# Predicting building age from urban form at large scale

Florian Nachtigall[1,2,*], Nikola Milojevic-Dupont[1,2], Felix Wagner[1,2], and Felix Creutzig[1,2]

[1] *Technical University of Berlin, Berlin, Germany*
[2] *Mercator Research Institute of Global Commons and Climate Change, Berlin, Germany*
[*] *Corresponding author (nachtigall@tu-berlin.de)*

## Abstract

To stay within 1.5°C of global warming, reducing energy-related emissions in the building sector is essential. Rather than generic climate recommendations, this requires tailored, low-carbon urban planning solutions and spatially explicit methods that can inform policy measures at urban, street and building scale. Here, we propose a scalable method that is able to predict building age information in different European countries using only open urban morphology data. We find that spatially cross-validated regression models are sufficiently robust to generalize and predict building age in unseen cities with a mean absolute error (MAE) between 15.3 years (Netherlands) and 19.9 years (Spain). Our experiments show that large-scale models improve generalization for predicting across cities, but are not needed to infer missing data within known cities. Filling data gaps within known cities is possible with a MAE between 9.6 years (Netherlands) and 16.7 years (Spain). Overall, our results demonstrate the feasibility of generating missing age data in different contexts across Europe and informing climate mitigation policies such as large-scale energy retrofits. For the French residential building stock, we find that using age predictions to target retrofit efforts can increase energy savings by more than 50% compared to missing age data. Finally, we highlight challenges posed by data inconsistencies and urban form differences between countries that need to be addressed for an actual roll-out of such methods.

*Keywords: Building age, Machine learning, Urban form, Energy modeling, Retrofit, Spatial autocorrelation*

## Highlights

- Assessing generalizability of age prediction across cities and countries based on urban form.
- Evaluating amount of local data needed to fill data gaps.
- Detailed analysis of predictive performance across regions, construction periods, settlement and buildings types.
- Assessing the usability of predictions to improve prioritization of large-scale retrofits.
- Highlighting climate relevance of scalable, spatially explicit building attribute prediction.

## Key findings

- Filling data gaps within countries is possible.
- 10% local data allow inference of remaining unknown building ages.
- Massive training data improves generalization across regions.
- Generalizing across countries is not (yet) possible.
- Age predictions can inform retrofit policies and may significantly increase energy savings.

# Contents

# 1 Introduction

The building sector is responsible for 31% of global final energy demand and 21% of global greenhouse gas (GHG) emissions [1]. Building emissions must be reduced by 80–90% by 2050 to reach the 1.5°C climate goal [2]. To implement solutions, policy-makers, administrations, and companies demand tailored approaches that fit their political, economic and climatic conditions [3, 4]. Yet, global assessments do not adequately reflect the different local conditions [5] making it difficult to provide geographically-differentiated and contextually relevant policy advice at local scale [6]. For an improved granularity of climate solutions, the availability of large-scale data is essential [5] and can help to translate high-level decisions into local actions. Particularly, demand-side climate solutions, which are highlighted in the IPCC's AR6 [7], require fine-grained analyses as they depend on local factors that vary across regions [8]. For targeted, large-scale retrofitting of the housing stock, one of the key demand-side solutions to decarbonizing cities, efficient identification of buildings with high energy saving potentials is essential.

Here, the construction year of buildings is highly relevant. Several studies have shown that the construction period is a key factor in modeling energy consumption of buildings [9–12] as it serves as a proxy for thermal insulation [13], ventilation rate [14], or glazing ratios [15]. As a result, old buildings often consume more energy than new buildings. Aside from energy modeling, the construction period has a variety of other important, climate change mitigation related applications. It is used in vulnerability and risk assessment for natural hazards, for example for earthquakes [16–18], floods [19], landslides [20] and extreme heat [21]. Moreover, it is important for material flow analysis in the construction sector [22].

Yet, for more than two thirds of EU buildings the year of construction is not publicly known. Only a few cities, regions and countries, such as the Netherlands, Spain and France, make the data publicly available, while the majority of age cadaster data needs to be bought, is subject to specific contractual conditions [23] or data protection restrictions [12]. To fill existing data gaps, previous research has shown that certain building characteristics can be inferred from its surrounding spatial context. This includes predicting building attributes like type [24–26], height [23, 27] or age [11, 28–30] based on urban morphology data, which is publicly available across Europe.

However, previous work has made little use of the availability of the existing open-access urban form data and was mostly focused on city level case studies [11, 28, 30]. Only a single study has aimed at generalizing across different cities, resulting in low prediction performance [29]. The potential for an improved cross-regional generalization by training on a large, heterogeneous dataset remains to be explored. Also, the generalization capacity for geographically distant and urban-typologically different regions has not been examined yet. Both are crucial to assess how well data gaps can be filled for whole cities or countries. In addition, the majority of studies lack adequate control for spatial-autocorrelation effects, resulting in overoptimistic generalization estimates [11, 28, 30–32].

Here, we propose a scalable machine learning approach to predict missing building age information across countries in the EU. Our goal is to help fill gaps in publicly available administrative data to enable large-scale studies on the building stock with high spatial resolution. In this regard, our main contribution is twofold: First, in contrast to local case studies, we investigate generalizability of building age prediction across regions in Europe. Second, we increase the spatial scope and identify benefits, challenges and usability of large-scale building age prediction, especially in light of massive training data and diverse urban form.

Overall, we examine the following subjects:

1. **Local inference with partial data availability** Certain regions have available data, but only for a percentage of buildings: we evaluate how well a model trained on urban form characteristics can augment missing building age information in areas with partial data availability.

2. **Regional generalization** We examine how well such models can generalize across cities and countries in the EU. Specifically, we want to see if it is possible to fill data gaps for entire cities or countries by learning a model in different regions for which we have data. To this end, we quantitatively assess how the generalization accuracy deteriorates over distance.

3. **Need for massive data** To evaluate the potential of large-scale studies, we analyze the impact of additional and diverse training data on the prediction accuracy. The goal is to determine how many training samples are needed for an optimal model, especially given that the current literature is limited to regional case studies.

4. **Inspection of prediction results** We inspect our results and compare the predictive performance between construction periods, regions, settlement and building types to provide more insights into the prediction making. We highlight differences in the feature importance and other country-specific challenges to make the predictions more understandable and inform applications that build on the inferred data.

5. **Applicability for retrofit policies** At last, we evaluate the usability of the inferred building age for energy modeling in order to inform retrofit policies on a national level.

## 2 Methods

We train a supervised machine learning model to predict the construction year of buildings in Netherlands, France and Spain based on publicly available 2D urban morphology data. In this section, we describe what data sources and features are used. We further describe which machine learning model is employed, how hyperparameters are optimized, and which metrics and cross-validation strategies are used for evaluation. Lastly, we introduce the set of experiments we conduct to answer our 5 research questions. Figure 1 provides an overview of the methods and machine learning pipeline. To validate and reproduce our approach, we made the source code available on GitHub: `https://github.com/ai4up/ufo-prediction`

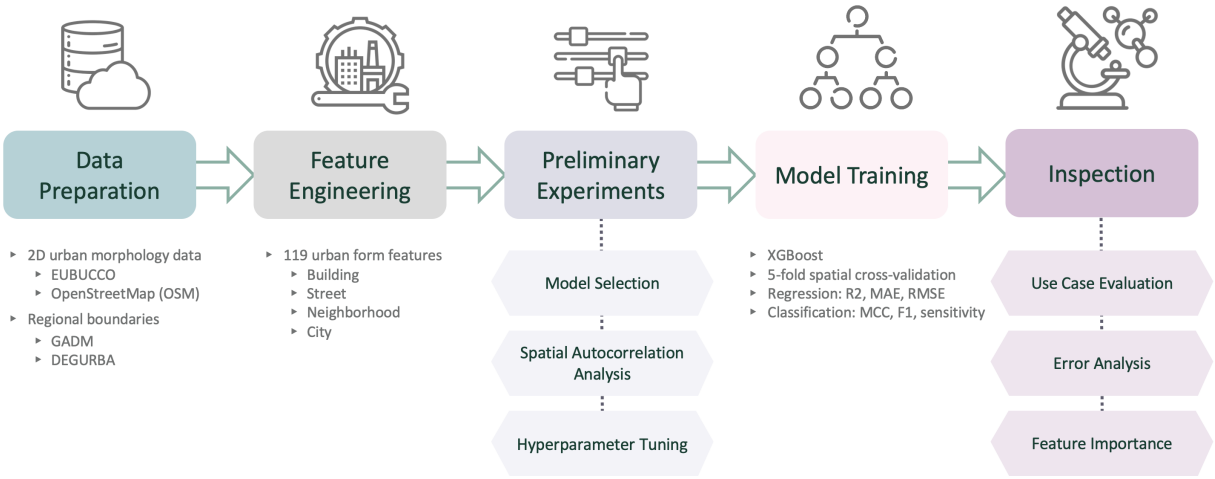


**Figure 1:** ***Methods overview.*** *The rounded boxes depict the main steps of our analysis, starting with data preparation and ending with the inspection of our prediction results. The boxes with sharp corners show the substeps for our preliminary experiments and prediction inspection. The light gray text highlights key aspects of the steps.*

## 2.1 Data

**Data source** Several data sources are used for the prediction task. Information on the building stock is derived from cadaster datasets from 2019 and 2021 downloaded via the harmonized database on the European building stock, called EUBUCCO [33]. The cadaster datasets include information about the buildings' geometry and attributes like construction year, residential usage and height. In addition, we downloaded OpenStreetMap (OSM) data [34] to retrieve street network information on all three countries. Both data sources are combined to construct the input features.

To analyze regional differences and calculate centrality features at the city level, we split up the three countries into regions and cities according to the boundaries of the Database of Global Administrative Areas (GADM) [35]. To assess differences between settlement types, we further classified these regions by their degree of urbanization (DEGURBA) [36] into densely (cities), intermediate (towns), and thinly (rural) populated areas.

We reuse the available DEGURBA classification [37] of Europe's Local Administrative Units (LAU) [38]. As they slightly differ from GADM regions, we map LAU and GADM regions based on their names and geometries. For ambiguous cases, we estimate the settlement type with spatially proximate regions.

