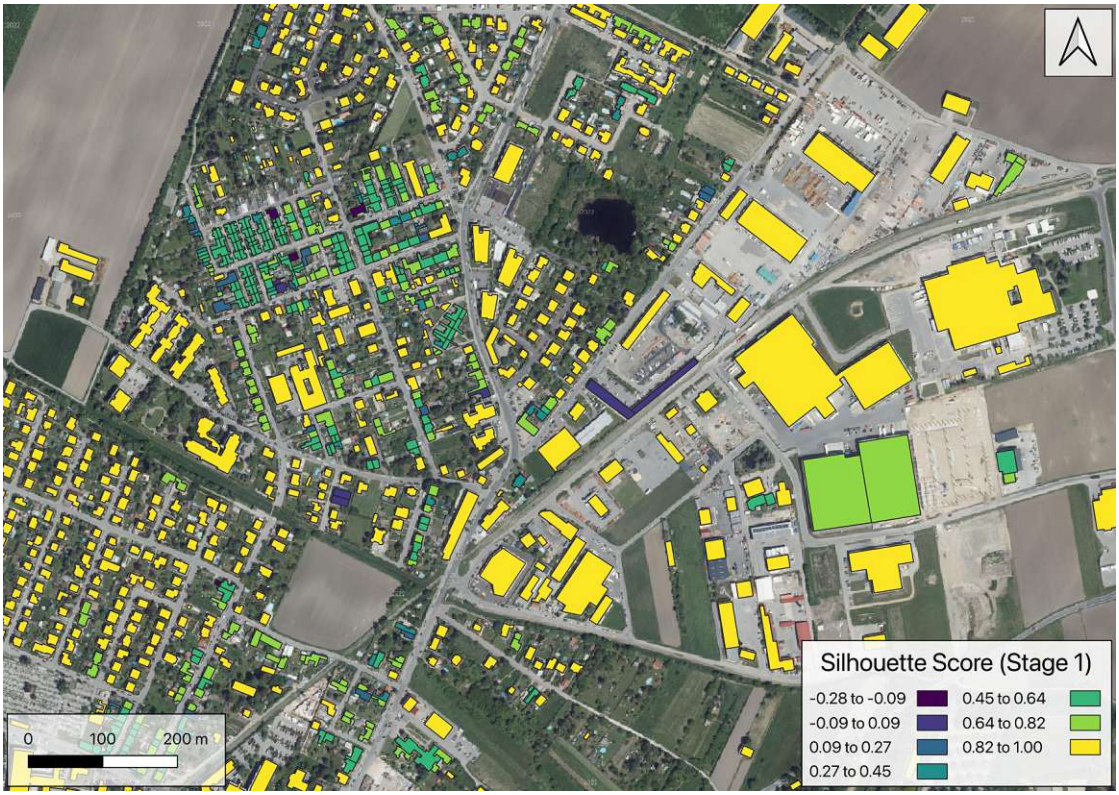


Figure 5.7: Spatial representation of median silhouette scores in stage 1 for run `S1_4`.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

### Stage 2: Subclustering within Attachment-Groups

The second stage of the clustering immediately follows the first step because the clustering, in this step, was carried out within the identified attachment patterns. Compared to the baseline clustering, the starting point was different because the clustering was not just mapped at the grouping level of the partitions. This means that with the Gower distance calculation, a distance matrix was computed for every stage-1 cluster. Additionally, the attachment features (see Table 5.9) were excluded for the distance calculation, since these had already been used in stage 1. The aim of this step was to find ideally interpretable subclusters, based on the rest of the feature space in Table 5.1, which could subsequently be mapped to archetypes.

The assessment, along with the selection function for selecting the best solution candidates, was taken mainly from the previous step, whereby the parameter `max_k_default=12` instead of $k = 4$ from step 1 was chosen for the maximum cluster number. The reason is



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

that in this clustering process, more finely differentiated solutions were desired in order to receive interpretable results. Stage-1 clusters, which fell below $min_group_size = 25$, were directly adapted to step 2. The quality metrics stayed the same for this step to choose the best solution candidates.

Because the already identified clusters were split further in this step, a new methodological decision option was added: either keep the cluster the same or split it into further subclusters. To pass this judgment, a function was implemented which controlled, based on the control-parameters from Table 5.11, whether a stage-1 cluster should be split. The comparison for this took place solely in the feature space of stage 2. First, for every cluster from the first step, a reference was calculated based on the middle paired Gower distances within the group. This reference was compared with the mean within-cluster distance of the proposed stage-2 assignment. This is calculated by summing the mean pairwise distances (see Equation 2.3) of the subclusters and divided by the number of building pairs they contain. A split was considered to be better if this value was reduced by at least the specified percentage from `accept_split_min_relative_within_improvement`.

Table 5.11: Parameters controlling the stage-2 split acceptance logic.

| Parameter | Interpretation |
|---|---|
| `accept_split_min_relative_within_improvement` | Minimum required reduction in within-cluster distance compared to the unsplit stage-1 group. |
| `accept_split_min_silhouette` | Minimum required silhouette quality of the proposed stage-2 split. |
| `accept_split_max_neg_silhouette_share` | Maximum allowed share of buildings with negative silhouette values. |
| `accept_split_max_low_gap_share` | Maximum allowed share of buildings with low assignment confidence according to the affinity gap. |
| `accept_split_exception_min_relative_within_improvement` | Relaxed improvement threshold for otherwise clear stage-2 splits. |
| `accept_split_exception_min_silhouette` | Higher silhouette requirement for accepting a split under the exception rule. |
| `accept_split_exception_max_neg_silhouette_share` | Stricter limit for negative silhouette values under the exception rule. |
| `accept_split_exception_max_low_gap_share` | Stricter limit for low-confidence assignments under the exception rule. |

For a relative comparison between stage 1 and stage 2, the silhouette coefficient could not be utilized meaningfully. This is due to the undivided stage-1 clusters, where each one only represents a single cluster. Because the silhouette puts the distance to the own cluster in proportion to the nearest cluster, it is undefined in this case. Instead, the other parameters from Table 5.11 were defined as quality criteria for the acceptance of a separation into the stage-2 clusters. These requirements should ensure a certain minimum silhouette as well as a maximum share of negative silhouette values and low



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

affinity gaps.

To prevent the compactness comparison using the mean within-cluster distance from eliminating all solution candidates that are only slightly below the threshold value, exceptions were introduced. These were controlled through the parameters which are marked by the word `exception`. The idea was that stage-2 split candidates, which were only slightly below the minimum improvement, could still be used if they had especially good values for all quality criteria.

If a solution candidate was chosen and subsequently accepted, the label structure of the cluster was extended to ensure the traceability of the individual steps. For example, `P05_S1_C01_S2_C02` should describe the second local cluster of stage 2 which was formed from the first cluster of stage 1 for partition 5. If the separation of a cluster was not accepted or if there was no valid solution candidate, then the stage-1 cluster remained unchanged. The created solutions of the second step thus contained clusters which were already created in stage 1 and were not split further and those subclusters, which were created in stage 2 through further division.

The aggregated metrics of the created solutions from the stage-2 iterations were only limitedly comparable to each other, because the stage-2 metrics were combined with those of the stage-1 clusters carried over. Because the stage-1 clusters provided good quality metric values (see Table 5.10), solutions with fewer splits had substantially higher quality metrics. To enable an ideally fair comparison between the results of the second step, the evaluation logic for the separation assessment was applied analogously to the final cluster labels of the solutions. This way, the runs with an identical stage-1 initial solution could be evaluated uniformly based on the distances of the stage-2 feature space regarding quality metrics and the improvement in within-cluster distance compared to the no-split reference. For no-split groups, the metrics were not taken and the relative improvement of the within-cluster distance was set to 0, because, in this case, no change of the clusters was made. Also, the non-defined metrics, such as silhouette and affinity gap, were not considered in this case.

Table 5.12: Fair comparison of stage 2 clustering solutions based on `S1_4`.

| Run | Clusters | Split groups | Split buildings | Silhouette | Negative. sil. | Low gap | Rel. improvement |
|---|---|---|---|---|---|---|---|
| S2_0 | 22 | 10/11 | 99.8% | 0.317 | 3.2% | 8.0% | 10.8% |
| S2_1 | 19 | 7/11 | 77.9% | 0.342 | 3.5% | 8.1% | 8.7% |
| S2_2 | 16 | 4/11 | 28.8% | 0.260 | 5.3% | 13.2% | 4.7% |
| **S2_3** | **17** | **5/11** | **76.6%** | **0.342** | **3.4%** | **8.1%** | **5.3%** |
| S2_8 | 17 | 6/11 | 75.5% | 0.322 | 5.2% | 10.7% | 5.5% |

The following analysis of the solutions of the second step, listed in Table 5.12, which are all based on the initial solution of `S1_4`, showed that no solution dominated all criteria. `S2_0`, `S2_1` and `S2_3` were all identified as solutions with three different allocation patterns. A subsequent spatial assessment then showed that no solution could achieve the aimed archetypal distribution. The additional separations in `S2_0` and `S2_1`, for example, in the partition Social Infrastructure, did not always lead to an added



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

substantive value. The run S2_3, however, showed the least fragmentation with similar quality metrics. This is why it was chosen as a base configuration to achieve the missing differentiation regarding the pursued archetypes by weighting the features.

For example, a targeted separation between row houses and block-edge residential buildings should be implemented through this. As shown in Figure 5.8, both building typologies were still assigned to the cluster displayed in dark red. The row houses are shown on the left image boundary and the block-edge residential buildings in the middle as well as the right side of the picture.

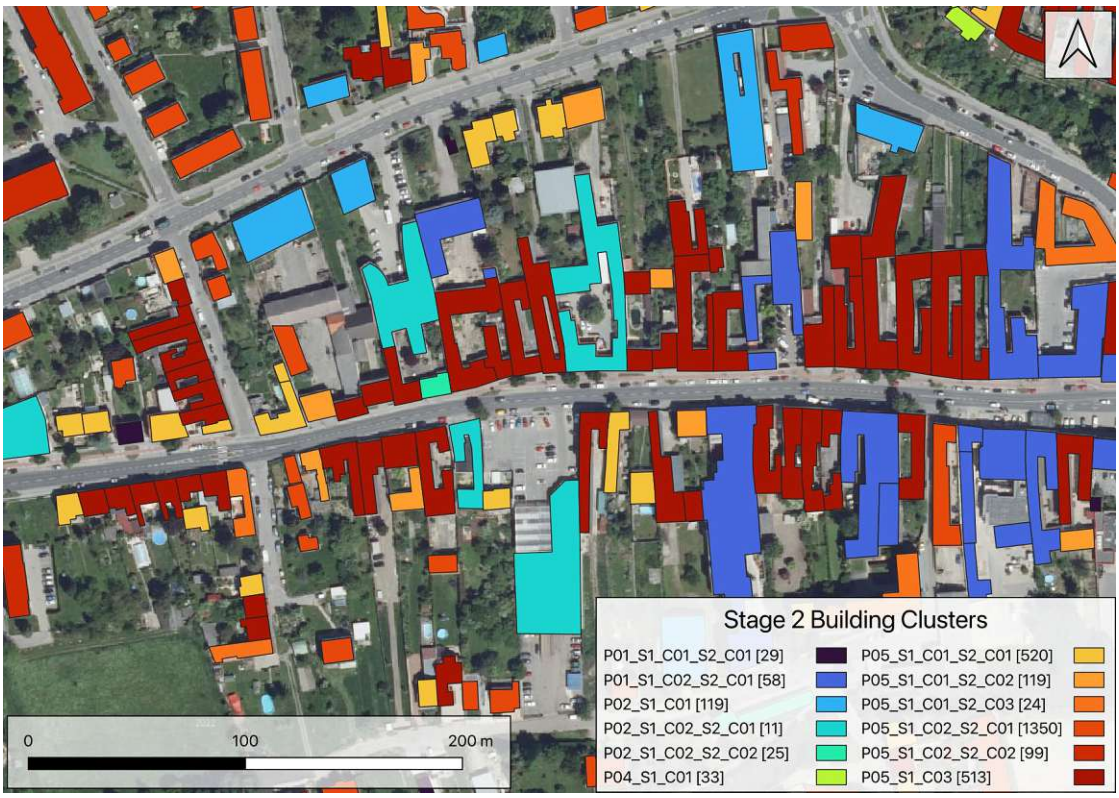


Figure 5.8: Spatial distribution of the stage 2 clustering result for run S2_3.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

The weighted approach did not use the already created clusters from S2_3, but used the parameters of this iteration and supplemented the creation of the clusters on the basis of S1_4 and applied feature-level weighting. This way, the clusters should be formed in a more deliberate manner in order to achieve a more interpretable cluster separation. The weights were defined group-specifically; for not explicitly defined features, the default value 1.0 was used. With the help of the weights the influence of individual features could be improved or limited. The refinement of the cluster solutions also included partition-specific boundaries of the cluster numbers as well as definitions of minimum size



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

and of exception rules for the split-acceptance. For both groups in the partition Social Infrastructure, the separation was refined in a rule-based manner, based on building size, height and compactness, to separate large buildings from the others.

The resulting refined solutions were also examined with the fair evaluation method. To be able to compare iterations with different weights, they were not taken into consideration in the assessment. Instead, the unweighed Gower distances were used to evaluate the effects of different weighting profiles in a unified manner. The iterations of this approach were lastly also checked spatially because already slight adjustments of the weight could change the respective distance matrix and thus the cluster separation. Thus the challenge was to adapt the weights iteratively until technically plausible potential solutions were available.

### Final Cluster Solution

As shown in Table 5.13, the created solutions tended to be within 19 and 22 created clusters, whereby on average they were more differentiated than the original stage-2 solutions. The quality metrics in this step only served the purpose of orientation and control variables, but not as a direct criterion for the choice of the best solution. This resulted from the logic that a stronger differentiation of the clusters, as was wanted for the archetypes, would go hand in hand with a certain worsening of the separation metrics. Given the background that interpretable archetypes should be depicted based on the created cluster labels, the slightly lower quality values were seen as appropriate. Thus, the term "refinement" refers only to an improvement in terms of morphological and content resolution, but not in terms of quality metrics.

Table 5.13: Fair comparison of representative weighted stage 2 clustering solutions.

| Run | Clusters | Split groups | Split buildings | Fair sil. | Neg. sil. | Low gap | Rel. improvement |
|---|---|---|---|---|---|---|---|
| S2_3 | 17 | 5/11 | 76.6% | 0.342 | 3.4% | 8.1% | 5.3% |
| S2_w_0 | 21 | 8/11 | 95.3% | 0.292 | 5.8% | 14.8% | 9.5% |
| S2_w_2 | 19 | 7/11 | 73.4% | 0.315 | 5.8% | 12.1% | 8.3% |
| S2_w_4 | 21 | 8/11 | 95.3% | 0.291 | 6.3% | 14.7% | 9.2% |
| S2_w_5 | 19 | 7/11 | 73.4% | 0.315 | 5.8% | 12.2% | 8.0% |
| S2_w_7 | 21 | 8/11 | 95.3% | 0.287 | 6.9% | 15.4% | 7.8% |
| S2_w_13 | 22 | 9/11 | 99.2% | 0.287 | 6.3% | 15.4% | 8.7% |
| **S2_w_14** | **22** | **9/11** | **99.2%** | **0.287** | **6.2%** | **15.3%** | **8.9%** |

After each iteration with feature weights, a spatial check took place, and the solutions could be separated into structural configurations that are as homogeneous as possible. This is shown, for example, in the separation of single-family houses and allotment gardens, as well as semi-detached homes and row houses. Exactly from this aspect of the best substantive separation and thus highest technical plausibility, the solution S2_w_14 was chosen as the final clustering solution. This solution proved to be the best compromise between archetypal interpretability, granularity and robustness from all created solutions. The clustering process was finished at this point, because further



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

adaptations would not have brought a recognizable substantive value and the existing solution `S2_w_14` was considered sufficient.

The map sections shown in Figure 5.9 illustrate the substantive value added by the final refinements. For example, in the solution `S2_w_14` row houses (magenta) and block-edge residential buildings (dark red) were assigned different clusters, while in `S2_3` they were still grouped in one cluster. Comparable examples for a higher typological resolution are also found in further partitions.

(a) The final solution `S2_w_14` assigns the two building structures to separate clusters.

(b) Run `S2_3` assigns row houses and block-edge residential buildings to the same cluster.

Figure 5.9: Spatial comparison of the differentiation between row houses and block-edge residential buildings. Both panels show the same spatial extent.

*Source:* Own representation based on basemap from Geoland.at [Geo25] and building geometries from SmartQ+ [BSR+25].

For a final assessment of the development from the beginning to the end of the clustering process, `S2_w_14` was compared to the baseline solution `B7`. In Figure 5.10, it is shown that the three original clusters in the residential partition in the coarse baseline solution are contrasted with eight more differentiated groups in the final version. Notably, the residential clusters in the final solutions are mostly composed of buildings from a single baseline residential cluster. Thus, the additional granularity does not go hand in hand with a substantial mixing between the groups, thereby preserving the original structure of the assignments.

## 5.4 From Cluster Labels to Building Archetypes

After the process for the formation of the cluster labels was described in the previous section, in the next step these 22 labels could be translated into technically meaningful building-archetypes. For this, in the next sections, the requirements and the assignment of the labels to these archetypes are described.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

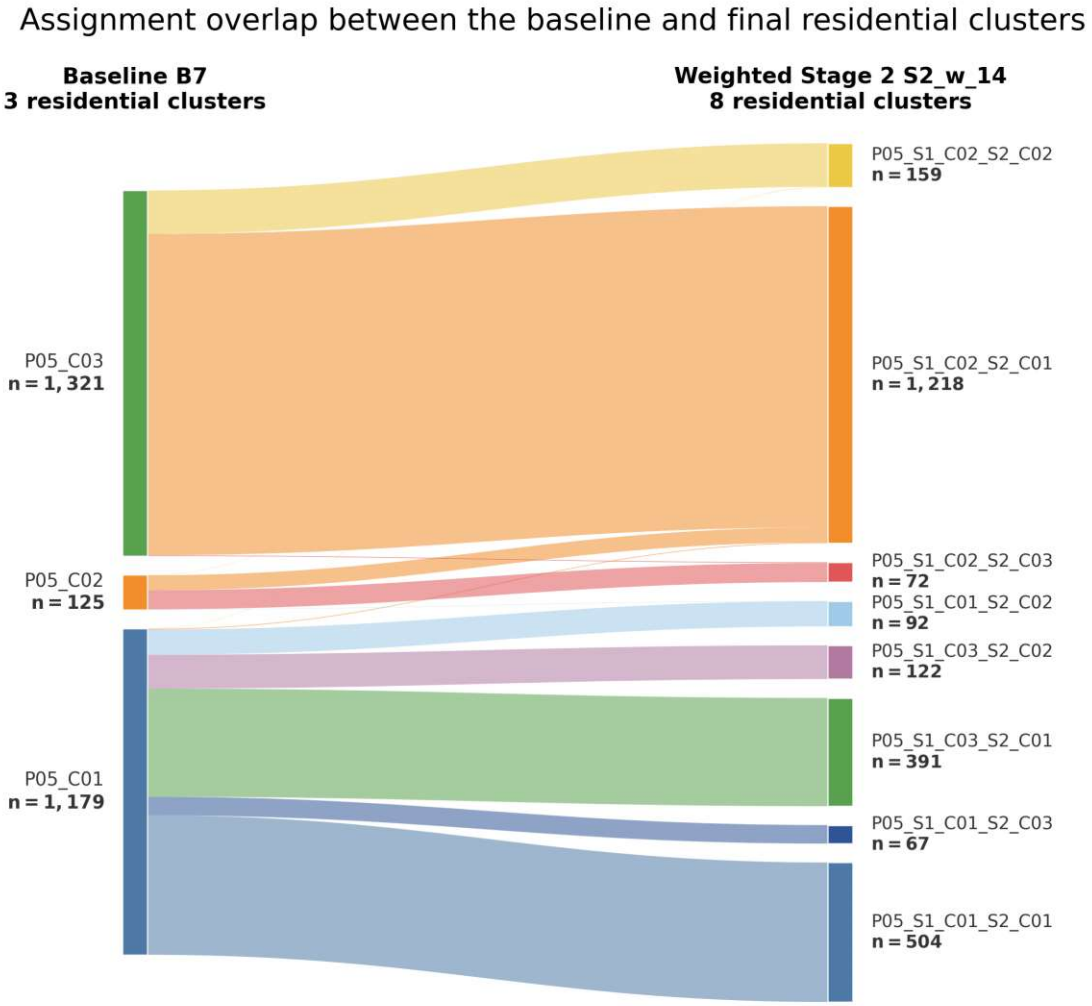


Figure 5.10: Assignment overlap between the residential clusters of the baseline solution `B7` and the final solution `S2_w_14`.