**Target variable** As the target variable, we choose the year of construction of a single building. This information is available in the cadaster data with varying coverage. We select the three countries with the highest coverage: France with 45%, Spain with 98%, and the Netherlands with 100%. Together, they represent 36% of

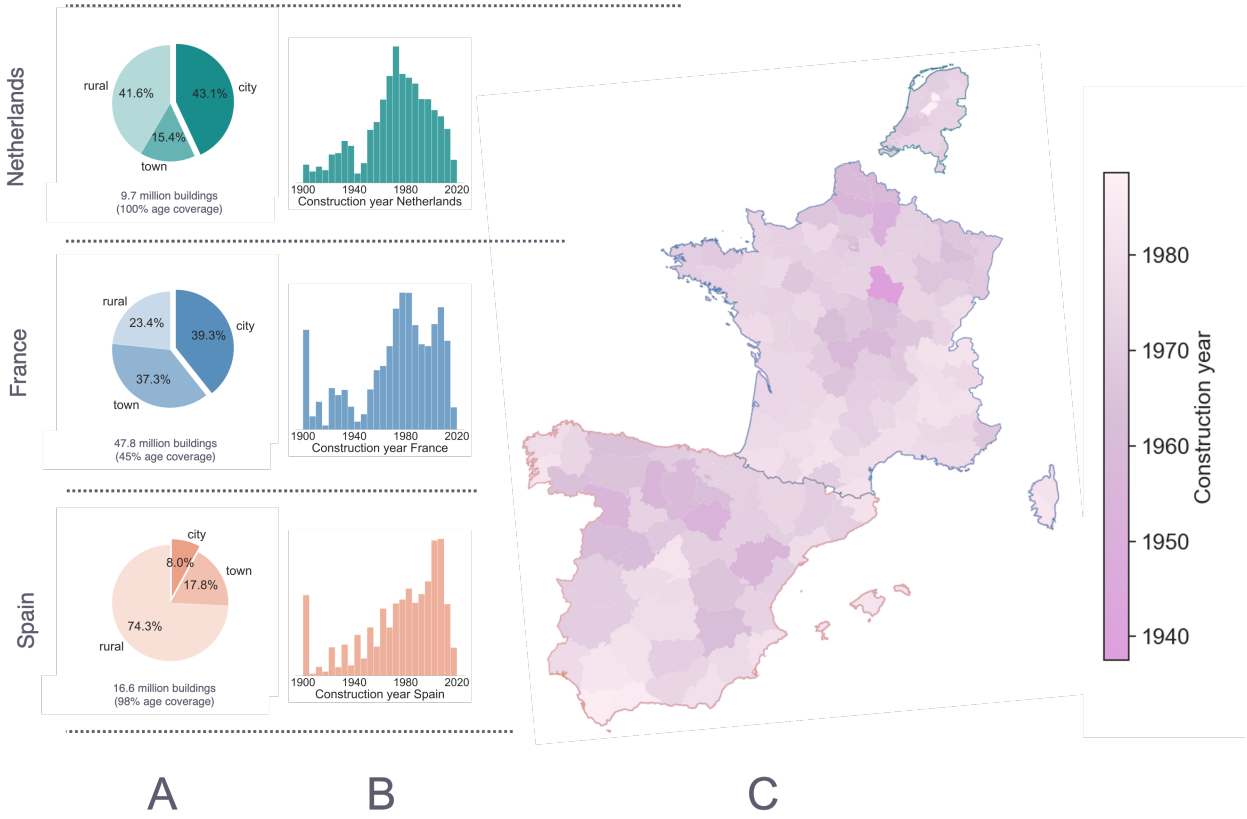


**Figure 2:** ***Data overview.*** *(A) Distribution of buildings from cities, towns and rural regions according to DEGURBA classification. (B) Distribution of the year of construction in the preprocessed dataset. (C) Average construction year per region in the Netherlands, France and Spain.*

| | | Construction year | | | | Other properties | | |
|---|---|---|---|---|---|---|---|---|
| Country | n | mean | std | Q1 | Q3 | residential [%] | footprint [$m^2$] | height [m] |
| Spain | 8,594,374 | 1975.6 | 32.4 | 1960 | 2001 | 78.3 | 181.2 | 5.0 |
| France | 10,638,655 | 1971.9 | 31.7 | 1958 | 1997 | 82.1 | 132.2 | 5.0 |
| Netherlands | 6,106,318 | 1972.3 | 26.7 | 1959 | 1992 | – | 124.5 | 5.6 |

**Table 1:** ***Summary statistics of preprocessed dataset.*** *The mean, standard deviation (std), 25% quantile (Q1), and 75% quantile (Q3) of the contruction year of all buildings in the preprocessed dataset are shown. Further, the share of residential buildings, the average height, and footprint area are depicted.*

the EU building stock [33]. For our analysis, we focus on buildings constructed after 1900, as these are of most interest for energy modeling and other climate mitigation related use cases. Also, for many older buildings the data quality deteriorates with the construction year being rounded to the nearest decade or century, which introduces unwanted noise into our dataset. After preprocessing, we utilize a sample of 25.3 million buildings with known year of construction information for our machine learning approach. In the following, we use building age as a synonym for the year of construction of buildings, as it improves readability and is commonly used in the literature.

Figure 2 and table 1 provide an overview of the distribution of building age and other key descriptive statistics of the data used in this study. Most importantly, the mean construction year is similar in all countries. However, in the Netherlands there are fewer new buildings and the overall distribution is more centered. In Spain the majority of buildings is located in thinly-populated, rural regions (74.3%), whereas in the Netherlands most buildings (43.1%) are located in densely-populated, urban areas. This is noteworthy because the mean construction year differs between settlement types, especially in Spain and France. In France, cities are almost 15 years older than rural areas and around 10 respectively 20 years older than cities in the Netherlands and Spain. In Spain, the opposite is true, with buildings in rural areas being older than buildings in cities and also older than buildings in rural areas in France and the Netherlands (see Appendix table 12).

**Feature engineering** To estimate building age, we utilize 119 features based on 2D urban morphology data on buildings and street networks, developed in [23]. While [23] developed these feature to infer building height, we hypothesize that their encoding of the surrounding urban context, as well as the building itself, will also be predictive of the building's construction year. The features can be divided into groups of building-, building block-, street-, street-based-block- and city-level features. Features describe either the building itself, e.g., its footprint area, properties of the closest street or intersection, e.g., betweeness centrality of the closest street, or summarize information about urban form within a 100 or 500 m squared buffer around the building, e.g., average building footprint area. Table 2 provides an overview of the features used for prediction. For a complete list of individual features including a brief description, variable unit, and definition, if applicable, see Appendix A. For a more contextualized description and rationale for their selection, refer to [23].

The features show several differences in the urban morphology between the Netherlands, France, and Spain, particularly in regards to footprint area characteristics. The share of buildings with small footprints is largest in France, while the density of buildings is highest in the Netherlands. An explorative analysis and visualization of how urban form features differ across construction periods and countries is performed in experiment 5 (see section 3.4F, table 14 and Appendix table 26).

| Feature group | Subgroup | Example | Count |
|---|---|---|---|
| *Buildings* | Building's own geometry | Footprint area | 10 |
| | Building's spatial location | Latitude | 2 |
| | Other buildings within 100 & 500 m | Mean buildings convexity within 100 m | $10 \times 2$ |
| *Building blocks* | Building's own block (if any) | Block elongation | 10 |
| | Other blocks within 100 & 500 m | Number of blocks 100 m within buffer | $9 \times 2$ |
| *Streets* | Closest street & intersection | Distance to closest street | 6 |
| | Street centrality | Betweeness centrality of closest street | 11 |
| | Streets & intersections within 100 & 500 m | Standard deviation length streets within 100 m | $6 \times 2$ |
| *Street-based blocks* | Building's street-based block | Street-based block's corners count | 3 |
| | Street-based blocks within 100 & 500 m | Average area street-based block within 500 m | $6 \times 2$ |
| *City level* | Buildings | Total building footprint area | 5 |
| | Blocks | Number of blocks consisting of 5 to 9 buildings | 4 |
| | Streets & intersections | Average street lengths | 3 |
| | Street-based blocks | Number of street-based blocks | 3 |
| Total | | | **119** |

**Table 2:** ***Overview of urban form features.*** *Breakdown of the 2D urban morphology characteristics used to predict building age into feature groups and subgroups. For each subgroup the number of features and one example is provided. See Appendix A for a complete list of individual features.*

## 2.2 Machine learning approach

**Evaluation metrics** We train a supervised machine learning model using the mean absolute error (MAE), the root mean squared error (RMSE) and the coefficient of determination ($R^2$) as evaluation metrics to assess the model's ability to infer missing building age information and generalize across regions.

**Model selection** While [23] found that XGBoost [39] models yield the smallest prediction error when inferring building attributes, the majority of prior studies [11, 28–30] used Random Forest [40] models. We conduct preliminary experiments to compare the predictive performance of XGBoost and Random Forest learners. We focus on decision tree ensemble methods as a comparative study [41] found them to still outperform deep learning approaches on tabular data. We find that the XGBoost regressor achieves a 2 percentage points larger $R^2$ for our prediction task while training twice as fast given our computational resources (see Appendix section B). Therefore, we utilize XGBoost for all further experiments.

**Hyperparameter tuning** Due to computational constraints in our experiments (see Appendix C), we perform the hyperparameter optimization as a preliminary step on a throw-away set of 10% of the data. The data used for any of the preliminary experiments including hyperparameter tuning will not be reused in any of the final experiments in order to avoid data leakage.

We utilize sklearn's random search [42] for hyperparameter tuning with 50 iterations minimizing the RMSE. We perform random 5-fold cross validation and city-wise 5-fold cross validation to account for the two major prediction use cases of this work, local inference and regional generalization (see section 2.3). We combine the results and select the hyperparameters, which performed best in both settings, while also taking the trade-off between prediction performance and training time into account. For all subsequent experiments, we select a maximal tree depth of 13 (`max_depth`), 1000 trees overall (`n_estimators`), a learning rate of 0.025 (`learning_rate`) and a random subset of 90% of the features for each tree (`colsample_bytree`) and 50% for each decision split (`colsample_bylevel`).

As it is difficult to make an educated guess for the optimal model complexity based on hyperparameter tuning on a subsample of the data, we verify, after conducting all experiments, if the chosen model complexity was appropriate and how large the potential improvement may be from performing nested cross-validation on 100% of the data (see Appendix C).

**Spatial cross-validation & autocorrelation** For all experiments, we perform 5-fold cross-validation. Motivated by our use cases, local inference and regional generalization across cities and countries, we perform random cross-validation, spatial city-based cross-validation, and spatial country-based cross-validation. An exemplary train-test split for the different cross-validation strategies is depicted in figure 3.

An important aspect to consider when choosing an appropriate model validation strategy, is the spatial autocorrelation of building age [29], meaning that neighboring buildings are more likely to be built around the same time than non-neighboring buildings. Depending on the use case, we recommend to exploit these effects to improve the prediction, e.g., for use cases where the age of neighboring buildings is known (local inference), or to prevent exploitation to avoid overoptimistic generalization estimates, e.g., for use cases where the age of neighboring buildings is not known (regional generalization). To evaluate the meaningfulness of features, for example to assess how well the urban morphology alone can explain building age, we also advice controlling for spatial autocorrelation effects to avoid the learning of spurious correlations. Correspondingly, we develop an additional cross-validation strategy that minimizes the spatial autocorrelation, while not making the prediction unnecessarily hard by having different architectural styles and urban morphology patterns in the train and test set as it is the case for cross-country or cross-city prediction. This allows predictive performance to be compared across building types, settlement types, and regions independent of a specific use case, while not being biased by different city sizes or different degrees of spatial clustering.

To prevent information leakage from the training to the test set due to spurious spatial autocorrelations and consequently overoptimistic generalization estimates, neighboring buildings must not be split into training and test set. To this end, spatial cross-validation methods ensure the spatial division between training and test samples by partitioning the data into non-overlapping spatial chunks with negligible spatial autocorrelation between them [43]. In addition, a spatial buffer between test and training set can be enforced [44, 45]. The minimal spatial size of the chunks and buffer depends on how far the spatial autocorrelation effect persists and what residual is acceptable [43].

Preliminary experiments show that spatial autocorrelation of building age decreases strongly for the first hundred meters and then levels off after 1 km (see section Appendix D). Consequently, we agglomeratively cluster buildings until a distance threshold of 1 km is reached. We refer to them as neighborhoods. On average, 15 neighborhoods make up a city and each neighborhood has an average size of 2 $km^2$. To investigate the effect of increasing spatial autocorrelation on predictive performance, we also test a smaller spatial cluster of buildings that we refer to as urban blocks, i.e., buildings surrounded by drivable roads. They average one-seventh the size of a neighborhood. While other approaches exist to obtain unbiased estimates of predictive performance for spatially autocorrelated data, e.g., probability sampling and design-based inference [46], we choose a spatial cross-validation approach because it is most aligned with the validation strategies dictated by our regional generalization use cases.

## 2.3 Experiments

We conduct the following five experiments to answer our main research questions:

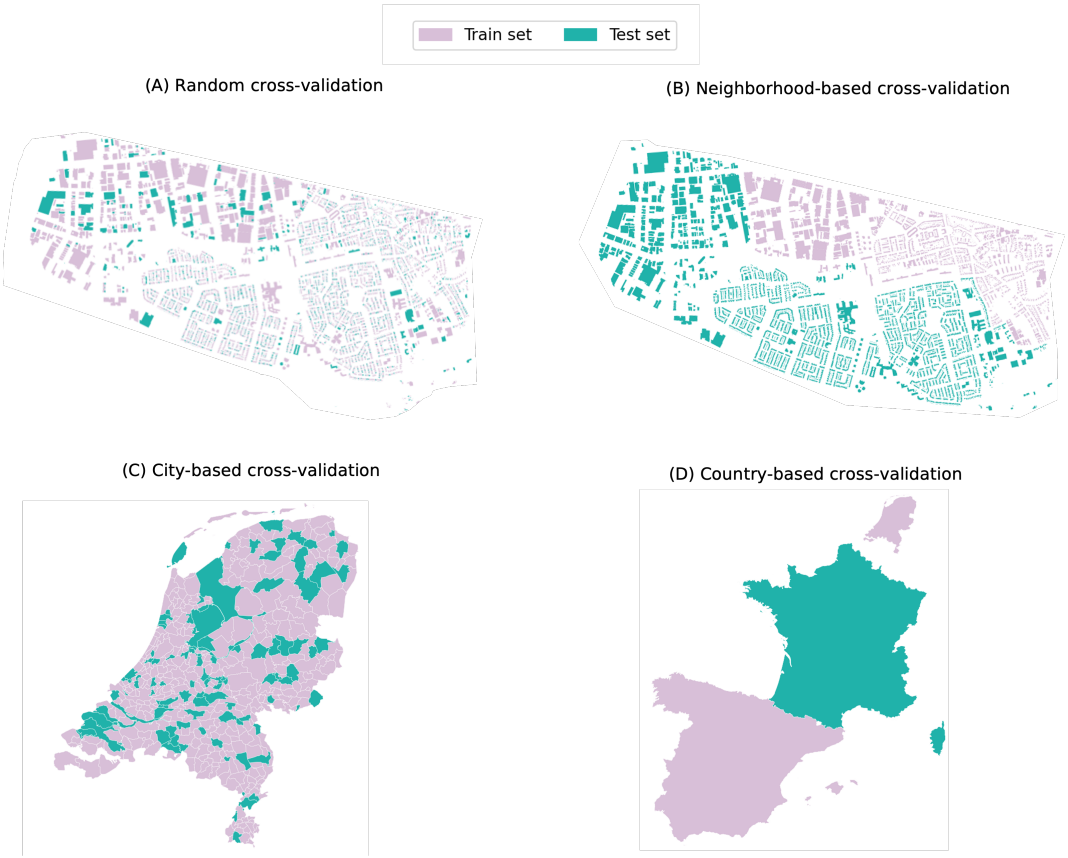


**Figure 3:** ***Illustration of an exemplary train-test split for different cross-validation strategies.*** *Purple colored buildings and areas are used for model training, green colored for testing. (A) and (B) show the footprint geometries of buildings in an area of Ede, Netherlands. The neighborhoods (B) are determined by agglomerative clustering of urban blocks with a maximum diameter of 1 km. GADM boundaries are used to map the cities in the Netherlands (C) and the countries Netherlands, France, and Spain (D).*

**Exp. 1 – Local inference with partial data availability** To address research question 1, we predict the construction year of buildings in areas with partial data availability. We assess the positive impact of spatial autocorrelation on the predictive performance by evaluating the model using different cross-validation techniques. We perform random cross-validation to allow for full exploitation of spatial autocorrelation effects between neighboring buildings. To prevent reporting overoptimistic accuracy estimates for buildings without neighbors and to assess how well urban morphology alone can explain the building age, we use neighborhood-based cross-validation.

To further assess the amount of local data needed to accurately predict the construction year for all other buildings in the region, and if training data from outside the region are needed, we learn a model that utilizes 80%, 50%, 20%, or 10% of the data from the region. We compare the prediction quality among the regional models and with a national model that utilizes country-wide data.

To differentiate between country-specific prediction challenges and to evaluate the effects of geographically diverse training data, we perform the experiments for each country individually and for all countries combined.

**Exp. 2 – Regional generalization** To answer research question 2, we perform city and country-based spatial cross validation with different spatial constraints. To evaluate the generalization performance across cities, we train the model on a random selection of cities within a country and predict in all remaining cities of the country, performing 5-fold spatial cross-validation on a city level. We conduct the experiments for the Netherlands, France, and Spain individually and for all three combined.

To assess if the geographic distance between buildings in the train and test set impacts the generalization performance, we repeat the experiments, but enforce an increasing spatial distance between the train and test set. We select a test region and divide the cities outside the region in up to 9 groups based on their distance from the test region (see figure 4). For each group, we train a separate model and analyze how the prediction quality varies over spatial distance. For a detailed description of the experimental set-up, refer to Appendix E.

To further test if it is possible to infer building age in a country where this information is not available, by learning a model in countries where it is, we perform 3-fold country-wise cross-validation for Netherlands, France and Spain. We train the model on two countries and predict in the third one. This extends the previous experiment on generalization across distances by assessing the additional challenge of national boundaries and thus differences in cadaster data quality and architectural style.

**Exp. 3 – Need for massive data** To investigate research question 3, we increase the number of training samples and quantify the impact on the prediction performance. We primarily focus our analysis on Dutch building data as the regions are of comparable size and exhibit a detailed and stable data quality.

We evaluate the potential of massive training data for a neighborhood cross-validated model and our three main use cases, local inference, generalization across cities, and generalization across countries. For cross-city and cross-country generalizing, we keep the test set fixed and only add buildings to the training set throughout the experiment. For local inference, we perform random cross-validation. We add buildings from one city at a time to ensure that in each iteration there are enough local buildings that can be exploited by the model for local inference. In addition, this ensures a smooth distribution of feature values, such as building density or street network centrality measures, which facilitates the learning of city structures. Once all cities from the country are utilized, we add buildings from a different country to assess the impact of transnational training data. We run and average all experiments across 10 iterations with different seeds to reduce the impact of the city sampling on the result.

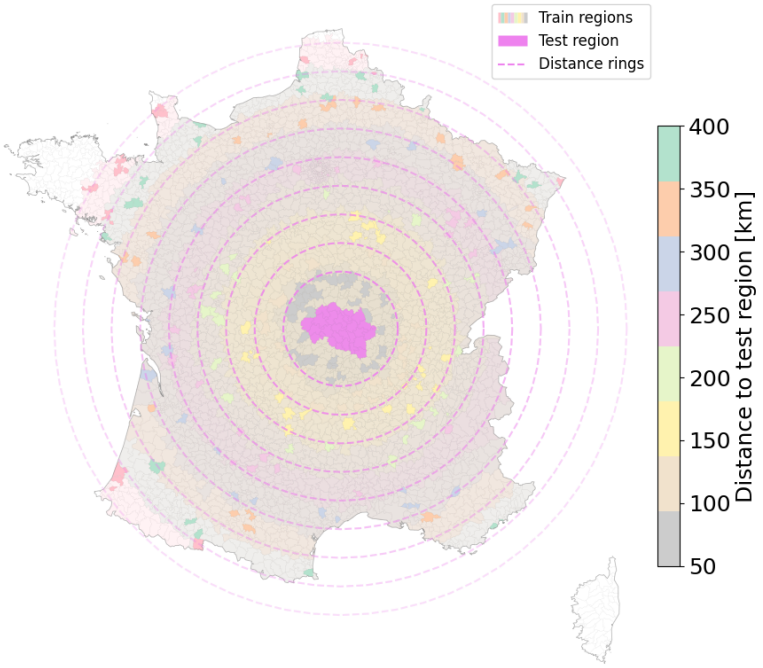


**Figure 4:** ***Division of French cities into 9 groups to assess the impact of geographical distance on the generalization performance.*** *Exemplary visualization for the department of Allier, France. The distances used for grouping the cities are visualized by the pink buffer rings. The cities from each group that were selected for training are highlighted with the color indicating their distance to the test region. Each group is 50 km further away from the test region. All highlighted cities within one distance ring have in total around 100,000 buildings.*

**Exp. 4 – Inspection of prediction results** To address research question 4, we inspect the results along six dimensions (A–F) that provide detailed perspective on where and why the model may perform better or worse.

(A) First, we compare the predictive performance between different construction periods. Since regression residuals are biased towards the mean, we redefine the task as a classification problem by grouping the construction years into construction periods (e.g., 1950–1965, see Appendix F). By doing so, we test whether old and new buildings are indeed more difficult to predict or whether this is an inherent problem of the chosen regression approach.

(B) In addition, we compare geographic regions defined by GADM boundaries and settlement types according to the EU DEGURBA classification. We test the correlation between prediction error and settlement size using Pearson's correlation coefficient.

(C) We also assess the challenges of predicting age of different building types. Information on residential and certain non-residential usages, i.e., commercial, agricultural, and industrial, are available in the cadaster data for France and Spain. Residential subtypes, i.e., single-family houses, terraced houses, multi-family houses, and apartment blocks, are estimated using a simple decision tree based on reference values for building height, footprint area, and number of adjacent buildings for these residential subtypes from TABULA [47] (see Appendix section G). We compare the model performance between residential and non-residential buildings and their subtypes in France and Spain.

(D) To follow up on a limitations of [29], which is to consider only urban blocks where all buildings are from the same construction period, resulting in an incomplete and biased building sample, we explore the challenge of predicting in-fill buildings. Generally, in-fill housing refers to new buildings constructed on underused lots in existing, older urban neighborhoods. For our experiment, we define in-fill housing as adjacent buildings whose construction year is more than two standard deviations above the mean of all buildings in the same urban block. See Appendix figure 21 for an illustration of example in-fill buildings.

(E) We calculate the feature importance to investigate which urban form characteristics are most informative. We calculate the relative contribution of each feature to the prediction based on their SHAP-values [48] as suggest by [49]. We assess their contribution individually and jointly as feature groups as defined in table 2. We validate that all features capture some information about the target variable by adding a random noise feature and testing that it has the lowest feature importance.

(F) Finally, we compare how the urban morphology, assessed by the 9 most important urban form features, differs across countries and construction periods to explore reasons for the differences in predictive power and generalizability.

**Exp. 5 – Applicability for retrofit policies** To answer research question 5, we use the predicted construction year to estimate the heating energy demand and savings potential from energy retrofits for all 8.7 million residential buildings in France present in the dataset. We evaluate to what extent our inferred data can improve the prioritization of large-scale retrofits by focusing on buildings with the highest energy demand. Our analysis is twofold:

(A) To assess the usability of our data, we first determine how much higher the energy savings per $m^2$ are for a prioritized retrofit approach compared to a non-prioritized, random approach. We compare the estimated savings potential of prioritization based using our inferred data and the ground truth construction year data.

(B) To further show the policy relevance of having such information at hand, we analyze the spatial heterogeneity of retrofitting needs and discuss the potential for targeted regional fund allocation and regional policy focus.

The energy performance of buildings is assessed based on an established energy model developed according to EN ISO 13790. We use precalculated heating energy estimates for specific building cohorts and different refurbishment conditions as de-

fined by TABULA [47]. The building cohorts are differentiated according to the construction period, climate region, and residential type, namely single-family houses, terraced houses, multi-family houses, and apartment blocks. We use a simple decision tree based on building height, footprint area and number of adjacent buildings to estimate the residential type (see Appendix section G). We focus our analysis on building data from France, as only there do we have appropriate coverage of the information necessary to classify buildings into residential types. In France, the energy estimates for heating are differentiated according to three climate zones (H1, H2 and H3) following French regulation, i.e., Réglementation Thermique de 2012. Given the residential type and building location, we use our age predictions to identify the TABULA cohort and obtain an estimate of the heating energy demand.