### 5.4.1 Construction of the Archetype Catalog

As described in Section 5.1, the heterogeneous building stock should be transformed into an ideally interpretable and relevant structure. The existing building typologies served as a useful starting point, but could not be applied directly. For this reason, an archetype catalog was developed, which not only considers residential buildings but also industrial and commercial buildings as well as social infrastructure, and can display their local morphological forms.

For the inclusion of an archetype, the following requirements were necessary: Based on the used clustering features, the categories had to be sufficiently distinguishable from different building structures and a potential relevance regarding the energy prognosis had to be



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

present. Additionally, the catalog's archetypes had to cover all buildings in the examined area. Because no detailed usage profiles were available, the link to the energy prognosis was not based on individual user behavior. Instead, the archetypes described recurring combinations of morphological and functional features, whose suitability for subsequent prognosis of the energy metrics was investigated in the course of this thesis. The attachment properties do play a special role because the varying structural configurations describe buildings, such as detached buildings, buildings attached on only one side, or buildings integrated into block structures. This way, they give hints regarding their share of exposed building envelopes which, in the context of the contemplated energy prognosis, are a relevant factor especially for heating and cooling numbers. Only then was an attempt made to make a more refined distinction between archetypes within these groups based on the remaining set of features. The methodological implementation of the two-step approach was already described in more detail in Section 5.3.

The archetype catalog was created simultaneously with the development of the cluster labels. A first draft, especially regarding the buildings of the residential category, was created on the basis of existing building typologies (i.e. TABULA/EPISCOPE) and building structures relevant to planning as well as a visual analysis of the building stock [Aus15]. This preliminary draft served as a subject-specific framework for evaluating the formed clusters for sufficient and meaningful differentiation. It did not depict a ground truth in the form of already labeled target classes as the cluster assignment remained unsupervised based on the defined feature space.

During the course of the clustering, the archetypes of the final catalog were refined based on the derived cluster labels. While some labels could directly be assigned an assumed archetype, some labels described the same building structure in slightly different forms and were summarized in one common archetype. Additionally, based on the clustering results, archetypes were sometimes identified and adopted in the final catalog. The original draft of the archetype catalog was thus specified based on the structures that emerged from the clustering and finalized under the consideration of the previously defined requirements.

The differences in granularity of the archetypes between the functional partitions resulted from the small proportion in the existing dataset. The residential buildings could be differentiated more finely because of the large number and the morphological diversity. In the other partitions, however, the differentiation between larger and smaller buildings as well as their structural integration was foregrounded.

The results of this process led to a catalog with a total of 14 archetypes, as depicted in Table 5.14. The defining distinguishing characteristics of the respective archetypes are also described there.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.14: Final building archetype catalog and defining characteristics.

| ID | Domain | Archetype | Defining characteristics |
|---|---|---|---|
| 1 | Commercial | Commercial, office and service buildings | Small to medium-sized commercial and service buildings, commonly located along streets or embedded in mixed-use and central settlement structures. |
| 2 | Commercial | Large-scale retail | Large, usually freestanding buildings with flat roofs and adjacent parking or service areas. |
| 3 | Industrial | Large industrial and commercial halls | Large to very large hall-like buildings associated with production, logistics, loading areas or operational yards. |
| 4 | Industrial | Small-scale workshop and commercial service buildings | Small to medium-sized workshops or service buildings, often attached or embedded in denser commercial structures. |
| 5 | Social infrastructure | Large institutional and community buildings | Large or multi-wing institutional buildings, often accompanied by schoolyards, driveways or landscaped open spaces. |
| 6 | Social infrastructure | Small-scale institutional and community buildings | Smaller community-oriented or ancillary buildings, sometimes with distinctive footprints, roofs or associated civic open spaces. |
| 7 | Other | Other | Technical infrastructure, ruins and special-purpose buildings that cannot be assigned to another archetype. |
| 8 | Residential | Detached single-family house | Detached residential building with a small to medium footprint, commonly surrounded by a private garden or yard. |
| 9 | Residential | Allotment garden/shed | Very small and low building located in an allotment or garden context, sometimes without direct formal street access. |
| 10 | Residential | Small building extension/garage | Small attached ancillary building connected to one neighboring building, but not forming a regular semi-detached pair. |
| 11 | Residential | Semi-detached house | Two directly attached residential units, frequently displaying a symmetrical or near-symmetrical building structure. |
| 12 | Residential | Row house | At least three attached residential units with narrow, repeated footprints, typically arranged along a street. |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

| ID | Domain | Archetype | Defining characteristics |
|---|---|---|---|
| 13 | Residential | Dense urban townhouse / block-edge residential building | Larger attached residential building forming a continuous street edge or part of a dense central or block-edge structure. |
| 14 | Residential | Apartment block | Comparatively large multi-dwelling building, usually freestanding or located in an open-set layout. |

### 5.4.2 Interpretation and Mapping of Cluster Labels

Because the archetype catalog and the clustering solutions were developed iteratively together, the final interpretation of 22 formed labels does not depict a completely retroactive designation. Moreover, for every cluster label it was checked, whether the assigned building matched with an archetype in the catalog. For this, the buildings assigned to each cluster were examined in greater detail using GIS to assess their spatial distribution and morphological characteristics. Thus, the cluster labels were mapped to the archetype that best matched the respective feature profile regarding their characteristics.

All 22 clusters were assigned exactly one of the 14 archetypes, while these could consist of multiple cluster labels. This semantic merge happened if a cluster differed in individual features of the stage-1 path, but described the same building typology. For example, the buildings of the commercial and industrial partition, as well as the social infrastructure, were summarized from different stage-1 clusters into archetypes for their respective domains. In detail, the existing differences of the structural integration did not justify a further differentiation in the archetype catalog. Examples for such a merge are `Commercial, office and service buildings` and `Large-scale retail`, which emerged from two cluster labels from different stage-1 clusters. In contrast, for the majority of the residential partitions, there was a direct 1:1 relationship between the cluster label and the archetype. For example `Detached single-family house`, `Allotment garden/shed`, `Row house` and `Apartment block` are all represented respectively through one label.

Through the mapping, each building received an archetype based on the affiliation of a cluster label. The complete assignment of the cluster to building archetypes is depicted in Table 5.15.

Table 5.15: Mapping of the final cluster labels to the building archetypes.

| ID | Archetype | Assigned cluster labels |
|---|---|---|
| 1 | Commercial, office and service buildings | `P01_S1_C01_S2_C01`, `P01_S1_C02_S2_C01` |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

| ID | Archetype | Assigned cluster labels |
|---|---|---|
| 2 | Large-scale retail | P01_S1_C01_S2_C02, P01_S1_C02_S2_C02 |
| 3 | Large industrial and commercial halls | P02_S1_C01_S2_C01, P02_S1_C02_S2_C01 |
| 4 | Small-scale workshop and commercial service buildings | P02_S1_C01_S2_C02, P02_S1_C02_S2_C02 |
| 5 | Large institutional and community buildings | P04_S1_C01_S2_C01, P04_S1_C02_S2_C01 |
| 6 | Small-scale institutional and community buildings | P04_S1_C01_S2_C02, P04_S1_C02_S2_C02 |
| 7 | Other | P03_S1_C01, P03_S1_C02 |
| 8 | Detached single-family house | P05_S1_C02_S2_C01 |
| 9 | Allotment garden/shed | P05_S1_C02_S2_C02 |
| 10 | Small building extension/garage | P05_S1_C01_S2_C03 |
| 11 | Semi-detached house | P05_S1_C01_S2_C01 |
| 12 | Row house | P05_S1_C03_S2_C01 |
| 13 | Dense urban townhouse / block-edge residential building | P05_S1_C03_S2_C02, P05_S1_C01_S2_C02 |
| 14 | Apartment block | P05_S1_C02_S2_C03 |

## 5.5 Validation Process of Archetypes

Finally, the derived building archetypes were subjected to a validation based on expert annotations and using the usage information of the research project SmartQ+. The annotations served the purpose of checking whether it was technically plausible, while the usage information was used as an additional plausibility check. Additionally, to assess stability, the cluster structure was examined via resampling under modifications to the building stock.

### 5.5.1 Expert Annotations

In this step, the existing buildings in the study area Bruck an der Leitha were manually annotated by two experts in architecture and spatial planning based on the Archetype Catalog. For this, a QGIS project was created, in which the buildings along with geometry and all attached attributes were provided. Additionally, the same orthophoto



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

basemap was provided, which was already used in the visual-spatial review of the cluster solutions [Geo25].

The annotation for each partition was carried out in a separate layer, so only the archetypes of the associated domain could be selected. This is due to the reason that through pure orthophotos, for example, a differentiation between a commercial warehouse or an industrial logistics center is almost impossible. This way it was ensured that the same basic requirements regarding archetype allocation as for the actual clustering were provided and that every building was allocated an archetype from its respective partition.

Thus, for every building an archetype label of the corresponding partition could be chosen, which was saved in the attribute `expert_label`. If the information from the building data and the orthophoto was not sufficient for the annotation, then the experts could also select the value *unknown*. Additionally, for every annotation, a confidence level was recorded, which describes the subjective certainty of the experts and which could take the values *high*, *medium* and *low*. If this value was not explicitly set, the confidence was classified as high by default. In addition, comments could be given as text to document reasons for uncertainty. Table 5.16 lists the recorded attributes of the annotation process.

Table 5.16: Attributes recorded during the expert annotation.

| Attribute | Possible values | Purpose |
|---|---|---|
| `expert_label` | 14 archetypes, *unknown* | Expert-assigned archetype |
| `confidence` | *high*, *medium*, *low* | Confidence in the assignment |
| `comment` | Text | Optional explanation or observation |

To ensure that the annotation process is as consistent as possible, guidelines were developed and made available to the annotators. These describe the archetypes and their distinguishing features as well as the handling of ambiguous buildings. The full guidelines are listed in the appendix Annotation Guidelines.

For the evaluation, the annotations by both experts E1 and E2 were merged via the building ID. A pairwise comparison was performed for all buildings, excluding annotations with the value *unknown* and the pre-annotated partition *Other* (A07). For E1 3,008 and for E2 3,007 valid annotations were available, because E2 included an annotation with the value *unknown*. First, the annotations by both experts were compared. Additionally, the clustering archetypes were directly compared with the expert annotations, along with a consensus set of a total of 2,560 buildings, for which both experts assigned identical annotations. To evaluate the agreement, standard metrics were calculated and analyzed for all comparisons.

Based on the agreement rate, the proportion of identical archetype labels between both observed groups was described. Cohen's $\kappa$ additionally considers the expected agreement given the respective prevalence of the archetypes. Additionally, a balanced accuracy was calculated, which matches the mean of the recall across all archetypes. It indicates whether archetypes are identified with similar accuracy, regardless of their frequency. The



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

final metric, which was calculated, was the Macro-F1-score, which averages the F1-scores of all archetypes and assesses the extent to which archetypes were recognized completely or without misclassifications. The last two key metrics are relevant for the application because all archetypes are equally included in the analysis, regardless of the number of occurrences. The results of the described standard metrics for the comparisons are listed in Table 5.17.

Table 5.17: Agreement between expert annotations and clustering archetypes.

| Comparison | $n$ | Agreement | Cohen's $\kappa$ | Bal. accuracy | Macro-F1 |
|---|---|---|---|---|---|
| E1 vs E2 | 3,007 | 0.851 | 0.811 | 0.848 | 0.826 |
| Clustering vs E1 | 3,008 | 0.846 | 0.804 | 0.802 | 0.788 |
| Clustering vs E2 | 3,007 | 0.872 | 0.837 | 0.804 | 0.812 |
| Clustering vs expert consensus | 2,560 | 0.934 | 0.914 | 0.874 | 0.869 |

It becomes apparent that the agreement between the clustering and the individual experts is in the same order of magnitude as that of the expert comparison. The comparison of the consensus set scored the highest scores, but due to the lower building set it does not directly relate to the accuracy of the full building stock. However, the high values show that the clustering could reproduce the archetype assignments well, which overlapped between the two experts. The lower macro-F1 and balanced accuracy scores imply that differences are not evenly spread across all archetypes.

This is why confusion matrices were created to analyze systematic confusions of the archetypes. Figure 5.11 included the confusion matrix for the comparison of clustering and the expert consensus, which depicts the composition of the deviations for the respective archetypes. The archetypes A10 and A14 stand out in particular, with 57% and 63% accuracy, respectively, while the rest achieve significantly higher accuracy rates. In the case of A14 it is especially apparent that 32% of the deviations were assigned the archetype A08. Larger deviations to individual archetypes point out less clear boundaries or structural similarities between two archetypes. Thus, the confusion matrix confirms the assumption that the deviations of the allocations are not evenly spread out across the archetype catalog, as indicated by macro-F1 and balanced accuracy. The complete confusion matrices for all four comparisons are shown in Archetype Annotation Validation.

Additionally, a supplementary check based on the internal cluster-quality metrics was carried out. As shown in Figure 5.12, the buildings with no consensus among the experts had lower values across all three metrics. This provides the insight that technically ambivalent cases are also less clearly delimited through the clustering.

### 5.5.2 Usage-based Plausibility Check

In this step, the final archetype labels were compared to the attribute `category_tu`, which contains detailed usage information of the building stock. Since the archetypes form a combination of the morphological and functional features, the usage information



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

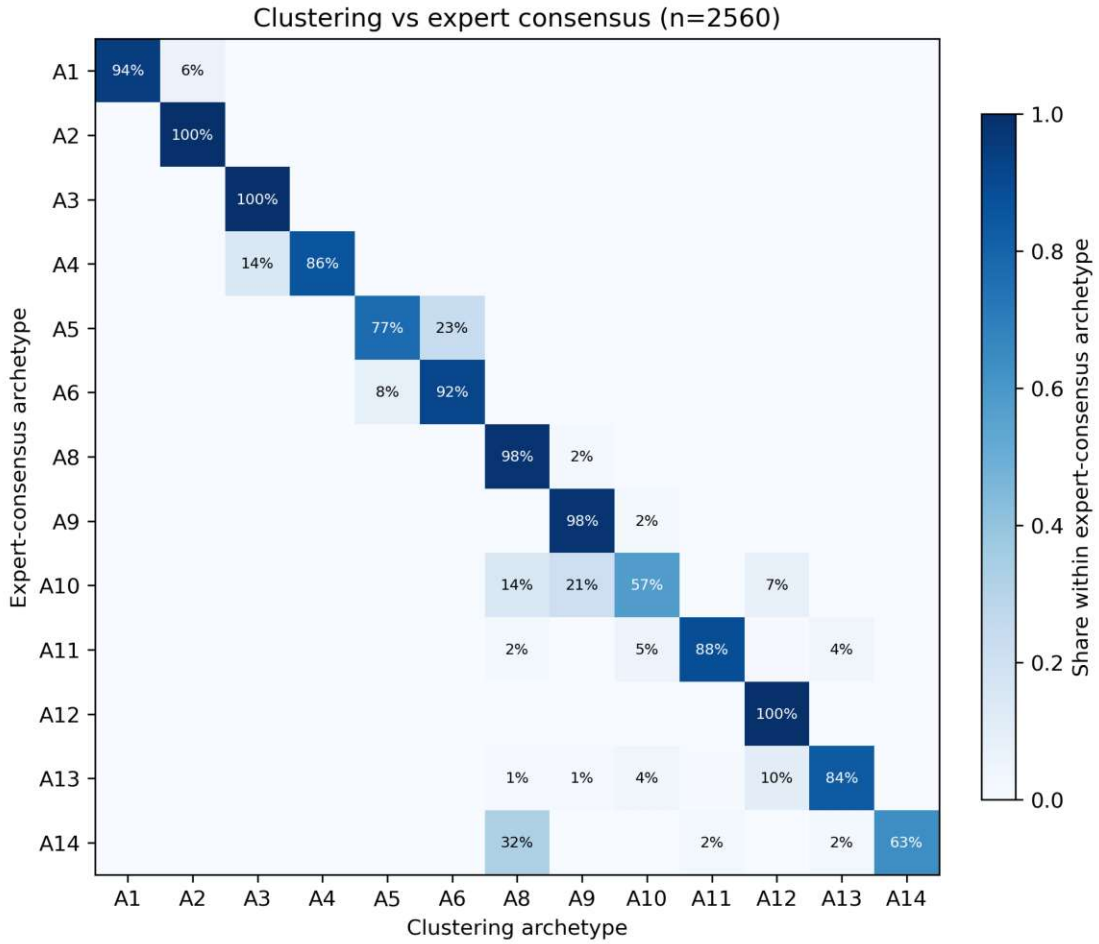


Figure 5.11: Confusion matrix of archetypes against the expert-consensus subset ($n = 2560$; pre-annotated A07 was excluded).

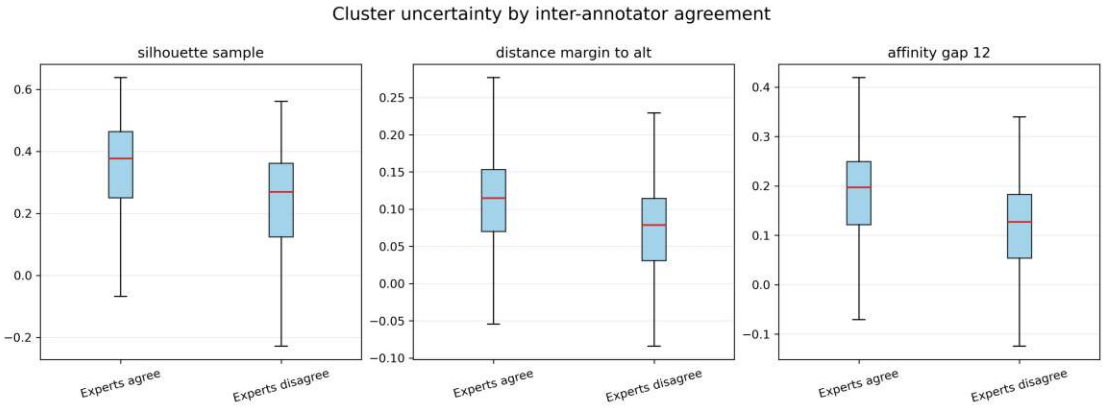


Figure 5.12: Cluster quality metrics by expert annotation agreement.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

can not be used directly as a reference. Instead, a mapping catalog (see appendix Usage-Archetype Compatibility Catalog) was created, which assigned one of the following ratings to each combination of `category_tu` and archetype within every partition: *expected*, *plausible* and *incompatible*. Since for the clustering already a partitioning using `grouped_category` was done, this crosswalk did only represent a conditional plausibility check but not an independent ground truth.

Of the total of 3,008 building allocations 78.7% were identified as *expected*, 20.8% as *plausible* and 0.5% as *incompatible*. Thus, 99.5% of the buildings were in a combination which was at least deemed to be plausible. The few incompatible mappings for the residential partition shown in Figure 5.13 resulted from some allotment-garden archetypes being mapped to multi-family buildings according to `category_tu`. The high rate of plausible allocations in the social infrastructure partition is striking. This suggests that the usages are generally consistent with the archetypes, but exhibit a weak semantic distinction for the social partition.

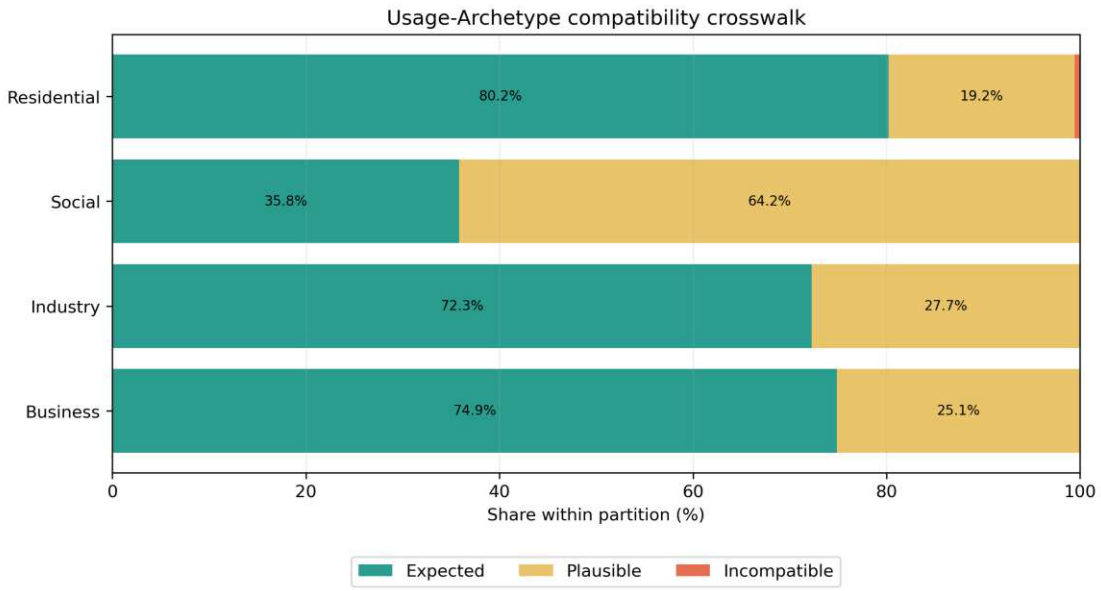


Figure 5.13: Compatibility crosswalk between SmartQ+ usage category and archetype by partition ($n = 3008$) to assess the assignment plausibility.

### 5.5.3 Cluster Stability

For the analysis of the cluster stability, 30 iterations with random subsamplings were carried out, in each of which 90% of the buildings per partition were selected without being returned. The resulting allocations were then compared with the final cluster solution. Additionally, the study area of the municipality was divided into five spatially connected blocks as shown in Figure 5.14 and every block was excluded for exactly one run. The slightly separated western section of the castral area constituted the first block



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

and comprised 540 buildings. The rest of the building stock, which consisted of 2,492 buildings, was divided into four further blocks with 623 buildings each; these blocks are also arranged from west to east. This made it possible to verify how the derived results are preserved when a part of the municipality is missing.

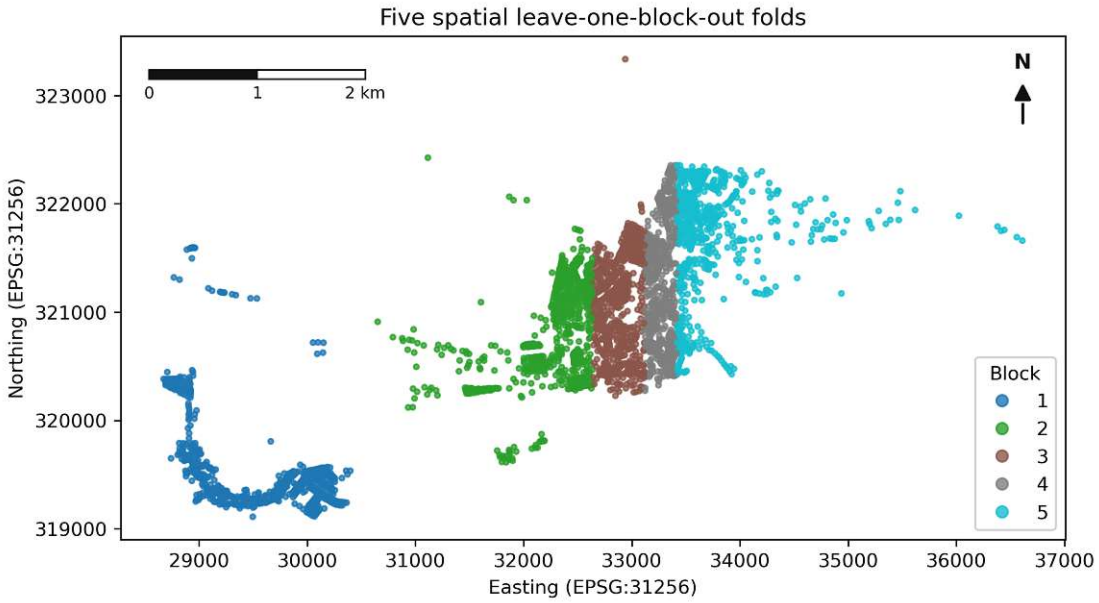


Figure 5.14: Five spatial leave-one-block-out folds. Blocks were formed from west to east.

As a measure of stability, adjusted rand index (ARI) at the level of the cluster labels and the archetypes was calculated to assess whether buildings in two solutions are grouped together or separately and to correct for randomly expected matches. Table 5.18 summarizes the median results for the four partitions. The very similar ARI-values for stage-1 show that the coarse structure in both analyses remained stable. For the cluster label, the ARI was 0.819 for the random subsampling and 0.802 for the spatial leave-one-out analysis across all four partitions. At the archetypal level, the respective ARI values were at 0.749 and 0.71. At the same time it is apparent that in both analyses more than 90% of the buildings kept their archetype allocation.

Table 5.18: Median stability across the four partitions.

| Scheme | Runs | Stage-1 ARI | Cluster ARI | Archetype ARI | Assignment agreement |
|---|---|---|---|---|---|
| Random 90% | 30 | 0.998 | 0.819 | 0.749 | 0.909 |
| Spatial LOO | 5 | 0.994 | 0.802 | 0.710 | 0.907 |

Thus, the results of both approaches point to a stable overall structure, although the finer boundaries of the clusters and archetypes were more responsive to changes in the building stock. Since the aggregated ARI values do not allow an assertion regarding individual archetypes, for further analysis, the Jaccard similarity per archetype was



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

assessed. This describes the agreement of the allocation of buildings to the archetypes between the cluster solution and the respective stability analysis. It corresponds to the relation between the intersection and the union of the observed group; the value 1 signals a complete consistency. The results in appendix Archetype Stability Validation show lower medians and thus lower stability, for the archetypes A04, A10, A13 and A14. This shows that the uncertainties are concentrated around single boundaries and do not affect the entire structure of the archetypes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

# Energy Prediction Model

This chapter describes the development and validation of a supervised regression model for predicting building-level energy metrics. The model is used as a statistical surrogate for the simulation results of the SmartQ+ research project [BSR+25]. Thereby, it should learn the relationship between known building features and the simulated energy values. An ablation experiment was conducted in which the model was trained once using the archetypes (see Archetype Catalog) and once without them. After the subsequent fine-tuning of the model, the model predictions were evaluated to answer research question RQ2. Finally, the spatial indices were derived to make statements on municipality level based on the archetypes and predictions.

## 6.1 Training of Prediction Model

### 6.1.1 Experimental Design

The regression task was split into a total of four separate submodels instead of training a multi-output model. This is due to, for example, a part of the building stock not being heated and thus purposefully excluded from the heating target value, or that 29 buildings did not have a complete annual electric profile. Based on these different building cohorts as well as different feature sets and units regarding the target values, the decision was made to train a separate CatBoost regressor per target variable.

As target variables, the annual energy intensities were chosen. Thereby, the yearly aggregation matches the building-specific and planning focus of this thesis: the forecast of summarized energy indicators for a simulated reference year. Additionally, the yearly values can be aggregated to a grid or municipality level and are thus suitable for the targeted spatial analyses. Through the choice of intensity instead of an absolute yearly value, the influence of the building size is decreased, enabling a better comparison across



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the heterogeneous building stock. To reach this, the simulated yearly values were divided by the approximated GFA of each building [DQ18]:

$$y_{i,t} = \frac{E_{i,t}^{\text{annual}}}{A_i^{\text{GFA proxy}}} \tag{6.1}$$

Here, $y_{i,t}$ refers to the yearly intensity of the target variable $t$ for a building $i$. $E_{i,t}^{\text{annual}}$ represents the respective annual value from the simulation and $A_i^{\text{GFA proxy}}$ the approximated value of the GFA. The resulting target and prediction values are thus given in the respective unit per square meter and year.

The determination of the target variables for the model was first done based on the contextual meaning and the available data. Thereby, first it was ensured that ideally different aspects of the simulated energy system are depicted, which have their own meaning for the planning context. Based on this, the following selection was made: *Primary Energy* for the additive energy demand, *Heating* for the thermal heating energy demand, *$CO_2$* represents the resulting emissions and *Electricity* describes the annual electricity demand. The values for primary energy, heating and $CO_2$ were available as annual values at building level, while the annual electricity value was aggregated from hourly load profiles, whereby only complete profiles were considered. Cooling was not chosen as a target because only 184 buildings had a simulated value for it, which was considered as insufficient coverage. The temporal representation of load profiles was not used as a target, because this would have required separate time-dependent predictions.

Subsequently, a variance test was administered to assess the suitability of the chosen targets. A target variable without or with only low variation enables barely differentiable model predictions. As shown in Table 6.1, for each of the four chosen targets, noticeable differences between the buildings could be noticed, so that none could be excluded based on the insufficient variation of the values. It is important to note, however, that the variances cannot be directly compared with one another due to the different units and scales; in this case, the coefficient of variation (CoV) is more meaningful, as it illustrates the relative dispersion.

Table 6.1: Descriptive statistics and variance of the annual energy-intensity targets.

| Target | $n$ | Mean | Standard deviation | Variance | CoV |
|---|---|---|---|---|---|
| Heating intensity | 2,929 | 57.889 | 25.270 | 638.557 | 0.437 |
| Primary-energy intensity | 3,032 | 103.028 | 45.104 | 2,034.391 | 0.438 |
| $CO_2$ intensity | 3,032 | 18.368 | 9.044 | 81.793 | 0.492 |
| Electricity intensity | 3,003 | 18.492 | 26.645 | 709.965 | 1.441 |

The respective units for the chosen target variables resulted from their content-specific meaning. Therefore, the intensities for heating, primary energy and electricity, calculated using Equation 6.1, are expressed in kWh/(m$^2$ a), while for $CO_2$ kg/(m$^2$ a) applies.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The choice of a model for the regression was CatBoost. As mentioned in Section 2.3, this approach is well-suited to tabular data and also enables a direct processing of categorical features without an upstream one-hot encoding. The latter play a special role in the forecasts, since factors such as the usage group, the type of heating system and especially the archetypes are present as such. Additionally, the internal coding is designed in a way that the targets of observations for the internal coding can be excluded, through which leakage can be counteracted [PGV+18]. In addition, the thought that between the building features a non-linear relation could be expected was essential, which CatBoost, as a Gradient-boosting measure, can also capture.

The influence of the archetypes on the predictions was tested by training two model variants:

**M0 – base model:** uses the basic building features and GIS data, as well as the heating system, depending on the target

**M1 – archetype model:** uses the full feature-space of *M0* plus the archetype label

For both models the exact same target values, cohorts, hyperparameters as well as validation steps were applied. This way, every difference between the prognoses of both models is directly traceable back to the inclusion of the archetype label for *M1*. This is an operational reduction, since the archetype condenses the morphological and spatial-structural information that was used for the derivation. The detailed clustering-features in the models are missing on purpose, because else both models would have the fine-grained information of these available and the implicit information of the archetypes would be negligible. Additionally, through this it is possible to make predictions for unknown buildings without having to calculate the features of the feature engineering from Chapter 4.

The full feature sets of both model variants *M0* and *M1* are depicted in Table 6.2. All listed data was used for the regression of the four target variables, with the exception of `Heating system`, which was only used for $CO_2$, primary energy and electricity. The reason for leaving out this feature for the heating model is that the thermal heating energy demand should be independently predicted from the used heating system. `Construction period` was purposefully kept as a sorted numerical value so the decision tree could form temporally connected threshold values. Therefore, the model can distinguish newer building periods such as 8 and 9 from older building periods such as 3. This ability is important regarding isolation and glazed areas of buildings, which in turn are energetically relevant building features. For consecutive construction periods, these characteristics are, on average, more similar than for construction periods that are further apart in time. Figure 6.1 shows how the building period in the yearly heating intensity based on the simulation results is reflected.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 6.2: Feature sets of the *M0* and *M1* prediction model variants.

| Feature | Description | Type | Model variants | Targets |
|---|---|---|---|---|
| Construction period | Ordered construction-period code from 1 to 9 | Numerical, ordinal | *M0* & *M1* | All |
| Footprint area | Building footprint area in $m^2$ | Numerical | *M0* & *M1* | All |
| GFA proxy | Approximate GFA calculated from footprint area and number of floors | Numerical | *M0* & *M1* | All |
| Number of floors | Number of building floors | Numerical | *M0* & *M1* | All |
| Building height | Max. building height in $m$ | Numerical | *M0* & *M1* | All |
| Grouped use category | Broad functional category | Categorical | *M0* & *M1* | All |
| Monument protection | Indicates whether the building is protected as a monument | Categorical | *M0* & *M1* | All |
| Main-building status | Distinguishes between main and side buildings | Categorical | *M0* & *M1* | All |
| Heating system | Assigned heating-system category | Categorical | *M0* & *M1* | Primary energy, $CO_2$, & Electricity |
| Archetype label | Archetype assignment | Categorical | *M1* only | All |

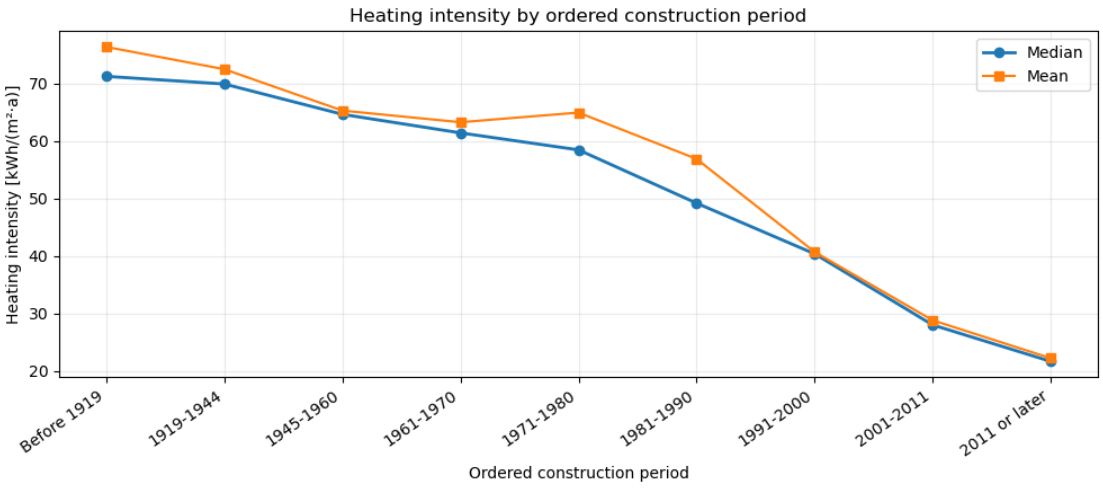


Figure 6.1: Simulated annual heating intensity by ordered construction period.

### 6.1.2 Baseline Training Workflow

Before the base model could be trained, the features had to be processed and checked. For this, a database table was created, which unified the aggregated annual target values as well as the derived intensities and served as the data source for the training. The resulting table was checked by a specified data contract, which contained the inspection of the 3,032 observations and the required data columns for each observation. The features were



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

first checked for missing values and, if needed, marked as `missing`. Additionally, `GFA proxy` and `footprint area` had to contain positive values. For `Construction period` it was checked that only numerical values between 1 and 9 were present. The categorical features had to be converted into strings for CatBoost. Because the model was trained on the original scales of feature and target variables, no standardization, outlier-clipping or imputation was done.

Based on the formed quality areas, subsequently, the target-specific building cohorts were formed. For the heating model, only 2,929 buildings with a known heating system were considered. For electricity, the 3,003 buildings with a complete annual load profile were used. For the other two target values the cohorts included the entire building set with 3,032 entries.

Subsequently, for each target value two CV schemes with five folds each were created. The first schema was a spatial CV. The five folds corresponded to the spatial blocks, which have already been used for the stability check of the archetypes (see Figure 5.14). In each of the five iterations, one of these blocks was completely left out for the training and was instead used as a training dataset. This way, each building was exactly once predicted by a model, which was not exactly trained on that particular building. As a second schema, a random CV was carried out. There, the buildings were distributed into five folds based on their archetype-label in order to reach an even distribution of the archetypes. The archetype information was also used for the distribution in the model variant *M0* while it was not used as an input feature for the training.

For both variants, within each target variable and every cross-validation schema, the same training and test buildings were used to maintain the comparability between *M0* and *M1*. Both CV approaches answered two different questions. The spatial cross-validation checks the robustness and shows how the model can be transferred to a left-out partial area of the municipality. The random cross-validation, however, answers how well the model can interpolate buildings of the same municipality; through the mixing based on the archetypes, spatially and structurally similar buildings could be present in the training and test data [RBC+17]. From the combination of the two model variants, four target variables, two CV schemes and five folds, respectively a CatBoost regressor was trained. The full baseline-training workflow thus included a total of 80 CV fits. These did not depict 80 different model approaches, but rather repeated training of the same configuration to different building cohorts and fold assignments.