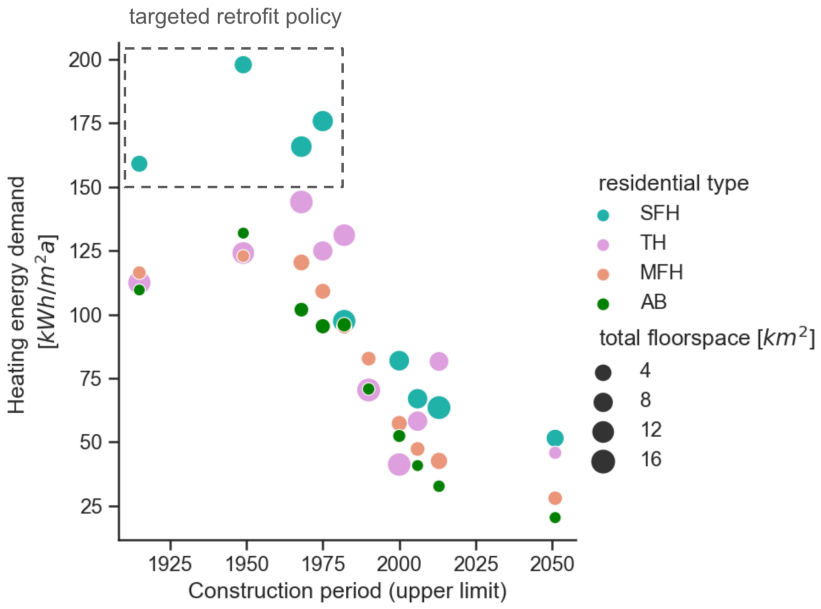


**Figure 5:** ***Building-level energy demand for heating in France.*** *Depiction of the average national heating energy demand per $m^2$ for different residential building types and construction periods in France according to the EN ISO 13790 compliant energy model from TABULA [47]. Building archetypes with a heating demand above 150 kWh/$m^2$a that are targeted for refurbishment are highlighted by the gray box. Buildings are matched to TABULA cohorts based on the construction year, geographic location and 2.5D shape from the EUBUCCO dataset [33]. The prevalence of each archetype in the building stock, as measured by total floor area, is coded by circle size.*

We choose an exemplary threshold to prioritize buildings with a heating energy demand above 150 kWh/$m^2$a for retrofit, which we estimate to represent about 16% of the total building stock. The buildings above the threshold correspond to single-family houses built until 1974 (see figure 5), approximately when the relevant energy regulations were introduced in France. To examine only the value of age information, not the residential type classification, we consider random retrofits of only single-family house as a second baseline for the energy savings assessment (A). We adopt a definition of retrofit that involves upgrading the thermal envelope and the heat supply system to an extent that is commonly realized during renovation and referred to as energy performance level (EPL) 2 by TABULA [50].

While this analysis does not account for the variability of energy demand across buildings from the same construction period, residential type, and climate region, it enables to assess the value of our age predictions compared to having no or perfect knowledge of the building age. Yet, all insights are constrained to the underlying energy model.

Finally, since primarily the construction period, not the precise year of construction, is relevant for energy modeling, we investigate whether a classification of the construction period can more effectively inform retrofit policies than a regression of the year of construction. (see Appendix section H).

# 3 Results

## 3.1 Filling data gaps within cities is possible

With access to local information by performing random cross-validation, the regression model is able to infer missing building age with a promising accuracy. The $R^2$ ranges from 0.63 for the Netherlands and 0.46 for France to 0.43 for Spain (see table 3). Overall, 69% of buildings in the Netherlands, 45% in France, and 41% in Spain are predicted with an MAE of less than 10 years.

Using only regional data for training yields more accurate predictions than using data from the entire country, with a $R^2$ on average between 2.4 and 4.0 percentage points larger (see figure 7 and Appendix table 14). This demonstrates that local models are appropriate and sufficient for local inference. Comparing different scenarios of local data availability, we find that the prediction error increases only moderately with less data (see figure 7). Between 80% and 10% of local data availability, the mean $R^2$ differs by 8.6 to 10.2 percentage points and the MAE by 1.5 to 2.1 in the Netherlands, France, and Spain (see Appendix table 14).

Performing spatial cross-validation and thereby reducing the availability of nearby buildings and preventing the exploitation of spatial autocorrelation degrades model performance (see figure 6). When conducting neighborhood-based cross-validation, the $R^2$ of our national models decreases by 19 percentage points in the Netherlands, by 5 in France and by 9 in Spain (see table 3). When performing spatial cross-validation on smaller spatial chunks, i.e., urban blocks that are one-seventh of the size of neighborhoods, the decline in predictive accuracy is smaller.

Further experiments demonstrate that our proposed method can also fill local data gaps for building type and building height. Using the same urban form features, we are able to predict type with a F1 score of up to 0.79 and height with a $R^2$ of up to 0.77 when performing neighborhood-based cross-validation (see Appendix section I).

## 3.2 Generalizing across regions has limitations

The generalization performance deteriorates with increasing geographical scope. Predicting building age in unseen cities is possible, but gets more difficult the further the cities are apart from the train region. For unseen countries, however, urban form is no longer predictive in our experiments.

Generalization across cities yields a $R^2$ of 0.32 for Spain, 0.38 for the Netherlands, and 0.39 for France (see table 3). This indicates that the construction periods of different regions are similarly manifested in the urban form, allowing the model, to some degree, to learn in one region and predict in another. Experiments for buildings type and height confirm this (see city cross-validation in Appendix table 9, 10 and 11). Precisely, the $R^2$ is between 2 and 6 percentage points lower compared to experiments where buildings from the same city are available for training (see table 3).

In general, spatial distance between buildings used for training and buildings from the test set negatively impacts the prediction accuracy (see figure 8). For a relatively small training set of 100,000 buildings, generalizing over more than 250 km is not possible in Netherlands and Spain. For larger distances the average $R^2$ becomes negative. We theorize that spatial distance has the same effect when more training data are used, just at a higher level of $R^2$.

As of now, learning in one country and accurately predicting in another country is not possible with our model. Regardless of which countries the model is trained on, the predictions scatter only around the mean of the distribution (see figure 6) and the $R^2$ never exceeds 0.2. Despite the harmonization efforts for the EUBUCCO dataset, urban form feature distributions differ noteworthy between the countries (see Appendix figure 25 and 26) making it difficult to generalize from one country to another.

| | Netherlands | | | France | | | Spain | | | All countries | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ |
| *Random cv* | 10.1 | 16.1 | 0.63 | 16.8 | 23.4 | 0.46 | 17.5 | 24.5 | 0.43 | 16.2 | 22.9 | 0.45 |
| *Urban block cv* | 12.5 | 18.5 | 0.51 | 17.6 | 24.1 | 0.42 | 18.5 | 25.5 | 0.38 | 17.0 | 23.6 | 0.41 |
| *Neighborhood cv* | 14.3 | 19.9 | 0.44 | 18.0 | 24.4 | 0.41 | 19.3 | 26.3 | 0.34 | 17.8 | 24.3 | 0.38 |
| *City cv* | 15.3 | 20.9 | 0.38 | 18.3 | 24.7 | 0.39 | 19.9 | 26.9 | 0.32 | 18.3 | 24.8 | 0.36 |
| *Country cv* | 23.4 | 28.8 | -0.18 | 23.8 | 29.5 | 0.13 | 28.6 | 34.7 | -0.15 | 25.3 | 31.2 | -0.02 |

**Table 3:** ***Regression model error.*** *Summary of the mean absolute error (MAE), the root mean squared error (RMSE) in years and the coefficient of determination ($R^2$) of the different cross-validation strategies. For random, urban block-based, neighborhood-based and city-based cross-validation (cv), training and prediction is conducted in the same country or in all countries at once. For country-based cross-validation the model was trained on two countries to predict in the third country. The table shows the prediction result for the country used as test set and for all countries, the average over all validation folds.*

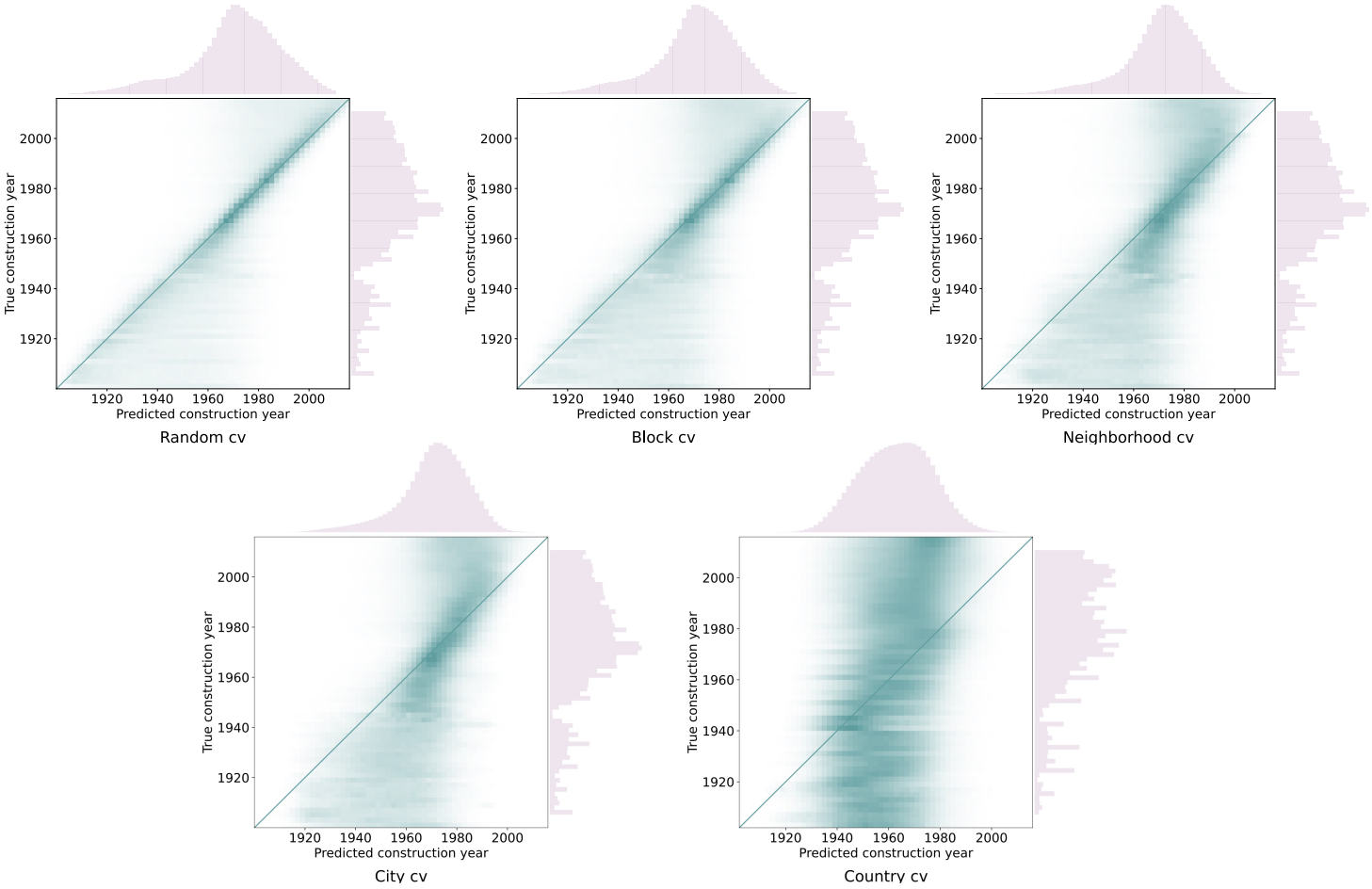


**Figure 6:** ***Increasing generalization distance amplifies the effect of old buildings being predicted as too new and vice versa.*** *Heatmap of a confusion matrix for binned years of construction. True values and prediction values are partitioned into 2-year intervals. Each row of the confusion matrix indicates the share of predictions that fall into each of the 2-year intervals. The sum over each row is 1. Higher color intensities indicate a higher share of prediction in the bin. The diagonal line represents perfect predictions. On the axis, the distributions of the true and predicted construction years are shown. Random, block, neighborhood and city cross-validation (cv) were performed in the Netherlands; country cross-validation on all three countries.*