The hyperparameters of the baseline configuration, which are listed in Table 6.3, were all kept identically across all 80 fits. As loss function, RMSE was chosen, since the four target values are continuous numerical values and based on the squaring in the function, larger prediction errors were weighted more strongly. Additionally, RMSE equals the default loss function of the CatBoost regressor. For each fit 1,000 boosting iterations with a learning rate of 0.03 were carried out. In every iteration the ensemble was supplemented with a further decision tree. The learning rate basically defines an adaptation rate of an individual tree, which is how the steps between the trees can be



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

controlled; thereby the choice of a smaller rate across 1,000 iterations forms a relatively conservative approach [Catb].

The tree depth was set to 6, allowing each tree to represent several non-linear allocations between the features. For symmetric CatBoost trees, this depth corresponds to a maximum of $2^6 = 64$ leaves. Additionally, `l2_leaf_reg` was set to 5 as L2-regulation, as a result of which very strong influences of individual leaves on the predictions were punished and thus should be prevented. Additionally, the baseline approach does not use an early-stopping mechanism, so for every fit all 1,000 iterations were performed. Thus, the test-fold was also not used for the determination of the number of trees [Cata].

Table 6.3: Hyperparameters of the baseline CatBoost configuration.

| Hyperparameter | Baseline value |
|---|---|
| Loss function | RMSE |
| Number of iterations | 1,000 |
| Learning rate | 0.03 |
| Tree depth | 6 |
| L2 leaf regularization | 5.0 |

The primary output of the model consists of out-of-fold predictions for the four annual energy-intensity targets for each building. These predictions were generated by a model, which was not used for the training of the respective building. For each building additional information such as the simulated energy intensities, the model variant, the cross-validation scheme and the fold-affiliation was stored.

Based on the comparison between the simulation-based value and the predicted value, for every building the absolute deviation $|y_{i,t} - \hat{y}_{i,t}|$ was calculated. The MAE defined in Equation 2.7 matches the mean of these deviations across all $n$ buildings of the respective cohort. Additionally, RMSE was determined as another margin of error and the coefficient of determination $R^2$ as a relative quality measure of prediction performance.

### 6.1.3 Model Fine-Tuning

After the baseline configuration provided acceptable results, the fine-tuning of the CatBoost model was purposefully kept to a minimum. This is because the focus was on archetype ablation rather than, for example, fine-tuning the prognostic model for a specific municipality. Thus, no extensive hyperparameter tuning was performed, but rather a relatively small set of plausible configurations was tested and checked regarding the respective effects. Only configurations whose spatial out-of-fold MAE did not exceed the respective baseline value in any of the eight combinations of four targets and two model variants were accepted. If one of the values did already fall through, then the execution was stopped and the results were not put into the database. The improvement of *M0* to *M1* was not used, which is how it could be avoided that the tuning favored the answering of RQ2.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

In a first run, the number of iterations was set to 2,000. Because this already led to a better result, based on that, three further configurations were tested: one with ordered boosting, one with a higher tree depth and increased L2-regularization of 10 and lastly an iteration with 1,500 and an increased learning rate (0.04). As shown in Table 6.4, the last run with ID 5 scored the best results. In the further course of this work, the results of run 5 were used for the derivation of the spatial indices and the validation.

Table 6.4: Comparison of the evaluated CatBoost configurations based on MAE.

| Run | Boosting | Iterations | Learning rate | Depth | L2 | MAE ratio | Decision |
|---|---|---|---|---|---|---|---|
| 1 | Plain | 1,000 | 0.03 | 6 | 5 | 1.0000 | Baseline |
| 2 | Plain | 2,000 | 0.03 | 6 | 5 | 0.9439 | Accepted |
| 3 | Ordered | 2,000 | 0.03 | 6 | 5 | 1.0679 | Rejected |
| 4 | Plain | 2,000 | 0.03 | 7 | 10 | 0.9647 | Accepted |
| 5 | Plain | 1,500 | 0.04 | 6 | 5 | 0.9424 | Selected |

## 6.2 Evaluation of the Prediction Model

As described in the previous section, run 5 from Table 6.4 was chosen as the final configuration of the model, which is why the validation refers to the out-of-fold prognoses which came from run 5. The models were not evaluated on their training data, because this could lead to an optimistic assessment of their prediction quality [JWH+21]. The focus of the validation is the general validity and robustness of the predictions as well as the influence through the archetypes as an additional input feature.

The cross-validation is based on two different schemes, as described in Section 6.1. For the validation, primarily the spatial CV was used because it examines how well the model can be transferred to an excluded area. The deciding factor is that the influence of spatially similar observations in the training and the test data is reduced [RBC+17].

### 6.2.1 Model Prediction Accuracy

#### Overall Prediction Accuracy

For the evaluation of the general accuracy of the predictions, first the forecasts of *M1* reflected upon. In addition to the introduced metrics from Section 2.3, the normalized mean absolute error (NMAE) was calculated:

$$\mathrm{NMAE}_t[\%] = \frac{\mathrm{MAE}_t}{\bar{y}_t} \cdot 100 \tag{6.2}$$

The MAE is normalized for every target value $t$ through the mean of the respective target cohort $\bar{y}_t$. The NMAE enables a direct comparison of the target values which are based on different scales [MMS+24]. The NMAE values within the spatial cross-validation were between 5.18% for heating and 6.57% for electricity. At the same time the values of $R^2$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

were between 0.911 and 0.982, whereby these values were affected by the variation of the respective target value and thus should not be used as quality comparison between the different targets. The metrics of the spatial out-of-fold predictions are summarized in Table 6.5. MAE and RMSE are represented in the unit of the respective target value, while $R^2$ is essentially unitless. The NMAE is expressed as a percentage due to the multiplication by 100 in Equation 6.3.

Table 6.5: Spatial out-of-fold prediction accuracy of the selected *M1* configuration. MAE and RMSE are expressed in the target-specific unit.

| Target | $n$ | MAE | RMSE | $R^2$ | NMAE [%] |
|---|---|---|---|---|---|
| Heating | 2,929 | 3.00 | 5.58 | 0.951 | 5.18 |
| Primary energy | 3,032 | 6.14 | 13.44 | 0.911 | 5.96 |
| $CO_2$ | 3,032 | 1.15 | 2.54 | 0.921 | 6.27 |
| Electricity | 3,003 | 1.22 | 3.53 | 0.982 | 6.57 |

In the random CV the results across all four target variables were better, as a result of which the expectation of a stricter test of the transferability through the spatial CV was fulfilled. In total, the NMAE values from 5.18% to 6.57% as well as the $R^2$ values of at least 0.911 point to a high prediction quality. The relatively high RMSE values, however, show that for individual buildings larger deviations occur.

### Paired M0–M1 Accuracy Comparison

To assess the influence of the archetypes on the predictions, the model variant *M0* served as a reference for the pairwise comparison. In the process the MAE was used as a primary comparison metric. For every building, the absolute error of *M0* and *M1* within the same folds was compared. Subsequently, the relative improvement of the MAE for every target cohort could be calculated in the following manner:

$$\Delta\text{MAE}_t = \frac{\text{MAE}_{t,M0} - \text{MAE}_{t,M1}}{\text{MAE}_{t,M0}} \cdot 100 \tag{6.3}$$

A positive value thus means that *M1* exhibits fewer errors than the reference model *M0*. As shown in Figure 6.2, the archetypes led to an improvement of the MAE for all target values in both CV schemes. However, depending on the target variable, only 51.81% to 55.65% of the buildings exhibited a lower absolute error at the building level. This thus means that the advantage did not affect all buildings, but was positive across all four target cohorts. Similarly, the relative change of the RMSE was calculated. Figure 6.2 also shows that *M1* exhibits a lower RMSE for the target cohorts across both CV schemes. Only electricity formed an exception in the spatial version: the RMSE improved from 3.39 at *M0* to 3.53 at *M1*, which equals a relative worsening of 4.03%. At the same time, the MAE for the same schema lessened by 3.03%. This suggests that the archetypes lowered the absolute error on average, but individual larger deviations were stronger.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Additionally, a noticeable high improvement of the RSME for heating in the random cross-validation was detected.

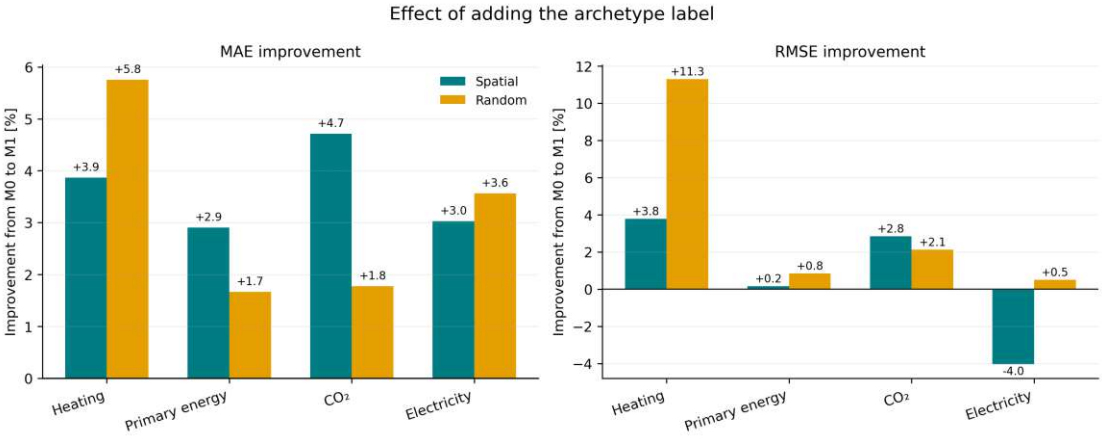


Figure 6.2: Relative changes in MAE and RMSE of all four targets with included archetype label.

The complete comparison of the aggregated metrics are found in appendix Out-of-Fold Prediction Diagnostics. There also figures for the representation of the out-of-fold predictions of the model variants *M0* and *M1* for both cross-validation schemes can be viewed.

### 6.2.2 Robustness of the Archetype Effect

For the review of the robustness of the model, the MAE comparison was additionally done for the individual folds of all targets. This way it could be checked whether the improvement was consistent across the entire fold-composition. As shown in Figure 6.3, for 17 out of 20 folds relative MAE improvements from *M0* to *M1* were scored within the spatial CV. Also for the random version 17 out of 20 folds were improved, but the distribution is different. Positive values here show a relative improvement expressed as a percentage, while negative values for a fold mean a corresponding worsening. The dots represent the single folds while the black diamonds symbolize the mean of the out-of-fold prognoses of the respective target variables.

It can be noted that, for $CO_2$, all ten folds showed an improvement regarding MAE. For heating this was the case in nine folds, for primary energy in eight folds and for electricity still in seven folds. Regarding the target variables and folds a performance plus could be observed.

In the next step, the effect of the archetype labels was also checked at the archetype level. For this, the relative change of the spatial out-of-fold MAE values was calculated independently for each archetype. Figure 6.4 shows that a total of 37 of the 56 combinations of archetypes and target variables are positive. Additionally, the distribution of the



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

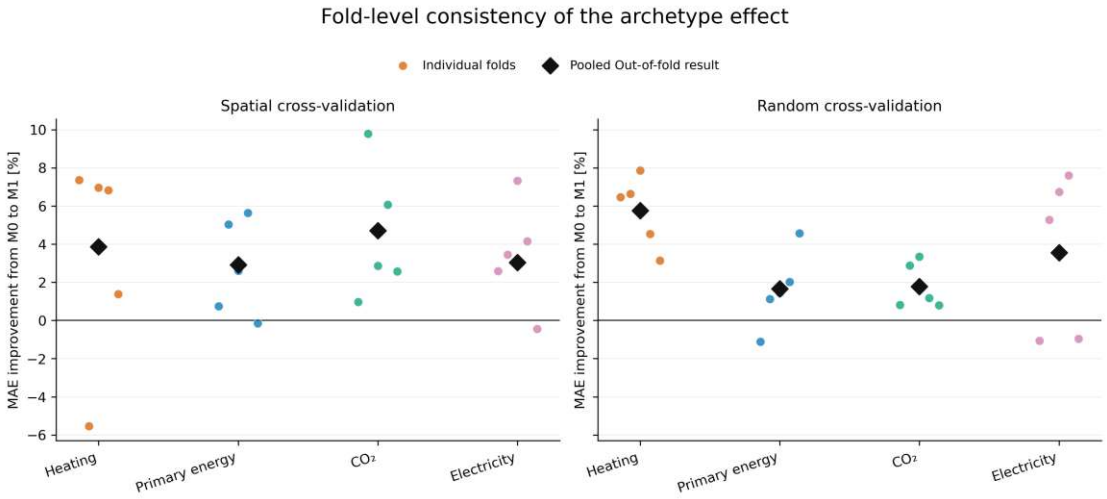


Figure 6.3: Fold-level consistency of the relative MAE improvement from model *M0* to *M1*. Dots represent individual folds, while black diamonds show the pooled out-of-fold result for the respective target.

negative values is observable, whereby it is noticeable that strongly negative values are mostly in smaller cohorts with less than 45 buildings. An example for that is archetype A04, which is negative across all targets, but also underrepresented regarding the building numbers. Groups with less than 45 buildings were marked with a star symbol, because these could only be interpreted descriptively.

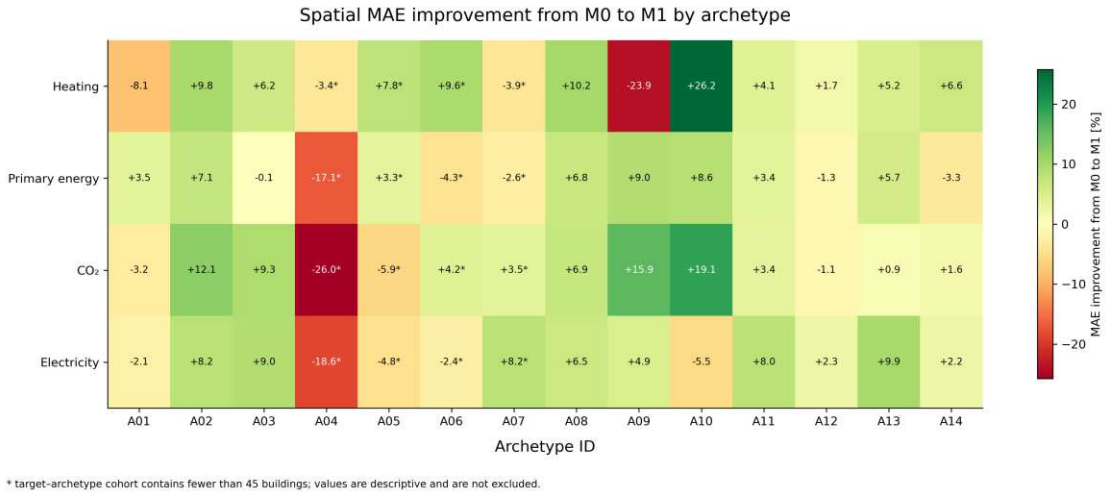


Figure 6.4: Archetype-specific relative changes of spatial MAE from *M0* to *M1*. Positive values indicate a performance gain in MAE, while negative values indicate a higher MAE.

The results across the folds and archetypes thus suggest that the inclusion of the archetypes



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

mostly leads to a robust advantage, especially for the adequately represented groups. At the same time, it is also noticeable here that the advantage is not applicable to all archetypes or even buildings.

### 6.2.3 Error Dispersion and Uncertainty

Since the CatBoost model only creates point predictions for the selected loss function, no individual prediction intervals were present. This is why the uncertainty was examined based on the spatial out-of-fold errors. For this purpose, it was examined how large the absolute deviations between the simulation-based and the predicted models were. For the comparison after target variables the errors were normalized through the mean of the target cohorts. In Figure 6.5, the curve shows, for each error value, the proportion of buildings which did not exceed this value. The curve of *M1*, which is far higher over wide ranges, shows that larger building proportions were under the respective error threshold, which confirms the observations regarding a lower MAE compared to *M0*.

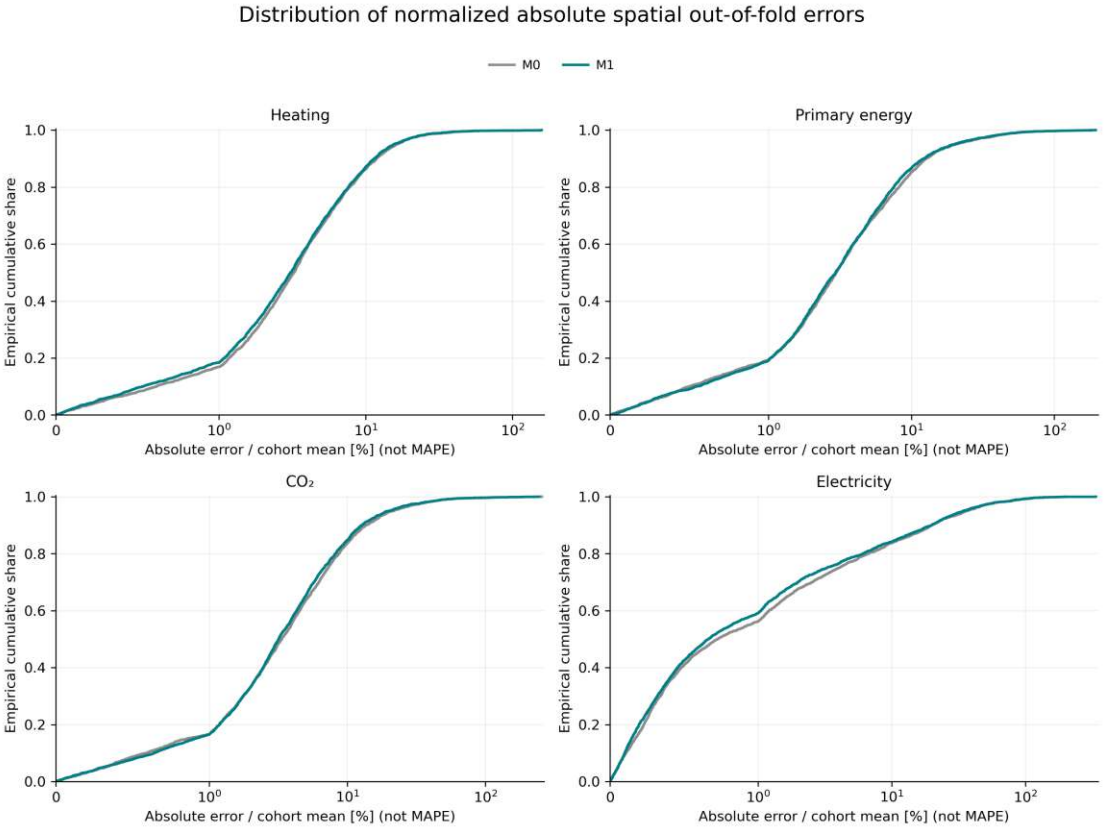


Figure 6.5: Cumulative distributions of the normalized absolute spatial out-of-fold errors for *M0* and *M1*.

The underlying evaluation showed that for *M1* around 95% of the absolute errors were



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

below 8.98 for heating, 20.41 for primary energy, 3.65 for $CO_2$ and 6.38 for electricity, in the respective unit. However, compared to *M0*, certain errors were larger. An example for this would be the maximum error for electricity, which increased from 36.84 in *M0* to 62.61 in *M1*. This observation matches the increased RSME of *M1*, which weights larger errors more strongly.

Additionally, the spatial distribution of the out-of-fold errors for the research area was analyzed. The spatial diagnosis in Figure 6.6 provides insight on areas with a high uncertainty, which were defined as an evaluation metric through the MAE. For this, the absolute error of *M1* was error-weighted and transferred to a 250-m grid. Tiles, which contained less than five buildings, are shown in a shaded manner. Here it shows that the increased MAE focuses on partial areas.

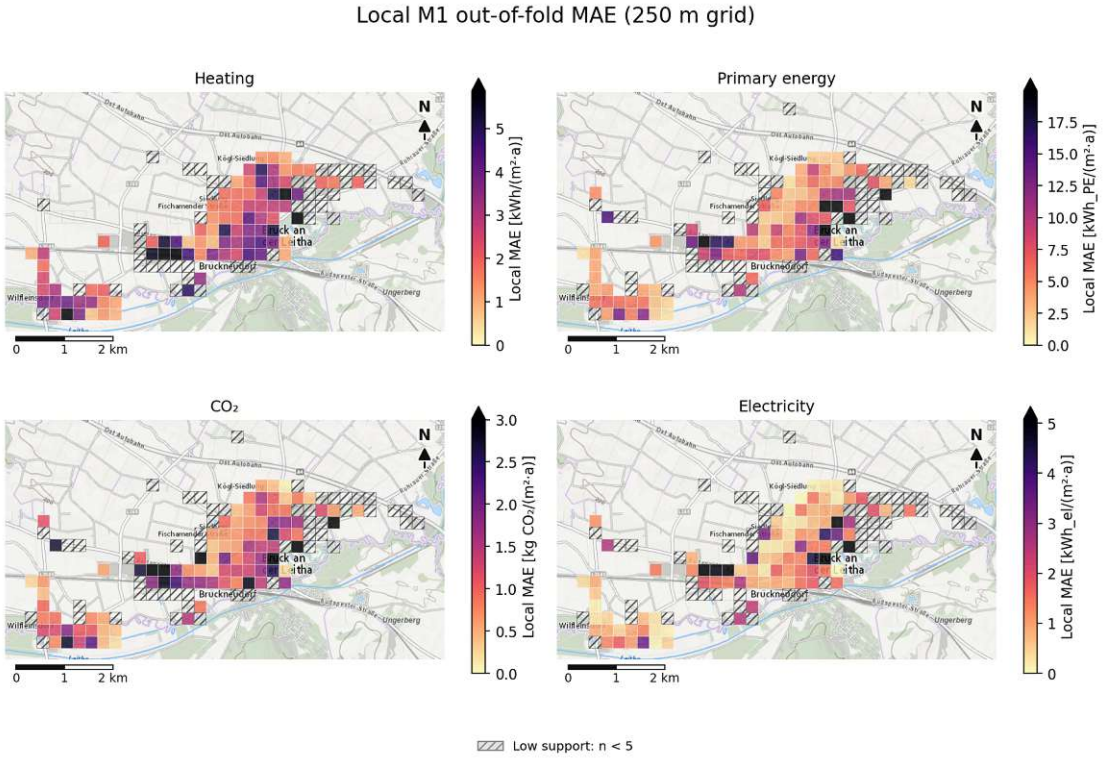


Figure 6.6: Local spatial out-of-fold MAE of *M1* on a 250-m grid; darker cells indicate higher prediction errors. Hatched cells contain $< 5$ buildings.

*Source:* Own representation based on basemap from Geoland.at [Geo26].

A spatial comparison of the two model variants showed an even distribution of the effect between the archetypes. Depending on the target metric, between 58.8% and 64.3% of the grid cells with at least five buildings achieved a lower local MAE. The figure is provided in appendix Spatial Out-of-Fold Error Diagnostics, with the exact shares of the grid cells summarized in Table 3.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

## 6.3 Derivation of Spatial Indices

In this section, the derivation of planning-relevant spatial indices on a 250-m grid level is described. Through this, the spatial analysis of the model prediction on the municipality level should be enabled, which, in turn, depicts the application for urban and spatial planning. For this, among other things, the predicted energy intensities and archetypes, but also the building features such as the GFA, the construction period and the geometries were used. The grid cell size of 250 meters is a trade-off between spatial detail depth and the number of buildings per grid cell. Across the 150 occupied grid cells, the average building count was 23.51, while 34.7% of the cells contained fewer than five buildings; these were visually indicated by hatching in the following figures. Buildings that could not fully be assigned to a cell were assigned based on their proportion of area.

Based on these building area shares, the GFA and the predicted energy intensities for the corresponding grid cells were calculated. The spatial depictions of the derived GFA-weighted energy intensities from that are shown in appendix Spatial Indices for the four target variables. Additionally, for each target value an annual load density was determined, which described the spatial concentration of the energy and emission load per grid cell. As shown in Figure 6.7, for example, high heat load densities are concentrated around the central residential area of the researched area.

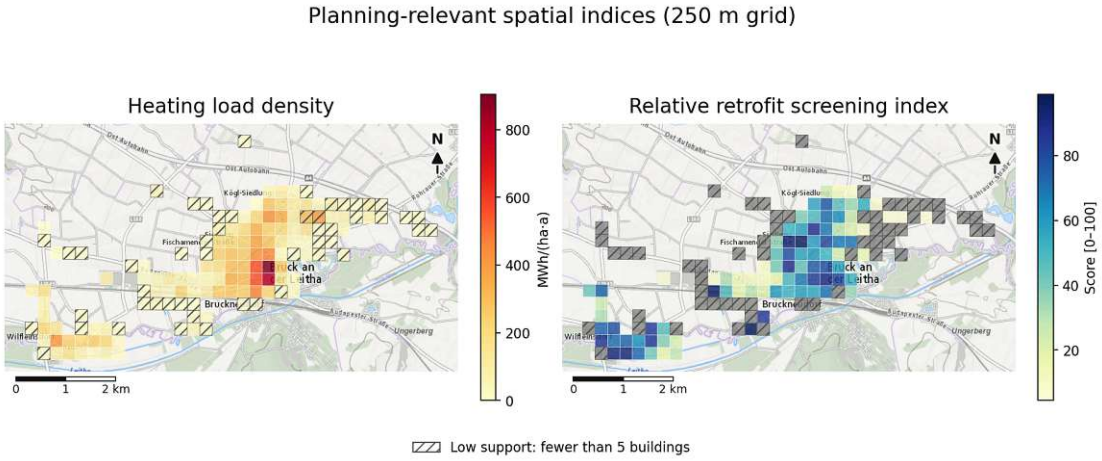


Figure 6.7: Spatial distribution of annual heating load density (left) and relative RSI (right) on a 250-m grid. Hatched cells contain $< 5$ buildings.

*Source:* Own representation based on basemap from Geoland.at [Geo26].

As an additional planning indicator, a relative retrofit screening index (RSI) was calculated, which combines the projected heating intensities with the GFA share of heated buildings constructed prior to 1981. Because the two values have different scales, they are transformed and weighted to a percentile range between 1 and 100. A high RSI shows grid cells, in which a high heating intensity, along with a high percentage of older buildings, is present. Through this, potential spatial analyses for the identification and prioritization



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of renovation candidates can be done [GLS25]. The spatial distribution of this screening measure is also depicted in Figure 6.7. Both indicators show complementary spatial perspectives, which are interpreted further in Section 7.3.

For the figures of the spatial building structure, for every grid cell, the dominant archetype with the highest GFA share was determined. Figure 6.8 thus shows which archetypes shape the building stock of the grid in terms of area, through which a characterization of partial areas after use and occupancy type is possible. The average number of buildings per cell varied heavily between the dominant archetypes. Grid cells dominated by the residential archetypes of dense urban townhouses (A13) and detached single-family houses (A08) contained, on average, 43.08 and 35.04 buildings, whereas the areas dominated by large-scale retail (A02) and the large-scale industrial (A03) contained only 9.41 and 6.64 buildings. This reflects the different levels of development density associated with the respective archetypes.

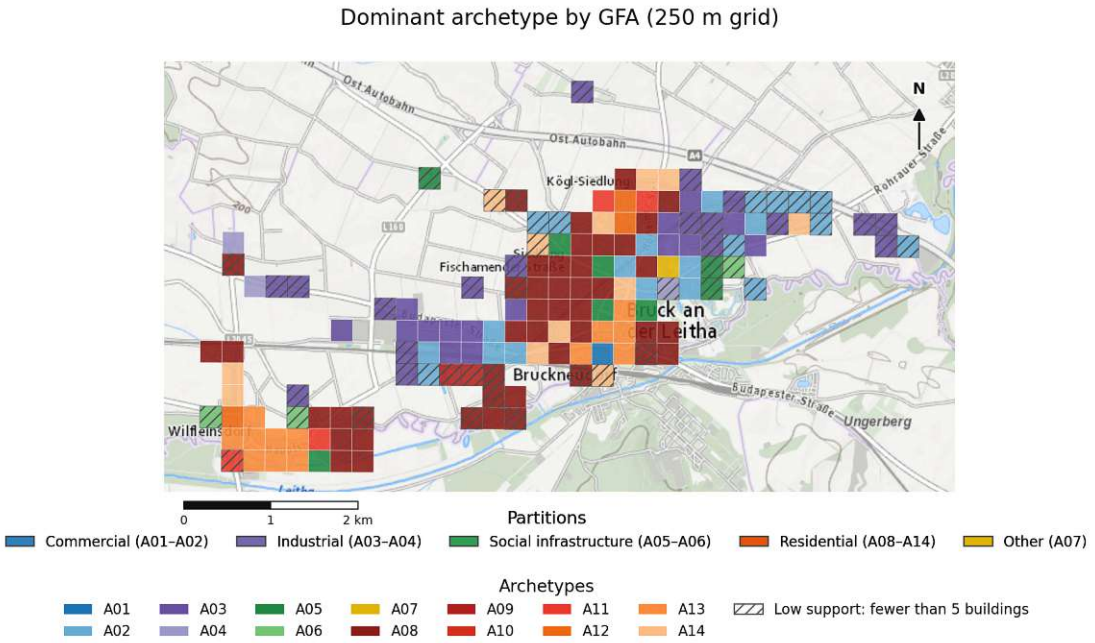


Figure 6.8: Dominant archetype based on GFA in each 250-m grid cell. Hatched cells contain $< 5$ buildings.

*Source:* Own representation based on basemap from Geoland.at [Geo26].

Additionally, from the GFA shares of the archetypes, a normalized archetype mix index was derived, which shows the degree of heterogeneity of the building stock. A value near 0 indicates a one-sided archetype composition, while a value near 1 shows a more diverse mix. Thus, the value is especially relevant for the evaluation, how representative a dominant archetype for the respective grid cell is. The spatial representation is found in appendix Spatial Indices.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CHAPTER 7

# Conclusion

This chapter concludes the pivotal results and observations of this thesis and categorizes them regarding the two research questions RQ1 and RQ2. Therefore, the formed building archetypes and their validation and subsequently their contribution for the energy prognosis model are discussed. Additionally, the implications from the derived planning-relevant indices are described. Finally, the boundaries of the experimental approach as well as possible future further research is mentioned.

## 7.1 Synthesis and Discussion of RQ1

The research question RQ1 assesses how building archetypes can be derived from morphological, functional and spatial-structural features and how stable and interpretable they are within the municipality. For this experiment, the refinement of available GIS data using feature engineering steps (see Section 4.2) in combination with the clustering process from Section 5.3 enabled a technically useful derivation. The clustering process consisted of a partitioning after coarse usage category, a distance calculation using Gower distances and a two-step, weighted, hierarchical clustering. For the clustering, first a flat approach was tested, which mainly represented the structural connection of the buildings. For this reason, the identification of the aforementioned was done in an upstream step to then differentiate based on the morphological and functional features. In this process, a total of 22 clusters were created, which were then mapped on the 14 archetypes (see Archetype Catalog). The final cluster solution had a silhouette of 0.287, which is lower than that of earlier results. However, this trade-off was accepted because, unlike the more coarse solutions, it achieved a typological resolution that was necessary for a domain-specific interpretation. The archetypes are to be understood as data-driven, domain-guided and technically consolidated representations of the local building stock of the researched area.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

For the review of the interpretability, a validation using expert annotation was done. Both experts agreed on 85.1% of the buildings (Cohen's $\kappa = 0.811$); the clustering had an agreement of 84.6% and 87.2% with the annotators ($\kappa = 0.804$ and $\kappa = 0.837$). The model-expert deviations were thus on a similar level as the deviations between the two experts, which entails a technically interpretable classification. The confusion matrices in appendix Archetype Annotation Validation show the archetypes, which had significantly fewer matching results. Here especially the low matching results for the archetypes A10 and A14 are obvious: A14 (apartment block) was often confused with A08 (detached single-family house), while A10 (small building extension/garage) especially had a large deviation to A09 (allotment garden/shed). Such confusions among morphologically similar building typologies show that some boundaries are oversimplified and thus do not capture these fine transitions. The low cluster metrics observed for the expert disagreement annotations further support this interpretation.

The plausibility check based on detailed usage information from the SmartQ+ project aided this analysis, whereby 78.7% of the classifications were identified to be *expected*, additional 20.8% as *plausible* and only 0.5% as *incompatible*; therefore, 99.5% were classified as functionally plausible (see Figure 5.13). The incompatible evaluations are based on buildings with the archetypes A09 and A10 that were classified as multi-family buildings according to *category_tu*. The high share of 62.4% of plausible evaluations shows that archetype A5 and A6 do not directly differentiate the concrete type of usage. However, this share is only limitedly reliable because of the partition size of only $n = 53$.

To evaluate the cluster stability, both a 90%-subsampling as well as a spatial leave-one-block-out analysis were done. The analyses show a robust coarse structure, but also partly sensitive archetype boundaries. In the process, stage-1 cluster labels and archetypes were assessed using ARI, whereby the values for stage 1 were at least 0.994 and at archetype level they sank to a minimum of 0.710, while the assignment agreement for both analyses remained at over 90%. The more detailed analysis of the archetypes using the Jaccard similarity showed that the following archetypes have a high stability: A01, A02, A05, A06, A08, A09 and A11; the archetypes A04, A10, A13 and A14 showed lower medians. This is due to blurred transitions, which were already discussed in relation to the confusion matrices. A09 and A10 contained the value 0 in individual iterations, which indicates a dependency of their boundaries on certain samples or subareas.

In summary, with the help of the developed workflow, a technically plausible and stable archetype catalog can be created. The results show a robust coarse structure with high interpretability and functional plausibility within the municipality. At the same time the reviews show that individual archetypes do not form discriminating building typologies but simplify transitions between structural and functional building forms. The derived archetypes are to be understood as a local representation of the research area, which is why the transferability to other municipalities must be validated separately.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 7.2 Synthesis and Discussion of RQ2

Research question RQ2 answers the examined effect of the inclusion of the archetype labels into the regression model regarding the accuracy, robustness and uncertainty. For this, the two model variants *M0* without archetypes and *M1* with archetypes were trained. The pairwise comparison of the model variants showed a small but consistent improvement of the accuracy. Table 7.1 lists the relative change regarding MAE and RMSE, whereby the MAE improved by 1.66–5.75%. The RMSE improved in seven cases; only for the target variable electricity in the spatial CV, the RMSE decreased by 4.03%. The reductions of the MAE concern between 51.81 and 55.65% of the building, which is why this advantage is to be understood as an improvement of the overall performance and not all individual predictions. Not all buildings benefitting is consistent with the results of RQ1: The archetypes simplify some transitions too strongly and thus do not map all building typologies equally well.

Table 7.1: Relative changes in prediction errors after adding the archetype labels. Positive values indicate an improvement of *M1* compared with *M0*.

| | Spatial CV | | Random CV | |
|---|---|---|---|---|
| Target | ΔMAE [%] | ΔRMSE [%] | ΔMAE [%] | ΔRMSE [%] |
| Heating | 3.87 | 3.78 | 5.75 | 11.29 |
| Primary energy | 2.91 | 0.16 | 1.66 | 0.85 |
| $CO_2$ | 4.71 | 2.84 | 1.78 | 2.13 |
| Electricity | 3.03 | -4.03 | 3.56 | 0.50 |

For the robustness evaluation, the archetype effect was assessed for the individual folds of the spatial and random CV as well as at the archetype level. For the cross-validation folds the MAE of *M1* improved in 17 out of 20 folds, respectively. Figure 6.4 shows the effects of the archetypes on the spatial MAE, whereby in 37 out of 57 cases an improvement was shown. The fine-grained residential archetypes could achieve an improvement in 23 out of 28 cases; the well-represented archetypes A08, A11 and A13 improved across all targets. This shows that the larger number of these building typologies enables a finer and more homogeneous differentiation, whose morphological and spatial-structural features offer an additional predictive use. Several strong declines, however, affect small cohorts with less than 45 buildings, for example A04. The lower robustness in places of the rest of the partitions could thus be related to the smaller building stock and thus limited typological separation.

Noticeable with the target variable heating is the strong decline for A09 by 23.9% and the strong improvement for A10 by 26.2%. The 15 multi-family buildings considered as incompatible in the plausibility check could have strongly affected the heterogeneity of the archetypes A09 and A10. Together with the mix-ups between A09 and A10 seen in Figure 5.11, this served as a plausible explanatory approach for the opposite effect.

The uncertainty of the predictions was assessed across the overall distribution of the



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

out-of-fold errors rather than for individual predictions. This is due to the used regression approach, CatBoost with RMSE loss, only creating point forecasts. Figure 6.5 shows that the error distribution of *M1* is slightly better in some places for all four target variables. At the same time, individual strong outliers increased; especially within the electricity partition the maximum error increased from 36.84 to 62.61, which is compatible with the RSME-decline.

In Figure 6.6 the spatial distribution of the errors of *M1* on a 250-m grid is shown. There it shows that the uncertainties are mostly on the boundaries. In addition, across all four target variables an error hotspot stands out; this encompasses the area "Industriegelände West", which is situated on the western outskirts of the central area of the municipality. There, mostly large-scale buildings of the partitions commercial and industrial as well as other are included. The recurring error hotspot could thus point to the usage profile only being limitedly mapped through the used entry parameters due to a larger functional and energetic heterogeneity. The spatial comparison to *M0* (appendix Spatial Out-of-Fold Error Diagnostics) shows that, depending on the target variable, in 58.8–64.3% of the grid cells a lower local MAE is scored. In the described error hotspot, the archetypes reduce the error for the targets primary energy and $CO_2$; for heating, both improvements and declines are present, whereby the latter is more strongly pronounced. The mostly lower local MAE shows that the archetype effect leads to a widely spread improvement; this observation is supported by the better error spread.

In summary regarding RQ2, the inclusion of the archetype labels leads to a small improvement in the forecast quality. This improvement is mostly robust across all four target variables, both CV variants as well as most folds and spatial subareas, but not for every single prediction. The moderate effect should be viewed in the light of the already high model quality of the surrogate model *M0*. The strongly represented archetypes of the residential partitions benefit especially consistently, while smaller cohorts are subject to larger fluctuations and have a lesser benefit of the archetype labels.

## 7.3 Planning implications

The derivations of the planning-relevant indices transferred the building-related prognoses into spatially aggregated values of the communal spatial and urban planning. The intensities describe the average energy and emission level of the respective local building stock, while the load density shows the spatial concentration of the total load. When combining both indices, for example, partial areas with striking energy or emission patterns can be identified, which can then be used for further planning analyses. Additionally, the model predictions theoretically allow for an initial orientational assessment of scenarios for planned settlement and quarter developments. The RSI marks the cells, in which a high predicted heating intensity meets a high share of buildings with an older building period. Especially in combination with a high load density, this indicator is thus a hint that a high renovation potential is present.