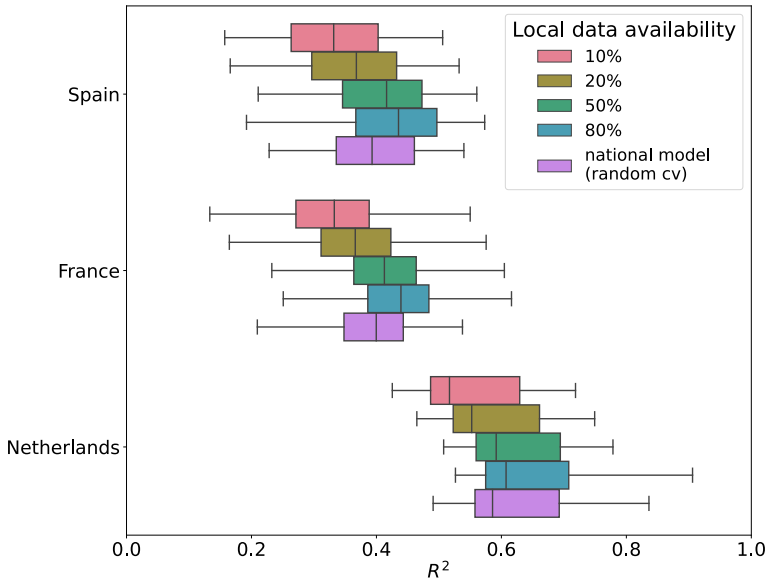


**Figure 7:** ***10% local data allows to fill data gaps locally.*** *Comparison of the prediction performance of regional models that are trained using 10%, 20%, 50%, or 80% of the local data from the region and predict all remaining buildings. Each boxplot depicts the $R^2$ distribution of all regional models in one country and for a specific level of local data availability. For 12 states in the Netherlands, 47 provinces in Spain, and 91 departments in France results are reported. The median number of buildings per region is about 125,000. For comparison, the prediction quality across regions of the national model evaluated by random cross-validation is depicted as well. See Appendix table 14 for the precise $R^2$ and MAE values.*

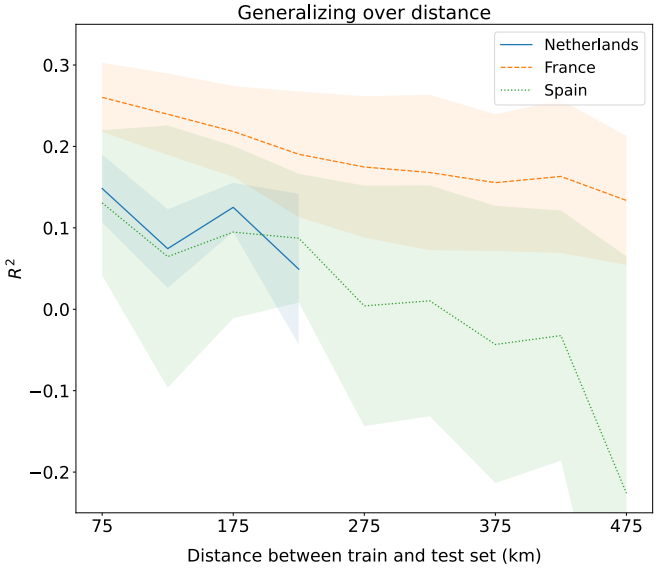


**Figure 8:** ***Increasing distance reduces generalization performance.*** *Illustration of the negative impact of increasing spatial distance between buildings from the train and test set on prediction accuracy. Around 100,000 buildings were used for training and at least 100,000 for testing in each distance experiment. Each experiment group represent non-overlapping circular clusters of cities with a maximal band width of 50 km (see figure 4 for a spatial representation of the experimental setup). The mean distance of each group is indicated on the x-axis. 9 states in the Netherlands, 55 departments in France and 24 provinces in Spain were used for validation. The reported $R^2$ is averaged across all test regions.*

## 3.3 Massive training data improves generalization performance

The generalization performance significantly benefits from additional training data, except for local inference, which exhibits a concave performance trend (see figure 9). Precisely, the predictive performance of local inference peaks at a $R^2$ of 0.68 when utilizing a training sample of ∼400,000 buildings (see Appendix table 15).

When predicting building age in unseen cities, more training data from the same country significantly decreases the prediction error. On average, four cities were required for training to generalize with a positive $R^2$ in the Netherlands. The impact of using more training data on cross-country generalization exhibits a positive trend as well, although the explanatory power is limited even with all data.

Using training data from a different country, in this case France, as indicated by the vertical line in figure 9, stops the positive trend in generalization performance and partially reduces the $R^2$. This highlights the challenges posed by national borders. Only the age prediction of French buildings benefits from adding training data from France.

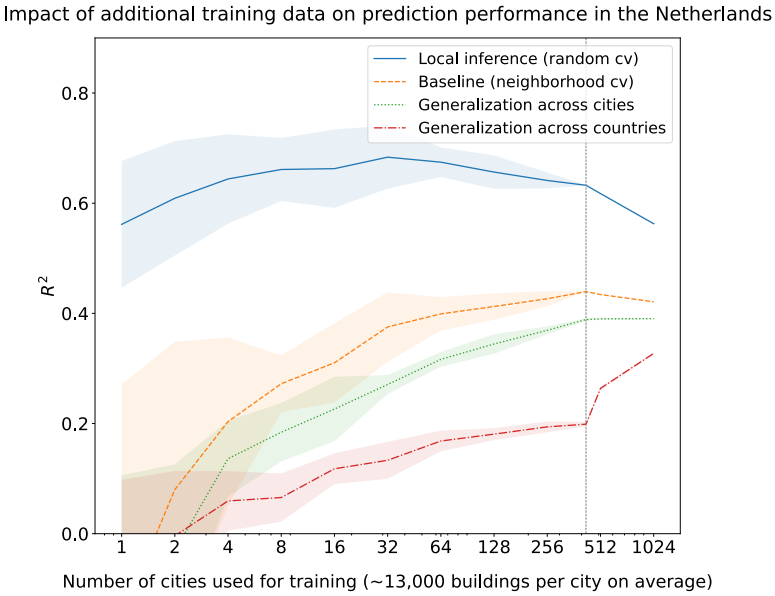


**Figure 9:** ***Additional in-country training data improve model performance in the Netherlands except for local inference.*** *Visualization of the impact of increasing the training sample size on the coefficient of determination, $R^2$. Building data from up to 424 dutch cities with ∼ 13,000 buildings on average are utilized for training. Once all cities from the Netherlands are used for training, buildings from France are added to assess the impact of additional transnational training data on the prediction. The model performance deteriorates thereafter except for cross-country generalization. The dashed vertical line indicates when all buildings from the Netherlands are utilized. The mean across 10 iterations with different seeds is indicated by the plotted line, the standard deviation over all iterations by the shadowed area. Maastricht, Almelo and Ede are used as test set for the cross-city generalization experiment; Aix-en-Provence, Rennes and Limoges for the cross-country generalization experiment. See Appendix table 15 for the exact $R^2$, MAE and RMSE.*

## 3.4 Prediction accuracy significantly varies across construction periods and regions

**A. Construction periods** The regression error varies across construction periods. Buildings built between 1960 and 1990 are predicted most accurately (MAE of 9.88 years), while accuracy deteriorates for buildings built before 1945 (MAE of 33.95 years) and after 2000 (MAE of 21.90 years) (see figure 10 and Appendix table 17). Old buildings being predicted as too new and new buildings as too old (see figure 6).

When redefining the prediction as a classification problem, errors become more balanced and country-specific. Buildings built before 1960 are still classified less accurately than those built after, but to a smaller extent (6 to 19 percentage points for 10 year bins, see Appendix table 13). Buildings constructed in the 1910s and 1940s are particularly misclassified (see Appendix figure 18), possibly due to small sample sizes (see histogram in figure 2) and historical events like World War I and II. Yet, in contrast to regression, prediction quality for buildings constructed after 1990 remains stable or even improves in Spain and France.

Overall, predicting old buildings is a challenging task due to the greater diversity of historic building geometries, smaller sample sizes, and less accurate construction date information. For instance, there are a suspicious number of buildings from 1900 compared to subsequent years: 5 times more in the Netherlands, 38 times more in France, and 66 times more in Spain.

**B. Regions & settlement types** The prediction quality varies widely between regions. Buildings in the Netherlands and northern France are predicted more accurately than in southern France and Spain (see figure 11). Overall, the standard deviation of $R^2$ between regions is 0.08.

Comparing prediction results across settlement types shows that rural regions exhibit the highest prediction errors with an average of 3.95 years above the country's overall MAE (see table 4). In the Netherlands and Spain, buildings in cities are predicted most accurate and in France, buildings in towns. Generally, the Pearson correlation coefficient $r$ indicates a low, but highly significant negative correlation between the MAE and the number of buildings per settlement, with $r = -0.22$ in the Netherlands, $r = -0.37$ in France and $r = -0.22$ in Spain. This especially affects prediction results in Spain, as its rural regions are comparatively small. For example, 69.8% have less than 1000 buildings, whereas in the Netherlands it is 11.8% and in France 22.7%. As a result, rural regions in Spain show the highest MAE with 26.59 years and also the highest standard deviation of the MAE between them with 9.26 years (see table 4).

**C. Building types** Depending on the country, different building types are predicted most accurately. In France, residential buildings are predicted more accurately than non-residential buildings, partly due to the greater diversity and challenging feature distributions of non-residential buildings (MAE of 17.20 years $<$ 22.48 years, see table 16). Among the residential and non-residential subtypes, industrial buildings have the lowest MAE, while other non-residential types are predicted worse than the average (see Appendix table 19). In particular, annex buildings, which are the second most common building type in France with a share of 14.4% of all buildings, significantly increase the overall MAE. Within the residential sector, single-family and terraced houses are predicted well, while apartment blocks are mostly predicted as too new (see figure 12).

In Spain, urban context and construction period determine whether residential or non-residential buildings are predicted better. In cities and until the 1980s, residential buildings show a lower MAE, whereas in towns, rural areas, and after 1980 non-residential buildings are predicted more accurate (see Appendix table 16 and 18). We suppose that this can be attributed to the high predictive performance for agricultural buildings, which are predominantly constructed in thinly or intermediately populated

regions after 1980. Similarly, predictions are particularly accurate for apartment blocks with a MAE of 10.9 years (9.2 years below the average). In contrast to France, industrial buildings and terraced houses are predicted with the highest MAE.

**D. In-fill housing** The construction year of in-fill buildings is predicted with a substantially higher MAE, 25.54 years in the Netherlands, 43.69 years in France, and 36.29 years in Spain, which on average is 18.06 years larger than the MAE of non in-fill buildings. We hypothesize that since the model heavily relies on neighborhood features for making predictions (67% in the Netherlands, 54% in France, and 49% in Spain, see table 5), it struggles to accurately predict buildings which differ from surrounding buildings, such as in-fill buildings.

**E. Feature importance** Different urban form features are most relevant in France, Netherlands and Spain. While the distance to the closest street and footprint area are most important for the prediction in France, the footprint area of the individual buildings plays a negligible role in the Netherlands, instead the elongation standard deviation of all buildings in a 100 m buffer is most predictive (see figure 13). In Spain, besides several features describing the footprint area, the shared wall length plays a noteworthy role in the prediction. Overall, few features are very decisive in France and Spain, whereas in the Netherlands the feature importance is more balanced across multiple urban form characteristics. Still, in all countries every feature utilized captures some information. The second and third least important features, orientation of the buildings and its urban block, are still twice as important for prediction as the random noise feature.