The composition based on the archetype mix index and the dominant archetypes per grid



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

cell may provide insights on the usage characteristics of partial areas. The differences in building densities between dominant archetypes also reflect common variations for the respective spatial structures. This, as well as the interpretability and plausibility checks of the created archetype catalog, support the applicability for the spatial characterization. The additional figures regarding the uncertainties help to estimate the reliability of the forecasts. The indices and predictions of the model thus allow a first spatial evaluation, but are not an evaluation of concrete planning measures and scenarios because they are based on simulation-based data and do not consider the concrete condition of the building stock. Additional methodical limitations are discussed in Limitations.

## 7.4 Limitations

First, all results are limited to the research area, which is due to the simulation-based results as well as the local building typologies and building regulations. The derived archetypes thus are to be understood as a representation of the building stock of the test municipality, Bruck an der Leitha, and not transferable to other municipalities without prior testing. Additionally, the significance of the formed Archetype Catalog is limited due to the scope and the composition of the data basis; this is especially true for the non-residential partitions. The uniform preprocessing of the clustering features could only consider partition-specific requirements to a limited extent and compressed informative heterogeneity within the social infrastructure partition. Statements regarding the transferability of the regression model are not possible due to a missing comparison municipality. Additionally, the absence of forecast intervals means that no coverage probabilities can be specified for the uncertainty of individual predictions. Furthermore, the validations were done using spatial folds, which have also been used for the representation of the results; thus, the results could be slightly optimistic regarding external validation.

## 7.5 Future Work

Future work should apply larger and, if possible, use datasets from more than one municipality, which include all partitions in a sufficient manner. Additionally, the steps for the preprocessing should be checked partition-specifically. Based on these improvements, a more detailed archetype catalog could be developed, which includes more differentiated and homogeneous archetypes across all partitions. Furthermore, through building stocks across municipalities, the cluster stability could be validated in external research areas. An implementation of the annotations with more experts would enable more reliable statements regarding the technical interpretability. For the energy predictions, alternative models and probabilistic loss functions should be researched. Additionally, an implementation of the experiment with measured target values or an external validation would provide additional insights regarding the predictions. Beyond the building-level application, such clustering approaches could also be applied to other spatial planning entities, such as land surfaces or spatial areas, to identify recurring spatial structures.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Overview of Generative AI Tools Used

OpenAI GPT-5 was used as a tool to support brainstorming, conceptual development and implementation. Furthermore, it was used to aid the structuring and wording of text and tables. To review wording, grammar and spelling, Grammarly was used. DeepL was used to support the translation of text passages and technical terms.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Figures


1.1 Research framework linking research approach, objectives and research questions. . . . . 4

2.1 Urban Taxonomy map of Bruck an der Leitha, illustrating the morphological heterogeneity within a rural municipality. . . . . 8

4.1 Example of overlapping building geometries. . . . . 23
4.2 Spatial distribution of the `attachment_ratio`. Blue indicates detached buildings; red tones indicate increasing shares of shared wall length. . . . 24
4.3 Distribution of coverage ratios for building–parcel intersections and corresponding confidence thresholds. . . . . 26

5.1 Distributions of four examined numerical features with pronounced upper tails, including the median and the 95th and 99th percentiles. . . . . 38
5.2 Distribution of `gfa_proxy` before and after $log1p$ transformation, illustrating the reduction of the variable's right-skewness. . . . . 39
5.3 Truncated dendrogram of the hierarchical clustering for the residential partition, based on average linkage. . . . . 42
5.4 Truncated dendrogram of the hierarchical clustering for the residential partition, based on complete linkage. . . . . 42
5.5 Comparison of the baseline clustering runs by total and residential cluster count, median silhouette score and shares of potentially uncertain assignments. 47
5.6 Spatial distribution of the baseline clustering result for run `B7`. . . . . 48
5.7 Spatial representation of median silhouette scores in stage 1 for run `S1_4`. 52
5.8 Spatial distribution of the stage 2 clustering result for run `S2_3`. . . . . 55
5.9 Spatial comparison of the differentiation between row houses and block-edge residential buildings. Both panels show the same spatial extent. . . . . 57
5.10 Assignment overlap between the residential clusters of the baseline solution `B7` and the final solution `S2_w_14`. . . . . 58
5.11 Confusion matrix of archetypes against the expert-consensus subset ($n = 2560$; pre-annotated A07 was excluded). . . . . 65
5.12 Cluster quality metrics by expert annotation agreement. . . . . 65
5.13 Compatibility crosswalk between SmartQ+ usage category and archetype by partition ($n = 3008$) to assess the assignment plausibility. . . . . 66




Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.14 Five spatial leave-one-block-out folds. Blocks were formed from west to east. 67

6.1 Simulated annual heating intensity by ordered construction period. . . . . 72
6.2 Relative changes in MAE and RMSE of all four targets with included archetype label. . . . . 77
6.3 Fold-level consistency of the relative MAE improvement from model *M0* to *M1*. Dots represent individual folds, while black diamonds show the pooled out-of-fold result for the respective target. . . . . 78
6.4 Archetype-specific relative changes of spatial MAE from *M0* to *M1*. Positive values indicate a performance gain in MAE, while negative values indicate a higher MAE. . . . . 78
6.5 Cumulative distributions of the normalized absolute spatial out-of-fold errors for *M0* and *M1*. . . . . 79
6.6 Local spatial out-of-fold MAE of *M1* on a 250-m grid; darker cells indicate higher prediction errors. Hatched cells contain $< 5$ buildings. . . . . 80
6.7 Spatial distribution of annual heating load density (left) and relative RSI (right) on a 250-m grid. Hatched cells contain $< 5$ buildings. . . . . 81
6.8 Dominant archetype based on GFA in each 250-m grid cell. Hatched cells contain $< 5$ buildings. . . . . 82

1 Confusion matrices for all archetypes against the expert-consensus subset ($n = 2560$; pre-annotated A7 was excluded). . . . . 109
2 Archetype-specific Jaccard similarity for random 90% subsampling and spatial leave-one-block-out. Random error bars show the interquartile range, while the spatial error bars show the full range across the five folds. . . . . 110
3 Simulated and spatial out-of-fold predicted annual energy intensities for model variant *M0*. . . . . 111
4 Simulated and spatial out-of-fold predicted annual energy intensities for model variant *M1*. . . . . 112
5 Simulated and random out-of-fold predicted annual energy intensities for model variant *M0*. . . . . 113
6 Simulated and random out-of-fold predicted annual energy intensities for model variant *M1*. . . . . 114
7 Relative MAE change from *M0* to *M1* on a 250-m grid. . . . . 115
8 GFA-weighted spatial energy intensities for the four targets on a 250-m grid. Hatched cells contain $< 5$ buildings. . . . . 116
9 Archetype mix index on a 250-m grid. Hatched cells contain $< 5$ buildings. 116



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Tables


2.1 Selected clustering approaches for archetype formation. . . . . 11
2.2 Overview of selected regression approaches for the prediction of energy indicators. . . . . 14

4.1 SmartQ+ building attributes . . . . . 20
4.2 Morphological building features . . . . . 22
4.3 Building attachment features . . . . . 23
4.4 Parcel related features . . . . . 25
4.5 Neighborhood structure (100$m$ radius) . . . . . 27

5.1 Archetype clustering features. . . . . 32
5.2 Distribution of buildings across the functional partitions used for clustering. 33
5.3 Caption . . . . . 33
5.4 Preprocessing decisions before distance calculation. . . . . 35
5.5 Identified missing values and subsequent handling. . . . . 36
5.6 Distribution indicators and transformation decisions for selected features. 37
5.7 Quality cluster metrics used for evaluating baseline candidate solutions. . 43
5.8 Criteria used for selecting the baseline partition-wise solutions. . . . . 45
5.9 Feature set used for the attachment-based stage 1 clustering. . . . . 50
5.10 Comparison of attachment-based stage 1 clustering solutions. . . . . 51
5.11 Parameters controlling the stage-2 split acceptance logic. . . . . 53
5.12 Fair comparison of stage 2 clustering solutions based on `S1_4`. . . . . 54
5.13 Fair comparison of representative weighted stage 2 clustering solutions. . 56
5.14 Archetype Catalog . . . . . 60
5.15 Mapping of the final cluster labels to the building archetypes. . . . . 61
5.16 Attributes recorded during the expert annotation. . . . . 63
5.17 Agreement between expert annotations and clustering archetypes. . . . . 64
5.18 Median stability across the four partitions. . . . . 67

6.1 Descriptive statistics and variance of the annual energy-intensity targets. 70
6.2 Feature sets of the *M0* and *M1* prediction model variants. . . . . 72
6.3 Hyperparameters of the baseline CatBoost configuration. . . . . 74
6.4 Comparison of the evaluated CatBoost configurations based on MAE. . . 75




Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.5 Spatial out-of-fold prediction accuracy of the selected *M1* configuration. MAE and RMSE are expressed in the target-specific unit. . . . . 76

7.1 Relative changes in prediction errors after adding the archetype labels. Positive values indicate an improvement of *M1* compared with *M0*. . . . . 85

2 Out-of-fold prediction metrics for *M0* and *M1* under both cross-validation schemes. Positive delta values indicate a lower error for *M1* compared to *M1*. 110

3 Share of supported 250-m grid cells with a lower local MAE for *M1* than *M0*. 115

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Glossary

**agglomerative approach** A bottom-up approach used in hierarchical clustering. Each data point is treated as a separate cluster at the start . 41

**building archetype** An interpretable class of buildings which is classified by an recurring combination of it's features . 1, 17, 83

**clustering** An unsupervised learning method for grouping observations based on their similarity and recurring patterns in the data . 7

**dendrogram** A tree-like diagram for visualizing hierarchical clustering. It displays how individual objects or already formed groups are merged step by step based on their respective distance . 11, 41

**energy indicator** A variable describing an energy-related performance or characteristic of a building, such as annual electric energy use . 1

**morphology** The physical form and spatial configuration of buildings and settlements. At the building level this includes features such as the building's shape, height and footprint; at the settlement level it describes characteristics such as the arrangement and density of buildings and built-up areas . 1, 7

**parcel** A legally and administratively defined unit of land, commonly used as basic unit in cadastral and planning contexts . 1

**regression** A supervised machine learning task in which the model predicts a continuous numerical target variable . 13

**surrogate model** A simplified model that approximates the behavior of a more complex model, such as a computationally expensive simulation model . 86



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Acronyms

**ARI** adjusted rand index. 67

**BCR** building coverage area. 25

**CoV** coefficient of variation. 70

**CV** An evaluation approach in which the dataset is divided into multiple folds that are iteratively used for training and testing. This way the performance can be validated on unseen data . 15, 73, 85

**FAR** floor area ratio. 25

**GFA** gross floor area. 22, 70

**GIS** geographic information system. 1, 4, 17

**MAE** mean absolute error. 15, 85

**NMAE** normalized mean absolute error. 75

**NN** neural network. 15

**RMSE** root mean squared error. 15, 85

**RSI** retrofit screening index. 81

**SVR** support vector regression. 15

**UBEM** urban building energy modeling. 9, 17



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Bibliography


[ABT22] Imdat As, Prithwish Basu, and Pratap Talwar. *Artificial intelligence in urban planning and design: Technologies, implementation, and impacts*. Elsevier, 2022.

[AM24] Mousa Alrasheed and Monjur Mourshed. Building stock modelling using k-prototype: A framework for representative archetype development. *Energy and Buildings*, 311:114111, 2024.

[ASH+21] Usman Ali, Mohammad Haris Shamsi, Cathal Hoare, Eleni Mangina, and James O'Donnell. Review of urban building energy modeling (ubem) approaches, methods and tools using qualitative and quantitative analysis. *Energy and Buildings*, 246:111073, 2021.

[ATL18] Salem Alelyani, Jiliang Tang, and Huan Liu. Feature selection for clustering: A review. *Data clustering*, pages 29–60, 2018.

[Aus15] Austrian Energy Agency. Eine Typologie österreichischer Wohngebäude: TABULA/EPISCOPE Gebäudetypologie, 2015. TABULA/EPISCOPE brochure.

[BH07] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[BP98] J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):301–315, 1998.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[BSR+25] Stefan Bindreiter, Yosun Sisman, Lukas Rast, Sabine Sint, Simon Hinterseer, Emanuel Selz, Karin Mottl, and Ralf Roggenbauer. Smartq+ bruck/leitha: Potenziale der quartiersentwicklungsplanung auf dem weg zum plus-energie-quartier. 2025.

[BTBPR22] Patricia Borges, Oriol Travesset-Baro, and Anna Pages-Ramon. Hybrid approach to representative building archetypes development for urban models – a case study in andorra. *Building and Environment*, 215:108958, 2022.




Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Cata] CatBoost Developers. Common training parameters. https://catboost.ai/docs/en/references/training-parameters/common. Accessed: 2026-08-05.

[Catb] CatBoost Developers. Parameter tuning. https://catboost.ai/docs/en/concepts/parameter-tuning. Accessed: 2026-08-05.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[CWJ21] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623, 2021.

[DB04] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

[DCYC22] Zhang Deng, Yixing Chen, Jingjing Yang, and Zhihua Chen. Archetype identification and urban building energy modeling for city-scale buildings based on gis datasets. In *Building Simulation*, volume 15, pages 1547–1559. Springer, 2022.

[DJLS20] Ina De Jaeger, Jesus Lago, and Dirk Saelens. A building clustering approach for urban energy simulations. *Energy and Buildings*, 208:109671, 2020.

[DQ18] Mac Van Dat and Tran Ngoc Quang. A study on energy consumption of hotel buildings in vietnam. *Journal of Science and Technology in Civil Engineering (JSTCE) - HUCE*, 12(5):109–116, Aug. 2018.

[EKS+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[eve11] *Measurement of Proximity*, chapter 3, pages 43–69. John Wiley & Sons, Ltd, 2011.

[Fri01] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[FSB+26] Martin Fleischmann, Krasen Samardzhiev, Anna Brázdová, Daniela Dančejová, and Lisa Winkler. The hierarchical morphotope classification: A theory-driven framework for large-scale analysis of built form. *Cities*, 174:107047, 2026.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

[FSK08] Alexander Forrester, Andras Sobester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

[GBCB16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[Geo25] Geoland.at. Orthofoto Österreich – basemap.at Orthofoto. Web Map Tile Service (WMTS), Layer `bmaporthofoto30cm`, `https://basemap.at/orthofoto/`, 2025. CC BY 4.0; Dienstaktualisierung vom 29.08.2025; Datenstand Flug 2024; accessed: 2026-08-07.

[Geo26] Geoland.at. basemap.at Verwaltungsgrundkarte Raster Österreich – GRAU. Web Map Tile Service (WMTS), Layer `bmapgrau`, `https://basemap.at/grau/`, 2026. CC BY 4.0; Release 63 vom 27.05.2026; accessed: 2026-08-07.

[GLS25] Ezgi Köker Gökgöl, Esther Linder, and Philipp Schuetz. A community-based decision support map for building retrofit towards a more sustainable future. *Energy and Buildings*, 335:115554, 2025.

[Gow71] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[GTTB21] M. Gholami, D. Torreggiani, P. Tassinari, and A. Barbaresi. Narrowing uncertainties in forecasting urban building energy demand through an optimal archetyping method. *Renewable and Sustainable Energy Reviews*, 148:111312, 2021.

[Hen15] Christian Hennig. Clustering strategy and method selection, 2015.

[Hen19] Christian Hennig. *Cluster Validation by Measurement of Clustering Characteristics Relevant to the User*, chapter 1, pages 1–24. John Wiley & Sons, Ltd, 2019.

[HLR+22] Daniel Heidenthaler, Markus Leeb, Patricia Reindl, Lukas Kranzl, Thomas Bednar, and Michael Moltinger. Building stock characteristics of residential buildings in salzburg, austria based on a structured analysis of energy performance certificates. *Energy and Buildings*, 273:112401, 2022.

[HML+25] Tianzhen Hong, Jeetika Malik, Jared Langevin, Zhaoru Liu, and Da Yan. Ten questions on building stock modeling to inform energy efficiency and sustainability. *Building and Environment*, 284:113465, 2025.

[Hua98] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Jai10] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[JWH+21] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning: with applications in R*, volume 2. Springer, 2021.

[KHP20] Martin Heine Kristensen, Rasmus Elbæk Hedegaard, and Steffen Petersen. Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling. *Energy*, 201:117687, 2020.

[KR09] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[Lan25] Land Niederösterreich. Digitale Katastralmappe (DKM): Parcel geometries. Data obtained via the NÖ GeoShop, `https://geoshop.noel.gv.at/`, 2025. Last accessed: 2026-08-07.

[LG19] Dongmeng Liu and Jinko Graham. Simple measures of individual cluster-membership certainty for hard partitional clustering. *The American Statistician*, 73(1):70–79, 2019.

[LHBR23] Z Le Hong, Z Berzolla, and C Reinhart. The more the better? archetype segmentation in urban building energy modelling. *Journal of Physics: Conference Series*, 2600(8):082004, nov 2023.

[LSD16] Tobias Loga, Britta Stein, and Nikolaus Diefenbach. Tabula building typologies in 20 european countries—making energy-related features of residential building stocks comparable. *Energy and Buildings*, 132:4–12, 2016. Towards an energy efficient European housing stock: monitoring, mapping and modelling retrofitting processes.

[MH26] Meshkinkiya, Maryam and Hashemi, Farzad. Beyond numerical features: Mixed-type clustering of urban building archetypes using hierarchical and k-prototypes methods. *E3S Web Conf.*, 716:10014, 2026.

[MMS+24] Jihoon Moon, Muazzam Maqsood, Dayeong So, Sung Wook Baik, Seungmin Rho, and Yunyoung Nam. Advancing ensemble learning techniques for residential building electricity consumption forecasting: Insight from explainable artificial intelligence. *PloS one*, 19(11):e0307654, 2024.

[NSM+23] Bibhu Kalyan Nayak, Sunil K. Sansaniwal, Jyotirmay Mathur, Tarush Chandra, Vishal Garg, and Rajat Gupta. A review of residential building



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

archetypes and their applications to study building energy consumption. *Architectural Science Review*, 66(3):187–200, 2023.

[OAAS+22] Razak Olu-Ajayi, Hafiz Alaka, Ismail Sulaimon, Funlade Sunmola, and Saheed Ajayi. Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, 45:103406, 2022.

[PGV+18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[PWK19] Oleksii Pasichnyi, Jörgen Wallin, and Olga Kordas. Data-driven building archetypes for urban building energy modelling. *Energy*, 181:360–377, 2019.

[PZ20] Yue Pan and Limao Zhang. Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy*, 268:114965, 2020.

[QL26] Xiaoyu Qu and Ziheng Liu. Forecasting the total building energy based on its architectural features using a combination of catboost and meta-heuristic algorithms. *Energy & Environment*, 37(1):282–310, 2026.

[RBC+17] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[SEKN17] Gernot Stöglehner, Hans Emrich, Helmut Koch, and Michael Narodoslawsky. Impulse für eine kommunale energieraumplanung. *Hg. v. BUNDESMINISTERIUM FÜR LAND-UND FORSTWIRTSCHAFT. Wien.*, 4:2023, 2017.