An analysis of feature groups reveals that building and building neighborhood features are most important in all three countries, accounting for 43% of the contribution (see table 5). Further, street features are highly important (22%), especially in France (30%). In France and Spain, the model relies similarly on neighborhood and spatially explicit features, while in the Netherlands, spatially explicit features contribute only 33%, possibly negatively affecting predictions in heterogeneous neighborhoods.

Introducing building height and type as additional features improves the weight of spatially explicit features and significantly enhances prediction performance. Building height improves the $R^2$ by 1–3 percentage points and in combination with building type by 2–3 percentage points, depending on the country (see Appendix table 21).

**F. Urban morphology** The urban urban characteristics show a notable spatio-temporal variation. Mean values of the 9 most predictive urban form features differ significantly between countries for most construction periods (see figure 14). When assuming normally distributed feature values, the distributions overlap on average by 71.4% between the three countries (see overlapping coefficient (OVL) in Appendix figure 26). Domestically, the built environment in Spain exhibits the largest dispersion. Between the countries, features that capture building footprint area characteristics differ the most (average OVL of 52.5% for *StdBlockFootprintArea* and 64.6% for *FootprintArea*). Analog, the importance of these features varies strongly between the three countries. While *FootprintArea* contributes 5.7% to the prediction in France, in only contributes 1.1% in the Netherlands (see table 20). Conversely, *StdBlockFootprintArea* contributes 7.8% in Spain and only 1.3% in France. Consequently, these differences in feature value distributions make it difficult to generalize to unseen countries and may partially explain the deteriorating generalization performance.

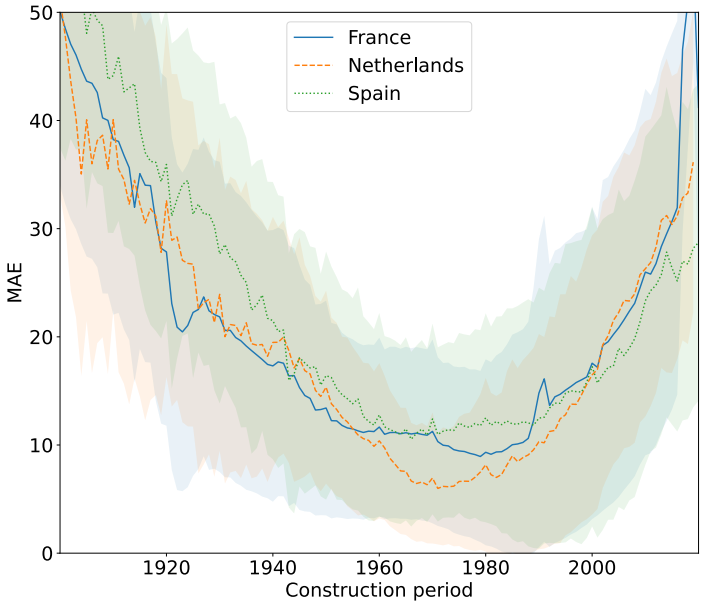


**Figure 10:** ***Regression error strongly varies across construction periods.*** *The mean absolute error (MAE) significantly varies across construction periods; buildings constructed in the second half of the 20th century are predicted most accurately, whereas for older and newer buildings the prediction accuracy deteriorates. The shaded area represents the standard deviation around the MAE.*

| Country | DEGURBA | n | MAE mean | MAE std |
|---|---|---|---|---|
| France | 1 (city) | 124 | 18.36 | 4.93 |
| | 2 (town) | 315 | 16.37 | 3.77 |
| | 3 (rural) | 2083 | 20.27 | 3.98 |
| | All | 2725 | 19.72 | 4.29 |
| Netherlands | 1 (city) | 98 | 14.91 | 4.45 |
| | 2 (town) | 238 | 15.28 | 4.00 |
| | 3 (rural) | 88 | 16.59 | 2.41 |
| | All | 424 | 15.47 | 3.88 |
| Spain | 1 (city) | 107 | 15.52 | 4.55 |
| | 2 (town) | 640 | 17.40 | 5.12 |
| | 3 (rural) | 3530 | 26.59 | 9.14 |
| | All | 5260 | 24.92 | 9.26 |

**Table 4:** ***Prediction error by settlement type.*** *Summary of the mean absolute error (MAE) of buildings located in cities, town, and rural areas according to the EU's DEGURBA classification. The number of settlements per type in the respective country is indicated by n. The mean and std column refer to the mean and standard deviation of the MAE across all settlements of a particular type. See Appendix figure 24 for a visualization of the error distribution per settlement type.*

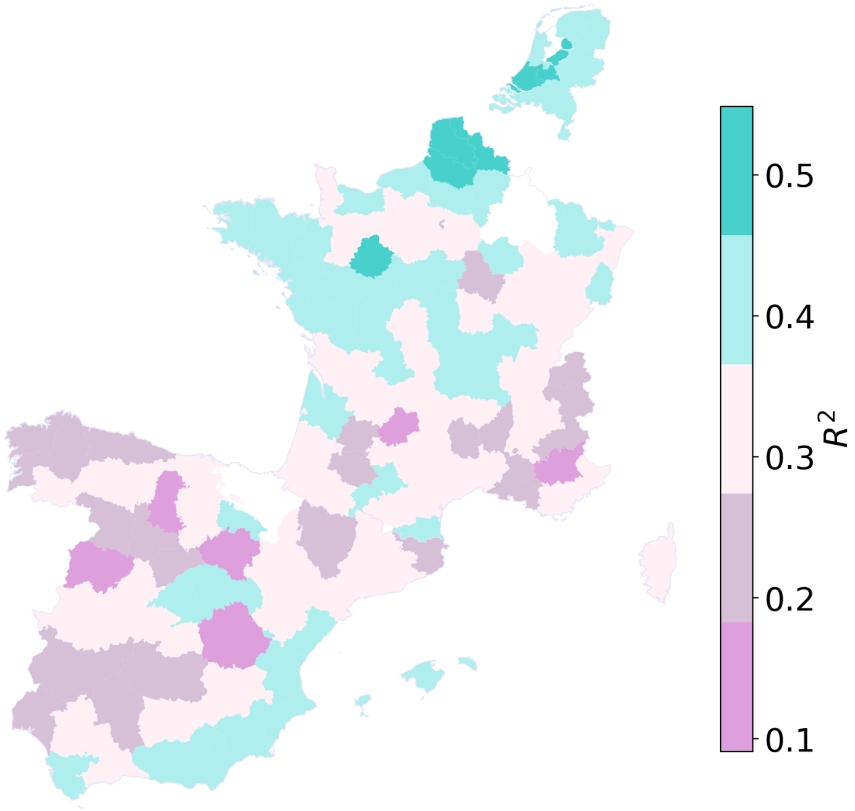


**Figure 11:** ***Prediction error varies across regions.*** *The coefficient of determination ($R^2$) varies across countries and regions, ranging from 0.09 in Melilla, Spain to 0.55 in Flevoland, Netherlands. See Appendix figure 23 for the spatial variation of the MAE.*

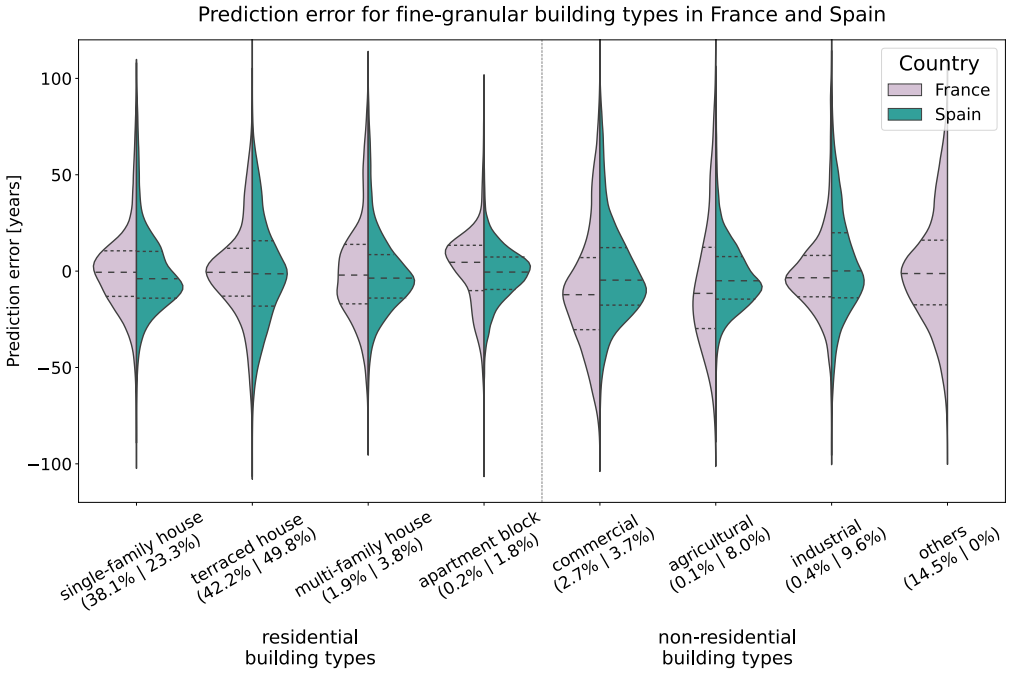


**Figure 12:** ***Prediction error differs between building types.*** *The building type is available in the cadaster data for more than 99% of all buildings in France and Spain. Of those, 78.7% are residential in Spain and 82.2% in France. The remaining non-residential building, were labeled as agricultural, industrial or commercial and service buildings. The respective percentages are indicated in the x-axis labels for France (left) and Spain (right). In France, we grouped annex, sports and religious buildings as others, since they are not available in the Spanish cadaster. The majority of those (14.4%) are annex buildings. The subtypes of residential buildings, single-family houses, terraced houses, multi-family houses, and apartment blocks, are estimated using a simple decision tree based on available building attributes such as height, footprint area and the number of adjacent buildings (see Appendix section G).*

| | Netherlands | France | Spain | All |
|---|---|---|---|---|
| Building | 0.13 | 0.23 | 0.28 | 0.21 |
| Building neighborhood | 0.26 | 0.19 | 0.19 | 0.21 |
| Building block | 0.12 | 0.09 | 0.14 | 0.12 |
| Building block neighborhood | 0.13 | 0.08 | 0.09 | 0.10 |
| Street-based block | 0.01 | 0.01 | 0.01 | 0.01 |
| Street-based block neighborhood | 0.05 | 0.04 | 0.03 | 0.04 |
| Street | 0.06 | 0.13 | 0.07 | 0.09 |
| Street neighborhood | 0.05 | 0.07 | 0.05 | 0.06 |
| Street centrality | 0.09 | 0.10 | 0.05 | 0.08 |
| City | 0.09 | 0.06 | 0.09 | 0.08 |
| Total: all building | 0.39 | 0.42 | 0.47 | 0.43 |
| Total: all building block | 0.25 | 0.17 | 0.23 | 0.22 |
| Total: all street-based blocks | 0.06 | 0.05 | 0.04 | 0.05 |
| Total: all street | 0.20 | 0.30 | 0.17 | 0.22 |
| Total: spatially explicit | 0.33 | 0.46 | 0.51 | 0.43 |
| Total: neighborhood & centrality & city | 0.67 | 0.54 | 0.49 | 0.57 |

**Table 5:** ***Normalized feature importance by feature group.*** *Summary of the summed feature importance for groups of similar features as defined in table 2 for the Netherlands, France, and Spain. Neighborhood feature groups refer to features which summarize information about urban form elements within a 100 or 500 m squared buffer. Non-neighborhood feature groups only contain spatially explicit features about specific urban form elements, i.e., buildings, building blocks, street-based blocks, and streets. Feature importance refers to the normalized, individual contribution of each feature to the prediction according to their SHAP-values. See Appendix table 20 for the precise individual feature contributions.*