[SS04] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[SW24] Pengyuan Shen and Huilong Wang. Archetype building energy modeling approaches and applications: A review. *Renewable and Sustainable Energy Reviews*, 199:114478, 2024.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[TKO+18] Giovanni Tardioli, Ruth Kerrigan, Mike Oates, James O'Donnell, and Donal P. Finn. Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. *Building and Environment*, 140:90–106, 2018.

[TPP25] G. Tocchi, M. Pittore, and M. Polese. Identifying urban and rural settlement archetypes: clustering for enhanced risk-oriented exposure and vulnerability analysis. *Natural Hazards and Earth System Sciences*, 25(9):3665–3692, 2025.

[Was24] Adam Wasilewski. Customer segmentation in e-commerce: a context-aware quality model for comparing clustering algorithms. *Journal of Internet Services and Applications*, 15(1):160–178, Jul. 2024.

[WWG+25] Haolin Wang, Zhi Wu, Wei Gu, Pengxiang Liu, Qirun Sun, and Wei Wang. A gis-informed deep learning framework to assess the effect of urban morphology on energy demand. *Energy and Buildings*, 349:116535, 2025.

[ZGP23] Pan Zhang, Debarchana Ghosh, and Sohyun Park. Spatial measures and methods in sustainable urban morphology: A systematic review. *Landscape and Urban Planning*, 237:104776, 2023.

[Ös17] Österreichische Raumordnungskonferenz. *ÖROK-Empfehlung Nr. 56: Flächensparen, Flächenmanagement & aktive Bodenpolitik – Ausgangslage, Empfehlungen & Beispiele*. Österreichische Raumordnungskonferenz (ÖROK), Wien, 2017. Materialien, Heft 4; Annahme durch die Österreichische Raumordnungskonferenz mit Dezember 2016.

[Ös18] Österreichische Raumordnungskonferenz. *Raumordnung in Österreich und Bezüge zur Raumentwicklung und Regionalpolitik*. Österreichische Raumordnungskonferenz (ÖROK), Wien, 2018.

[Ös21] Österreichische Raumordnungskonferenz. *Österreichisches Raumentwicklungskonzept ÖREK 2030: Raum für Wandel*. Österreichische Raumordnungskonferenz (ÖROK), Wien, 2021.

[Ös23] Österreichische Raumordnungskonferenz. *Bodenstrategie für Österreich: Strategie zur Reduktion der weiteren Flächeninanspruchnahme und Bodenversiegelung bis 2030*. Österreichische Raumordnungskonferenz (ÖROK), Wien, 2023. Entwurf, Stand: Juni 2023.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Appendix

## A Annotation Guidelines

Please open the .qgz file in QGIS to open the Annotation Environment project. The annotation environment is based on QGIS layers, which represent the partitions. Each partition has Archetype labels that can be assigned to only buildings of that partition.

### How to label building archetypes

Please select **one layer** at a time (and also click on **show layer**) and enter **edit mode** for that layer to annotate the buildings within the partition. For the annotation, simply select the building footprints and assign the corresponding archetype from the catalog to the field **expert_label**. This can be done by selecting 'editing attributes of selected objects' to open the edit form where the values for the annotation fields can be set.

**Note:** It is highly recommendend to **set a Shortcut** for **Editing attributes of selected objects**.

After every building of a partition is annotated, please make sure to **save the edits** when leaving the edit mode - **otherwise all annotations will be lost!**

**Note:** Even better, please make sure to **save the edits regularly** (especially for the 'Residential' partition). This can be done by **clicking on the 'Save Edits' button** in the toolbar next to the edit-mode button.

The buildings are grouped into different colors for better visibility of the labeling status of the buildings.

Colorcode:

- **Pink**: No archetype label assigned yet
- **Green**: Archetype label assigned (expert_label field is filled and not 'unknown')
- **Blue**: Archetype label is set to 'unknown'

### Annotation Values

**expert_label** This field should be filled with the corresponding archetype label from the Archetype Catalog.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

**expert_confidence** This optional field can be used to indicate the confidence level of the annotation.

If no confidence value is assigned, the **default is high confidence**. Assignable values are **low confidence** and **medium confidence**.

**Note:** Only assign the confidence value if you are unsure about the accuracy of the archetype label and it is **not high**.

**comment** Furthermore, there is a field for optional comments and notes. Please use this field to note any special cases, uncertainties or observations that may be relevant for the annotation process.

## Further Annotation Rules and Guidelines

### Row-house structures

- The end buildings of row-house structures should be annotated as **semi-detached houses**, but they should still be counted when identifying a row-house structure.
- Row houses do not necessarily need to be attached to other row houses in the context of clustering. They **could be between two semi-detached houses or edge-block buildings**.
- The **separation of row-house structures & block-edge structures** in the clustering context is **mostly based on morphology and size**.

### Incorrect footprints

- For some buildings, the footprints may not be correct (e.g. a row of houses may be represented as one large footprint). In such cases, please **annotate the building based on the footprint/partition** and not orthophotos or other data sources.

## Baselayer

There are two baselayers, which are activated by default:

- **Orthophotos**: Aerial imagery that provides detailed visual information about the buildings and their surroundings.
- **Building footprints**: This layer shows the outlines of the buildings, so also buildings of non-active partitions are visible.

## How to submit the annotations

Please save the QGIS project in the end and share the folder including the .gpkg and .qgz files (can be zipped or not).



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# B Usage-Archetype Compatibility Catalog

| TU code | TU label | Archetype ID | Class | Included |
|---|---|---|---|---|
| 3 | Bürogebäude | 1 | expected | True |
| 3 | Bürogebäude | 2 | plausible | True |
| 8 | Pension | 1 | expected | True |
| 8 | Pension | 2 | incompatible | True |
| 9 | Hotel | 1 | expected | True |
| 9 | Hotel | 2 | plausible | True |
| 13 | Verkaufsstätte | 1 | expected | True |
| 13 | Verkaufsstätte | 2 | expected | True |
| 22 | Industrie und Lagergebäude | 3 | expected | True |
| 22 | Industrie und Lagergebäude | 4 | plausible | True |
| 33 | Tierhaltung/Ställe | 3 | plausible | True |
| 33 | Tierhaltung/Ställe | 4 | plausible | True |
| 34 | Maschinenhalle/ Lagergebäude | 3 | expected | True |
| 34 | Maschinenhalle/ Lagergebäude | 4 | plausible | True |
| 23 | Garage/Gartenhütte | 7 | expected | False |
| 29 | Ruine | 7 | expected | False |
| 30 | Trafos | 7 | expected | False |
| 4 | Kindergarten | 5 | plausible | True |
| 4 | Kindergarten | 6 | expected | True |
| 5 | Höhere Schule | 5 | expected | True |
| 5 | Höhere Schule | 6 | plausible | True |
| 6 | Krankenhaus | 5 | expected | True |
| 6 | Krankenhaus | 6 | plausible | True |
| 11 | Veranstaltungsstätte | 5 | expected | True |
| 11 | Veranstaltungsstätte | 6 | plausible | True |
| 12 | Sportstätte | 5 | expected | True |
| 12 | Sportstätte | 6 | plausible | True |
| 14 | Hallenbad | 5 | expected | True |
| 14 | Hallenbad | 6 | plausible | True |
| 15 | Pflichtschule | 5 | expected | True |
| 15 | Pflichtschule | 6 | plausible | True |
| 16 | Hochschule | 5 | expected | True |
| 16 | Hochschule | 6 | plausible | True |
| 24 | Kirche | 5 | expected | True |
| 24 | Kirche | 6 | plausible | True |
| 25 | Kapelle | 5 | plausible | True |
| 25 | Kapelle | 6 | expected | True |
| 26 | Friedhofbauten | 5 | plausible | True |
| 26 | Friedhofbauten | 6 | expected | True |
| 27 | Freiwillige Feuerwehr | 5 | expected | True |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

| TU code | TU label | Archetype ID | Class | Included |
|---|---|---|---|---|
| 27 | Freiwillige Feuerwehr | 6 | plausible | True |
| 28 | Stadttürme und Stadtbefestigung | 5 | plausible | True |
| 28 | Stadttürme und Stadtbefestigung | 6 | plausible | True |
| 31 | Museum | 5 | plausible | True |
| 31 | Museum | 6 | expected | True |
| 1 | Einfamilienhaus | 8 | expected | True |
| 1 | Einfamilienhaus | 9 | plausible | True |
| 1 | Einfamilienhaus | 10 | plausible | True |
| 1 | Einfamilienhaus | 11 | expected | True |
| 1 | Einfamilienhaus | 12 | expected | True |
| 1 | Einfamilienhaus | 13 | plausible | True |
| 1 | Einfamilienhaus | 14 | plausible | True |
| 2 | Mehrfamilienhaus | 8 | plausible | True |
| 2 | Mehrfamilienhaus | 9 | incompatible | True |
| 2 | Mehrfamilienhaus | 10 | incompatible | True |
| 2 | Mehrfamilienhaus | 11 | plausible | True |
| 2 | Mehrfamilienhaus | 12 | plausible | True |
| 2 | Mehrfamilienhaus | 13 | expected | True |
| 2 | Mehrfamilienhaus | 14 | expected | True |
| 20 | Wohngebäude für Gemeinschaften | 8 | plausible | True |
| 20 | Wohngebäude für Gemeinschaften | 9 | incompatible | True |
| 20 | Wohngebäude für Gemeinschaften | 10 | incompatible | True |
| 20 | Wohngebäude für Gemeinschaften | 11 | plausible | True |
| 20 | Wohngebäude für Gemeinschaften | 12 | plausible | True |
| 20 | Wohngebäude für Gemeinschaften | 13 | expected | True |
| 20 | Wohngebäude für Gemeinschaften | 14 | expected | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 8 | plausible | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 9 | expected | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 10 | plausible | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 11 | incompatible | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 12 | incompatible | True |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

| TU code | TU label | Archetype ID | Class | Included |
|---|---|---|---|---|
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 13 | incompatible | True |
| 32 | Kleingartensiedlung (mit Wohnnutzung) | 14 | incompatible | True |

# C Archetype Annotation Validation

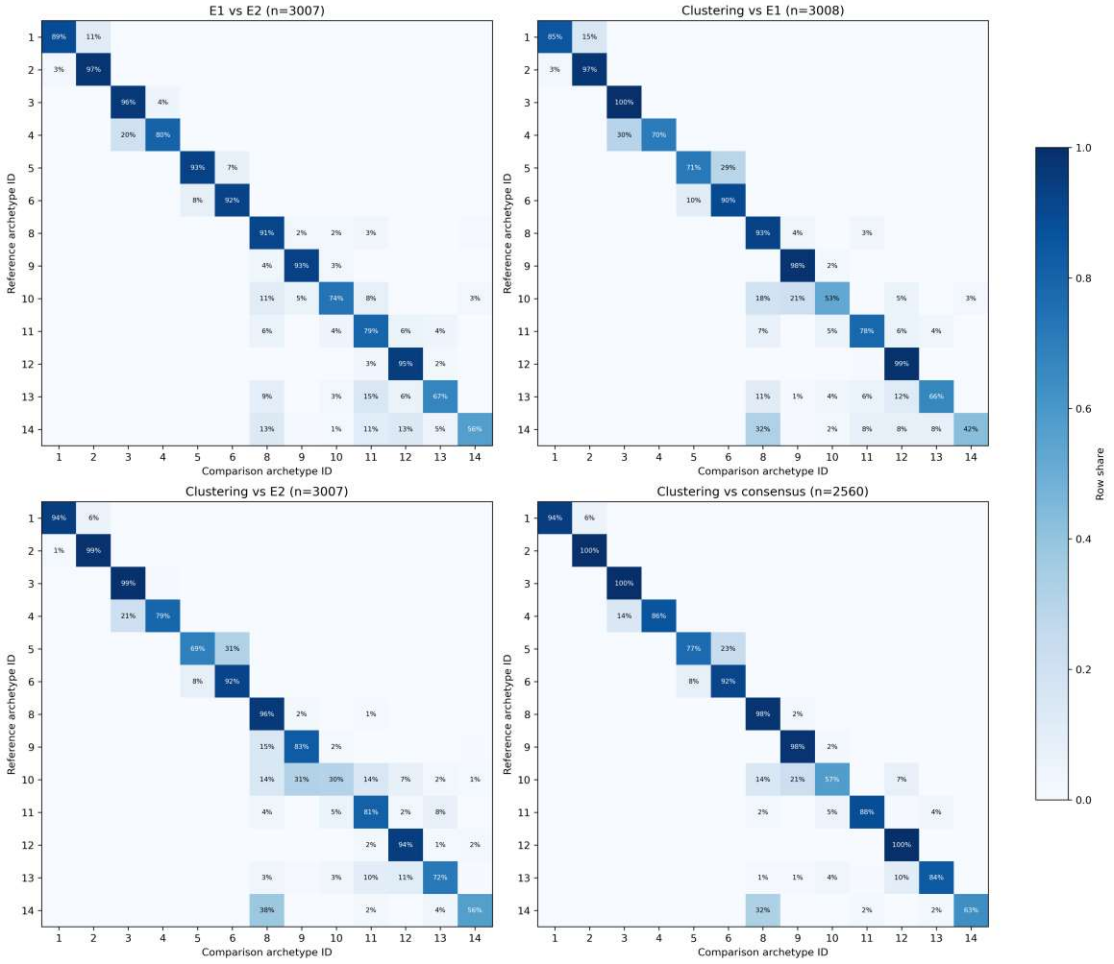


Figure 1: Confusion matrices for all archetypes against the expert-consensus subset ($n = 2560$; pre-annotated A7 was excluded).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# D Archetype Stability Validation

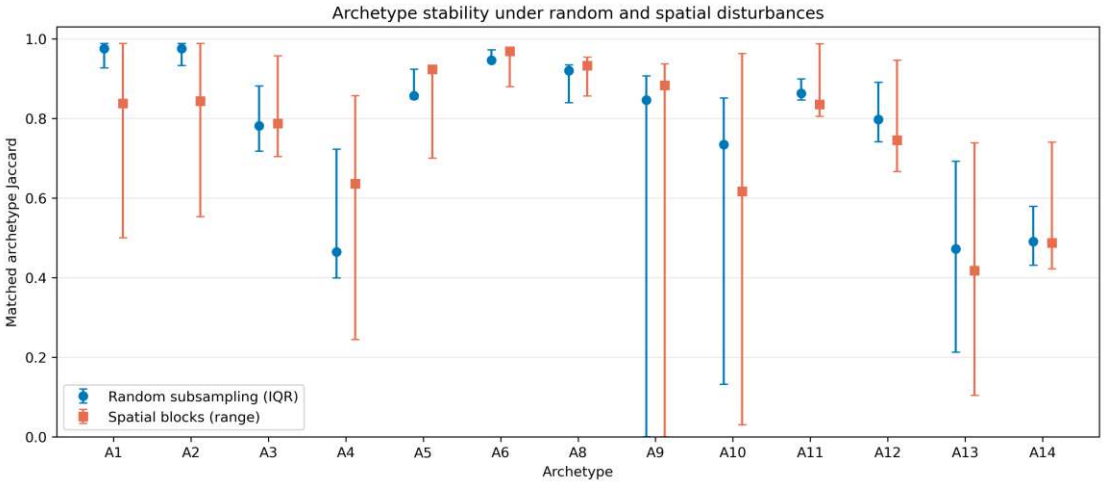


Figure 2: Archetype-specific Jaccard similarity for random 90% subsampling and spatial leave-one-block-out. Random error bars show the interquartile range, while the spatial error bars show the full range across the five folds.

# E Out-of-Fold Prediction Diagnostics

Table 2: Out-of-fold prediction metrics for *M0* and *M1* under both cross-validation schemes. Positive delta values indicate a lower error for *M1* compared to *M1*.

| | | MAE | | | RMSE | | | $R^2$ | |
|---|---|---|---|---|---|---|---|---|---|
| Scheme | Target | *M0* | *M1* | Δ [%] | *M0* | *M1* | Δ [%] | *M0* | *M1* |
| Spatial | Heating | 3.12 | 3.00 | 3.87 | 5.80 | 5.58 | 3.78 | 0.9472 | 0.9511 |
| Spatial | Primary energy | 6.33 | 6.14 | 2.91 | 13.46 | 13.44 | 0.16 | 0.9109 | 0.9112 |
| Spatial | $CO_2$ | 1.21 | 1.15 | 4.71 | 2.62 | 2.54 | 2.84 | 0.9162 | 0.9209 |
| Spatial | Electricity | 1.25 | 1.22 | 3.03 | 3.39 | 3.53 | -4.03 | 0.9838 | 0.9825 |
| Random | Heating | 2.90 | 2.73 | 5.75 | 5.34 | 4.74 | 11.29 | 0.9554 | 0.9649 |
| Random | Primary energy | 5.71 | 5.61 | 1.66 | 12.57 | 12.47 | 0.85 | 0.9223 | 0.9236 |
| Random | $CO_2$ | 1.05 | 1.03 | 1.78 | 2.45 | 2.39 | 2.13 | 0.9268 | 0.9299 |
| Random | Electricity | 1.11 | 1.07 | 3.56 | 3.14 | 3.13 | 0.50 | 0.9861 | 0.9862 |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The following figures compare the simulation-based target values with their respective out-of-fold predictions. The dashed line indicated a perfect match for the prediction and target.