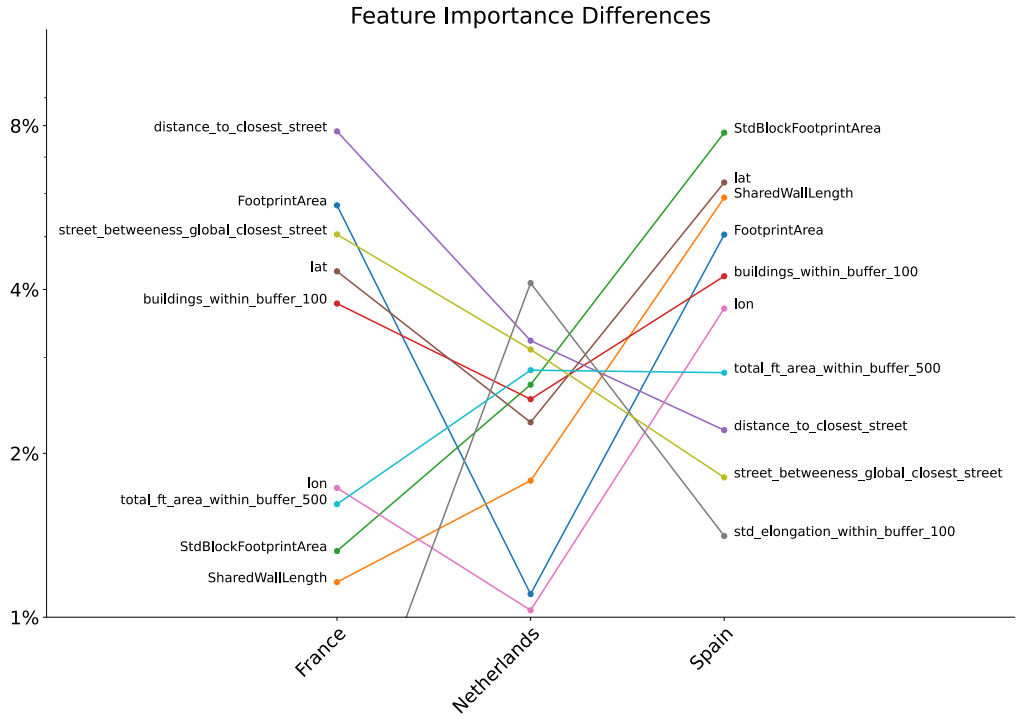


**Figure 13:** ***Different urban form features are distinguishing in France, Netherlands and Spain.*** *Feature importance comparison between France (left), Netherlands (center), and Spain (right) for the 10 most decisive features when performing neighborhood-based cross-validation. Feature importance refers to the normalized, individual contribution of each feature to the prediction according to their SHAP-values. The displayed range is cropped at 1% to improve the readability. The precise feature contributions can be obtained from Appendix table 20.*

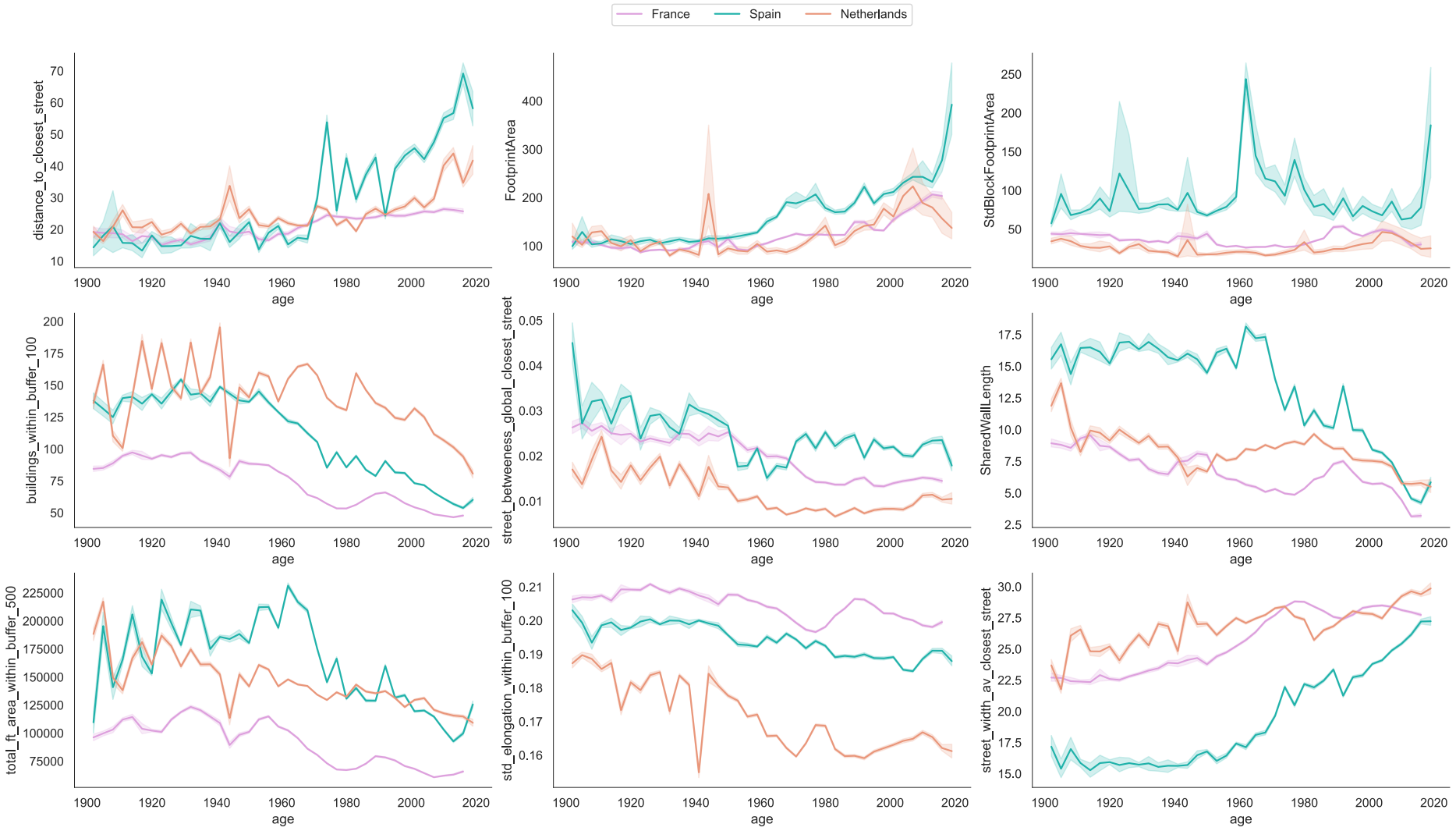


**Figure 14:** ***Feature values of most predictive urban form characteristics differ between the Netherlands, France and Spain.*** *The mean values of the 9 most important urban form characteristics differ significantly between the Netherlands, France and Spain over the construction periods. The shaded area indicates the 99% confidence interval of the mean estimate. A moving average of 3 years is applied to smooth the function. The ordering is according to their average feature importance across all three countries, ranging from distance_to_closest_street with 4.4% to street_width_av_closest_street with 1.9% on average according to their SHAP-values across all three countries. In total, they account for 28.4% of the prediction contribution.*

## 3.5 Predictions can inform prioritization of large-scale retrofits

**A. Energy savings from retrofit prioritization** Using the predicted construction year to estimate the energy demand for heating and identify retrofit candidates increases energy savings by more than 50% compared to a random retrofit strategy. More specifically, 61.9% more energy is saved per refurbished $m^2$ floor space when buildings with an annual heating demand above 150 kWh/$m^2$a are targeted compared to a random prioritization of single-family houses (see figure 15A). Savings compared to a random prioritization across all residential types are notably larger. With ground truth knowledge of the construction year, projected savings of targeted retrofits increase from 90.9 to 120.4 kWh/$m^2$a. We estimate the total energy savings potential for this strategy to be 0.0187 EJ, of which 75.5% may be saved using the inferred building age.

These energy savings estimates are based on locally inferred data, i.e., when building age of neighboring buildings is available. Nevertheless, when using data from generalizations across cities, the energy savings are still 53.3% above the baseline. Even if the energy modeling is informed by cross-country predictions, the savings are still noteworthy above the baseline. This is largely due to the fact that predictions for single-family houses in France are comparatively accurate - even when the regression model is trained in different countries. When a classification approach is adopted instead of a regression, the MAE of energy demand estimates can be reduced by 7% (see Appendix section H).


**Figure 15:** ***Building age prediction improves retrofit prioritization.*** *Results overview of a targeted prioritization of buildings with a heating demand above 150 kWh/$m^2$a. (A) Energy savings per $m^2$ of floor space for different levels of knowledge of building age: ground truth knowledge (left), predicted age according to different cross-validation approaches (middle), and no knowledge (right). (B) Regional heterogeneity in the share of buildings in need of retrofitting in France. (C) Illustration of accuracy in determining retrofit needs exemplified by the city of Valence, France.*

**B. Regional retrofit prioritization** According to ground truth age data, pronounced differences exist between regions in the share of buildings that need to be retrofitted. Our approach can estimate this share with a MAE of 5.3 percentage points, thus helping to focus retrofit efforts on specific regions. Figure 15B highlights the spatial heterogeneity between regions in France according to a targeted prioritization based on the building age prediction with local data availability. The lowest need for retrofitting exists in the south-east of France, specifically in the climate regions H3 with mild winter temperatures. The greatest need is in the center of France, where urban density is low and old single-family houses make up a large portion of the building stock. Given the high spatial resolution, this approach can also identify neighborhoods within cities with high retrofit need (see figure 15C).

# 4 Discussion

## 4.1 Filling data gaps to upscale European climate policies

Methods with high spatial resolution enable climate mitigation research to overcome generic recommendations and inform policy measures at urban, street and building scale, while being adapted to specific contexts on the one hand and scalable to European and global levels on the other [6].

Our experiments demonstrate that the urban form is predictive for building age across regions, architectural styles and settlement types. We show that the findings of previous regional case studies generalize to the Netherlands, France and Spain and are likely transferable to other western European countries. Further, large-scale experiments on building height and type suggest that urban form information can be used to predict buildings attributes in general (see Appendix section I). Our model has the potential to serve as a scalable, low-cost data generation tool to infer missing building attributes in existing datasets using only publicly available urban morphology data.

When local data are available, the predictive accuracy ($\overline{R^2} = 0.51$, see table 3) is sufficient to inform policies at different levels, from local to European. Particularly policies targeting residential buildings in urban areas (see Appendix table 16) can be informed with high confidence. An important practical implication is that no large-scale model is needed for local inference. Building age in a specific region can be precisely inferred with just 10% of the data from the region (see figure 7).

Inferring missing building attributes in unknown regions where no local data are available remains difficult ($\overline{R^2} = 0.36$), especially when the regions are far away from the training region (see figure 8). Generalization performance across distances can be improved by additional, geographically diverse training data (see figure 9). Nevertheless, the model performance varies significantly between provinces (see figure 11). This makes comparisons of predictive performance between regional case studies inconclusive, as regional differences in performance are greater than differences between case studies in the existing literature. Therefore, we refrain from direct performance comparisons with previous work. In order to reliably use the cross-regional predictions for downstream applications, further analysis is needed to better understand the reasons for the varying performance.

Learning building age from urban form in one country and accurately predicting in another country is not possible with our current model. With an $R^2$ around 0 (see table 3), climate policies cannot be informed. Instead, to upscale climate solutions across countries, local data in all countries are needed. Future research may evaluate what amount of local data is required and if sparse local data from OpenStreetMap (OSM) can be utilized in regions where no administrative data are available.