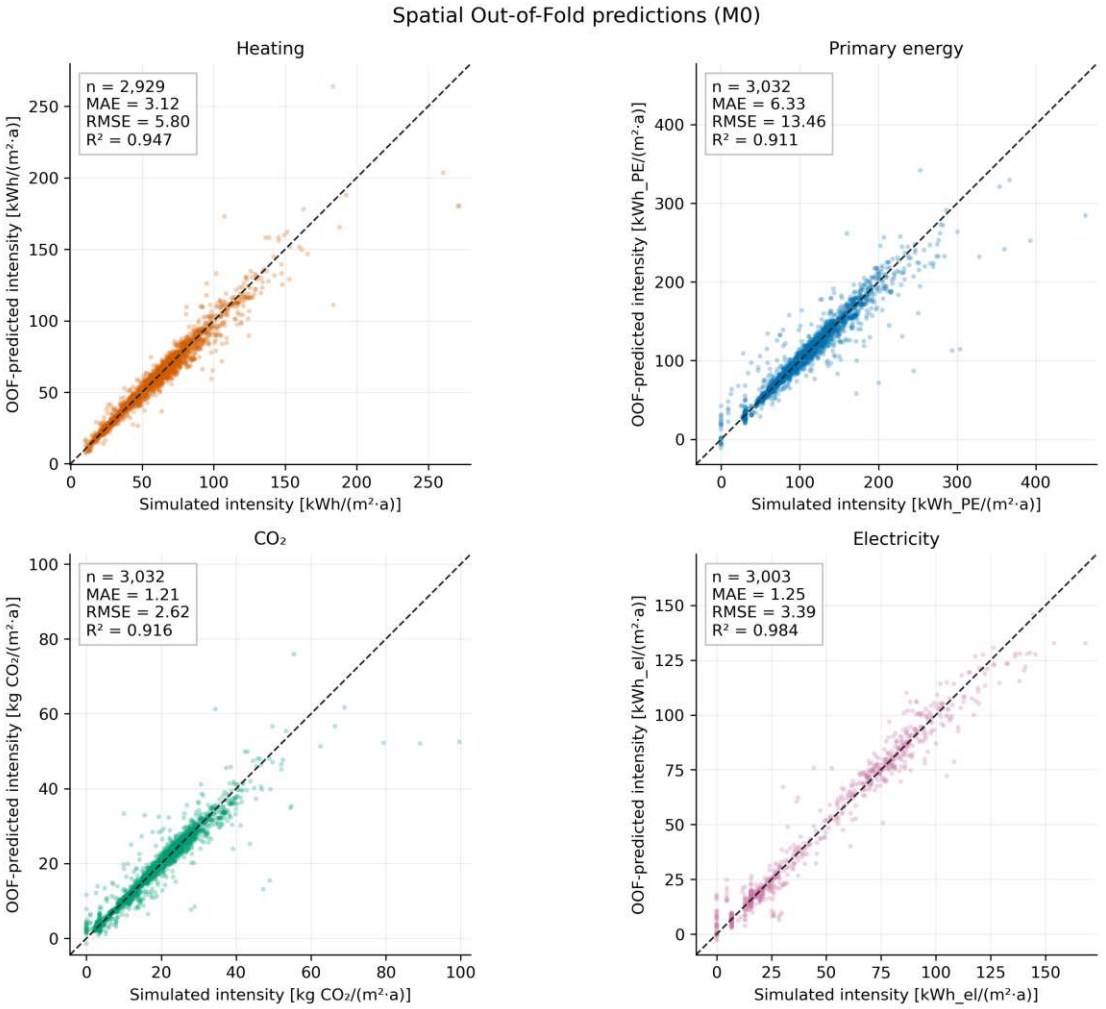


Figure 3: Simulated and spatial out-of-fold predicted annual energy intensities for model variant *M0*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

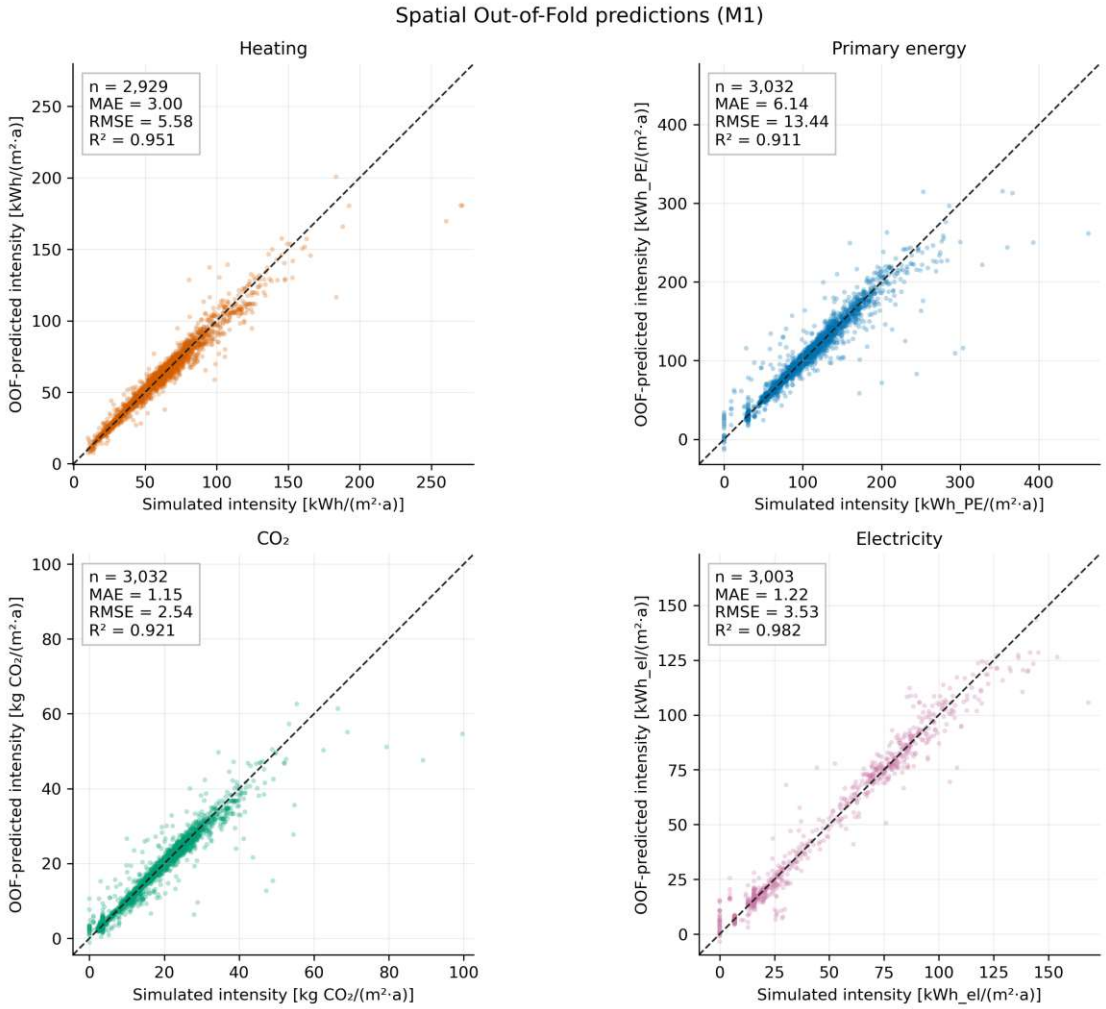


Figure 4: Simulated and spatial out-of-fold predicted annual energy intensities for model variant *M1*.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

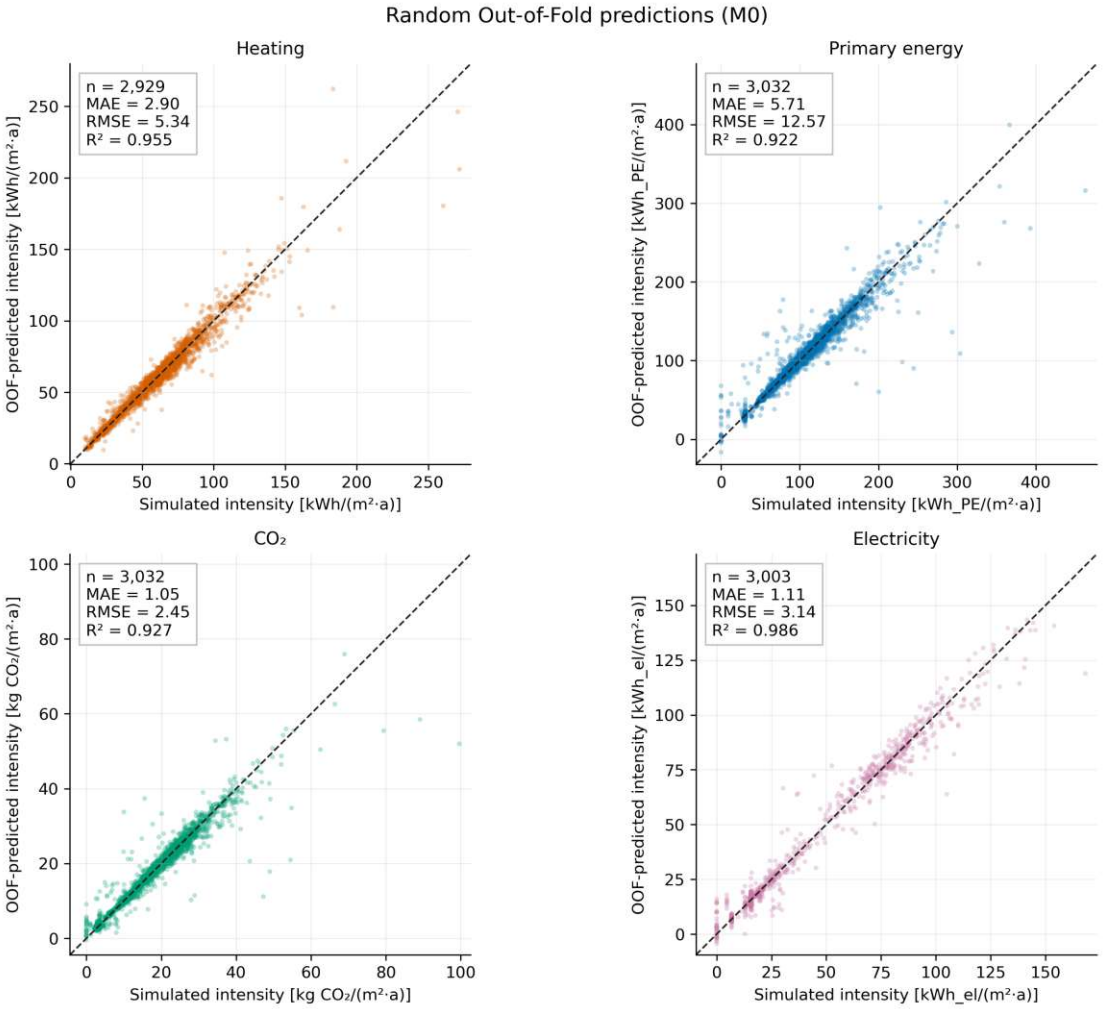


Figure 5: Simulated and random out-of-fold predicted annual energy intensities for model variant *M0*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

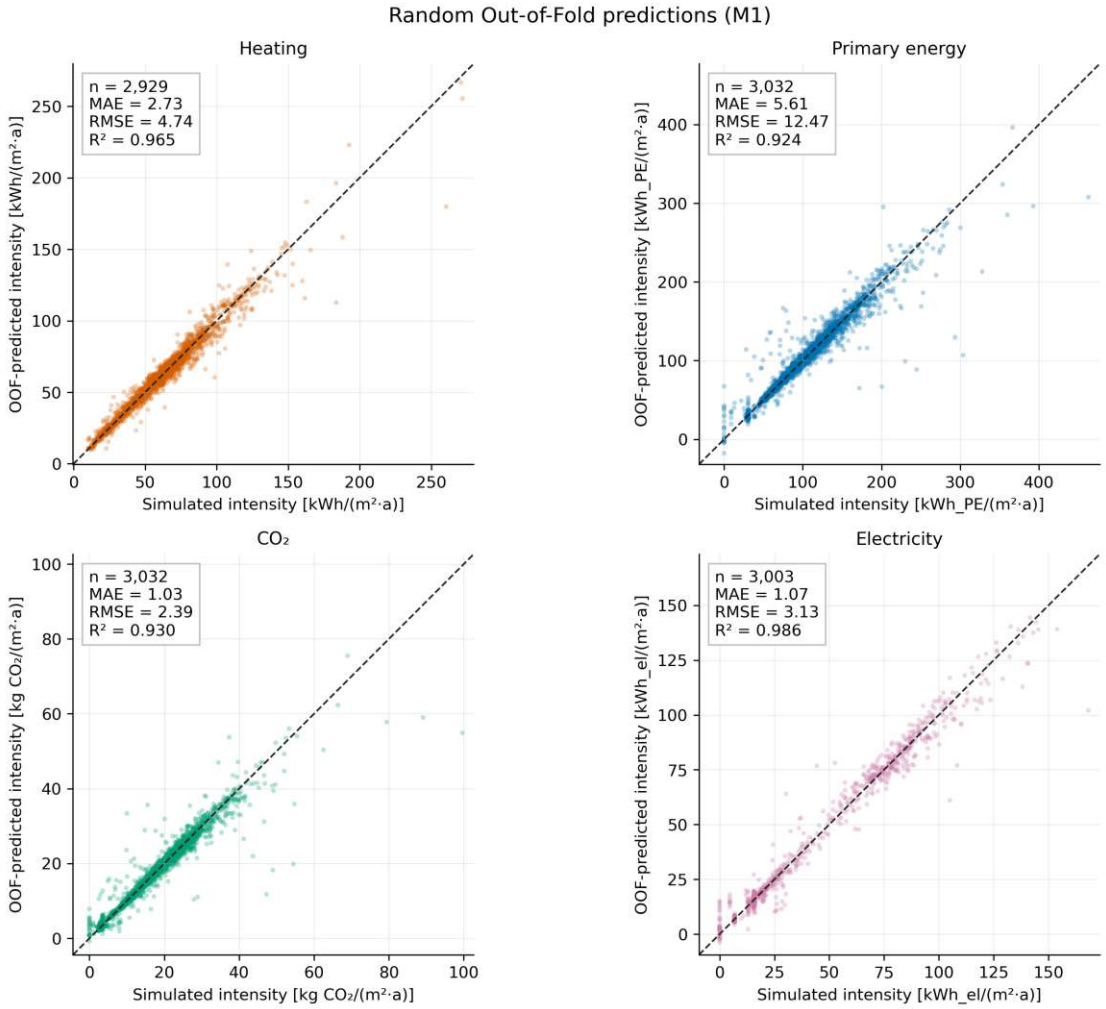


Figure 6: Simulated and random out-of-fold predicted annual energy intensities for model variant *M1*.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# F Spatial Out-of-Fold Error Diagnostics

The following map the relative change from *M1* compared to the *M0* variant, both on a 250-m grid. Green cells indicate a lower local MAE for *M1*, while red cells indicate a higher error compared to *M0*. Hatched cells contain $< 5$ buildings.

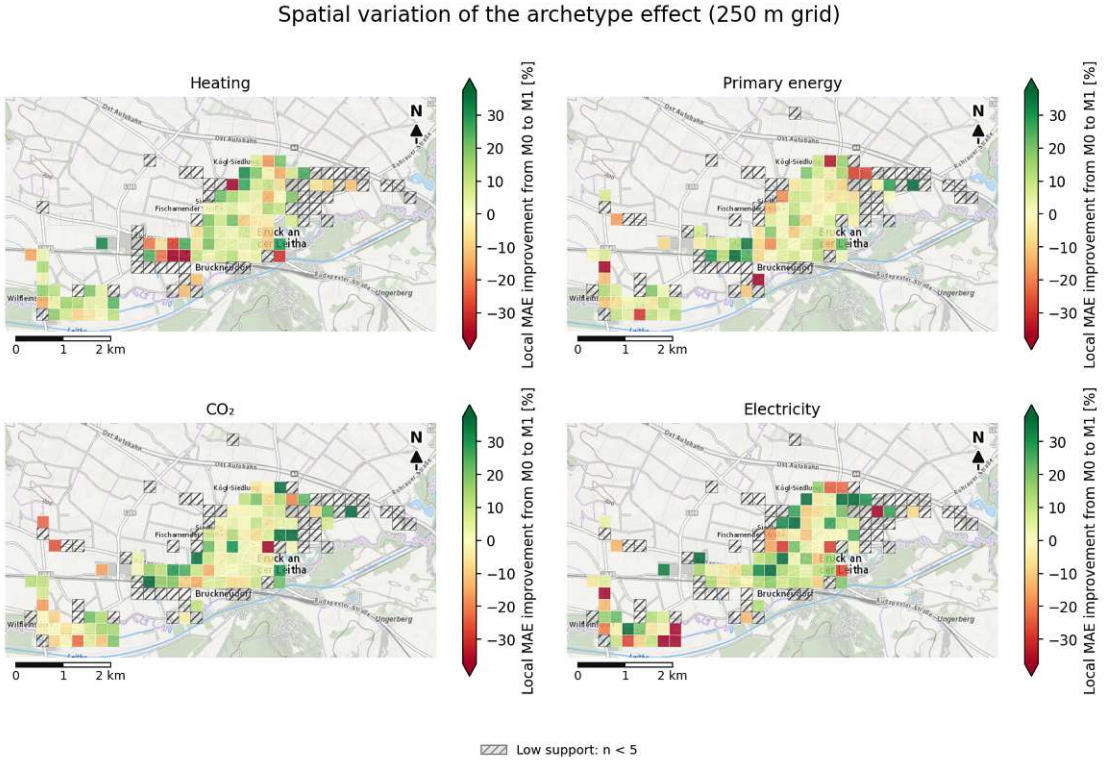


Figure 7: Relative MAE change from *M0* to *M1* on a 250-m grid.

*Source:* Own representation based on basemap from Geoland.at [Geo26].

Table 3: Share of supported 250-m grid cells with a lower local MAE for *M1* than *M0*.

| Target | Supported cells ($n \geq 5$) | Cells with lower *M1* MAE | Share [%] |
|---|---|---|---|
| Heating | 92 | 59 | 64.1 |
| Primary energy | 98 | 60 | 61.2 |
| $CO_2$ | 98 | 63 | 64.3 |
| Electricity | 97 | 57 | 58.8 |



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# G Spatial Indices

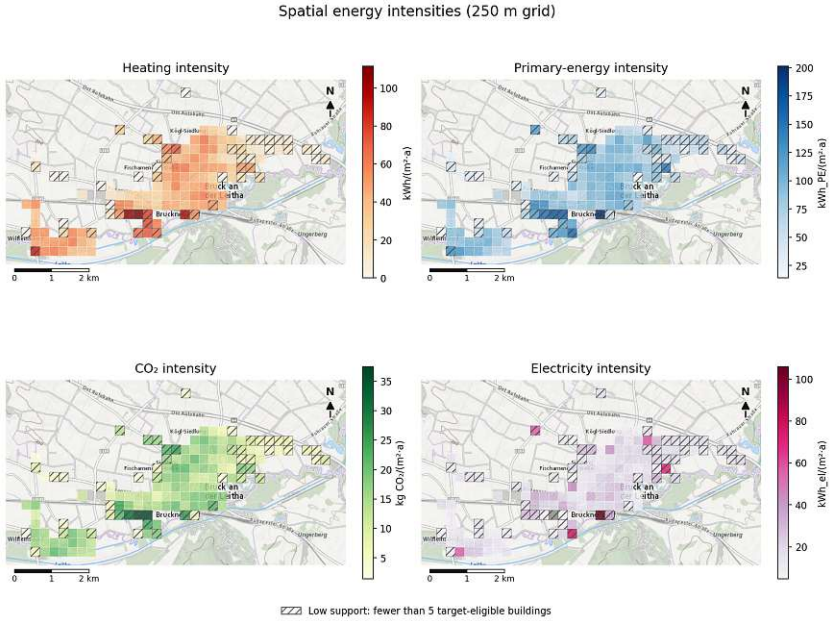


Figure 8: GFA-weighted spatial energy intensities for the four targets on a 250-m grid. Hatched cells contain $< 5$ buildings.

*Source:* Own representation based on basemap from Geoland.at [Geo26].

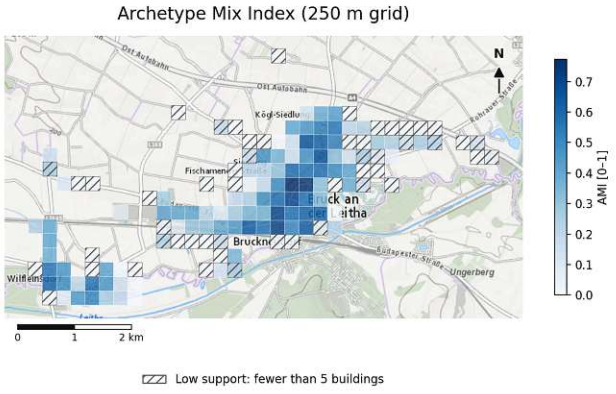


Figure 9: Archetype mix index on a 250-m grid. Hatched cells contain $< 5$ buildings.

*Source:* Own representation based on basemap from Geoland.at [Geo26].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.