Regarding feature engineering, incorporating building height as a feature improves the overall prediction $R^2$ by up to 2.5 percentage points (see Appendix table 21). A more detailed representation of 3D urban morphology could potentially yield further benefits. While studies have demonstrated the informative nature of building facade visual appearance [31, 32], the combination of urban morphology and visual features remains unexplored. The high feature importance of latitude and longitude and the spatial autocorrelation between prediction residuals suggest that a more sophisticated encoding of spatial relations could enhance the predictive performance and reduce misleading correlations of other features caused by the spatial proximity of buildings. For example Space2Vec [51] and eigenvector spatial filtering (ESF) [52] could be tested to incorporate spatial relationships explicitly as additional features and graph neural networks (GNNs) could be tested to encode the building surroundings as a spatial graph.

## 4.2 Towards large-scale, spatially explicit energy modeling

Particularly in the EU, where 50% of the building stock was constructed before 1970, when energy efficiency restrictions were beginning to be introduced, precise energy modeling and the identification of old buildings is essential to the effective implementation of retrofit strategies [53]. To reach the 1.5°C goal, the current annual renovation rates in the EU ranging from 0.4% – 1.2% have to be increased to 5% [54]. To support large retrofit strategies, such as the European Renovation Wave [55], scalable, spatially explicit methods are needed to identify renovation candidates across the EU.

With an average mean error of above 10 years in all countries (see table 3), our approach is yet not applicable for use cases that are very sensitive to the specific year of construction. However, it can help scale use cases where primarily the construction period is of interest, such as energy modeling and the identification of retrofit candidates.

Experiments in France show energy savings from large-scale retrofits may be improved by more then 50% when age predictions are used to target old, energy-inefficient buildings. If prediction accuracy is improved, these savings can be up to doubled (see figure 15A). Currently, precise age prediction is more difficult for non-residential buildings, with a few exceptions such as agricultural buildings. Particularly in France, residential buildings were predicted 4.75 years more accurately than non-residential buildings (see Appendix table 16). Therefore, energy savings from prioritizing non-residential retrofits based on our predictions are likely smaller. Within the residential sector, our analysis suggests that single-family houses can most reliably be targeted by EU-level retrofit measures using our methodology (see Appendix table 19). In Spain, apartment blocks can be targeted as well with very high accuracy. For infill housing, the model performance deteriorates. However, since they are by definition newer than the buildings surrounding them, this is less consequential as they are not the first candidates for energy retrofits.

Furthermore, our approach can help to identify regional clusters of buildings in need of refurbishment. A spatial analysis of France shows that the share of energy-inefficient buildings differs up to a factor of 5 between regions (see figure 15C). Here, our approach can help focus policy efforts and fund allocation on regions with high refurbishment needs. A particular value is that the spatially explicit modeling of building energy demand allows for assessment and prioritization of retrofit needs at any spatial resolution, i.e., at building, neighborhood, city, or regional level.

Since the regression approach is biased towards the mean and overestimates the age of old buildings and underestimates the age of new buildings (see figure 6), we recommend testing a classification approach tailored to residential buildings and the age classes required for the specific use case. Our experiments show that the MAE of energy estimates can be reduced by 7% using classification (see Appendix section H) and suggest that this may further improve prioritization of large-scale retrofit efforts.

While this highlights the enormous potential for decarbonizing the building sector, we acknowledge that all of our results are constrained to the underlying energy model. As we have no actual data on wall materials and thermal transmittance, and only consider 4 residential construction types, we call for future research to validate and refine our results with more detailed energy models.

## 4.3 Tear down national borders (in the data)

Harmonized and transnational data infrastructure is required to enable global urban climate science [5]. The lack of standardization, quality assurance, and availability of data poses a major challenge to comparing climate risks and policies between cities. Researchers, [5, 56–58], have repeatedly appealed for the development of global urban science, for which a harmonized data foundation is key. The EUBUCCO dataset [33] provides a starting point to enable studies like this with ideally consistent data quality.

Yet, we find that national borders in the cadaster data predetermine and limit the possibilities of cross-country age prediction. Our experiments show that generalizing

across regions is possible, but not across national borders with our current approach (see figure 6). For local inference, regional generalization (see table 3), and our experiment on additional training data (see figure 9) the prediction accuracy drops noteworthy when using transnational data compared to using only data from a single country. Examining the results indicates that our model relies on different urban form characteristics to predict building age in different countries (see table 5 and figure 13). We find that not only the predictiveness but also the value distributions of urban form characteristics differ significantly between the countries (see Appendix figure 25 and 26). While the urban morphology naturally differs between the countries, we suppose that this effect is amplified and biased by inconsistent data acquisition and preparation.

Inspection of the cadaster datasets reveals multiple discrepancies that could be harmonized. Data sources deal differently with missing data or data uncertainty. In Spain, for instance, the year of construction is estimated to the nearest decade or century, resulting in strong country-specific noise. Further, we notice different spatial granularity in terms of how an individual building is defined and whether adjacent built structures, e.g., annexes, are considered to be the same building or separate ones. In Spain, the average footprint area of buildings for some periods is 50% larger than in France and the Netherlands. In France, annex buildings are considered individual buildings with a distinct non-residential type. In general, building types are inconsistently defined and the resulting distributions of building types indicate conspicuous variations, e.g., 60 times more agricultural buildings in Spain than in France. Thus, harmonization of data quality and spatial resolution of the building stock may substantially help for reliable cross-country comparisons of urban climate solutions, policy advice and policy evaluation in the future.

# 5 Conclusion

Spatially explicit data that are publicly available at scale are the foundation to moving beyond generic climate recommendations and providing fine-grained solutions for decarbonizing cities. Since the year of construction of buildings is central to energy modeling, we developed a scalable method that is able to predict missing building attributes in available administrative data within different countries using only open urban morphology data, enabling large-scale energy modeling of the building stock with high spatial resolution.

We find that filling data gaps within known cities is possible with a MAE of 9.6 years in the Netherlands, 15.8 years in France, and 16.7 years in Spain. While higher local data availability improves the prediction, 10% are sufficient to infer the remaining 90% comparatively accurately, with an average MAE ranging from 11.0 years in the Netherlands to 20.0 years in Spain. Across all regression experiments, the best predictive performance is achieved for residential buildings constructed between 1960 and 1990 in dense urban areas.

Generalizing across regions is more difficult. The further apart the train and test region are, the stronger the effect. For predictions across cities, the $R^2$ is 7 to 25 percentage points lower compared to predictions with local data availability. Though, this effect is mitigated by larger and more diverse training data. Yet, when trying to generalize across national borders, urban form is no longer predictive in our experiments. We find that the feature value distributions of urban form characteristics differ significantly between countries, making it very difficult to train in one country and predict in another. In parts, this is the result of inconsistent data sources. To foster global urban climate science, data harmonization and standardization between countries must be improved.

## Acknowledgements

This work received funding from the CircEUlar project of the European Union's Horizon Europe research and innovation program under grant agreement 101056810. We further thank the Potsdam Institute for Climate Impact Research for providing the computing infrastructure. Map data is copyrighted by OpenStreetMap contributors and is available at `https://www.openstreetmap.org`. Figures have been designed using icons from Flaticon.com.

## Code availability

The code of this work is available on GitHub: `https://github.com/ai4up/ufo-prediction`

## Declarations of interest

None.

# References


[1] L. Cabeza *et al.*, *Buildings. ipcc, 2022: Climate change 2022: Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change*, 2022.

[2] T. Kuramochi *et al.*, "Ten key short-term sectoral benchmarks to limit warming to 1.5 c," *Climate Policy*, vol. 18, no. 3, pp. 287–305, 2018.

[3] D. Reckien *et al.*, "How are cities planning to respond to climate change? assessment of local climate plans from 885 cities in the eu-28," *Journal of cleaner production*, vol. 191, pp. 207–219, 2018.

[4] Y. Shan *et al.*, "City-level climate change mitigation in china," *Science advances*, vol. 4, no. 6, eaaq0390, 2018.

[5] F. Creutzig *et al.*, "Upscaling urban data science for global climate solutions," *Global Sustainability*, vol. 2, 2019.

[6] N. Milojevic-Dupont and F. Creutzig, "Machine learning for geographically differentiated climate change mitigation in urban areas," *Sustainable Cities and Society*, vol. 64, p. 102 526, 2021.

[7] F. Creutzig *et al.*, "Demand, services and social aspects of mitigation," in *IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2022, pp. 752–943.

[8] A. Grubler *et al.*, "A low energy demand scenario for meeting the 1.5 c target and sustainable development goals without negative emission technologies," *Nature energy*, vol. 3, no. 6, pp. 515–527, 2018.

[9] R. Nouvel, M. Zirak, V. Coors, and U. Eicker, "The influence of data quality on urban heating demand modeling using 3d city models," *Computers, Environment and Urban Systems*, vol. 64, pp. 68–80, 2017.

[10] M. Aksoezen, M. Daniel, U. Hassler, and N. Kohler, "Building age as an indicator for energy consumption," *Energy and Buildings*, vol. 87, pp. 74–86, 2015.

[11] T. R. Tooke, N. C. Coops, and J. Webster, "Predicting building ages from lidar data with random forests for building energy modeling," *Energy and Buildings*, vol. 68, pp. 603–610, 2014.

[12] M. Zirak, V. Weiler, M. Hein, and U. Eicker, "Urban models enrichment for energy applications: Challenges in energy simulation using different data sources for building age information," *Energy*, vol. 190, p. 116 292, 2020.

[13] S. K. Firth, K. J. Lomas, and A. Wright, "Targeting household energy-efficiency measures using sensitivity analysis," *Building Research & Information*, vol. 38, no. 1, pp. 25–41, 2010.

[14] É. Mata, A. S. Kalagasidis, and F. Johnsson, "Building-stock aggregation through archetype buildings: France, germany, spain and the uk," *Building and Environment*, vol. 81, pp. 270–282, 2014.

[15] J. F. Rosser, G. Long, S. Zakhary, D. S. Boyd, Y. Mao, and D. Robinson, "Modelling urban housing stocks for building energy simulation using citygml energyade," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 163, 2019.

[16] S. Steimen, D. Fäh, D. Giardini, M. Bertogg, and S. Tschudi, "Reliability of building inventories in seismic prone regions," *Bulletin of Earthquake Engineering*, vol. 2, no. 3, pp. 361–388, 2004.

[17] M. Wieland, M. Pittore, S. Parolai, J. Zschau, B. Moldobekov, and U. Begaliev, "Estimating building inventory for rapid seismic vulnerability assessment: Towards an integrated approach based on multi-source imaging," *Soil Dynamics and Earthquake Engineering*, vol. 36, pp. 70–83, 2012.

[18] M. Liuzzi, P. Aravena Pelizari, C. Geiß, A. Masi, V. Tramutoli, and H. Taubenböck, "A transferable remote sensing approach to classify building structural types for seismic risk analyses: The case of val d'agri area (italy)," *Bulletin of Earthquake Engineering*, vol. 17, no. 9, pp. 4825–4853, 2019.

[19] M. Fedeski and J. Gwilliam, "Urban sustainability in the presence of flood and geological hazards: The development of a gis-based vulnerability and risk assessment methodology," *Landscape and urban planning*, vol. 83, no. 1, pp. 50–61, 2007.

[20] M. Uzielli, F. Catani, V. Tofani, and N. Casagli, "Risk analysis for the ancona landslide—ii: Estimation of risk to buildings," *Landslides*, vol. 12, no. 1, pp. 83–100, 2015.

[21] M. J. Nahlik, M. V. Chester, S. S. Pincetl, D. Eisenman, D. Sivaraman, and P. English, "Building thermal performance, extreme heat, and climate change," *Journal of Infrastructure Systems*, vol. 23, no. 3, p. 04 016 043, 2017.

[22] R. Ortlepp, K. Gruhler, and G. Schiller, "Materials in germany's domestic building stock: Calculation model and uncertainties," *Building Research & Information*, vol. 46, no. 2, pp. 164–178, 2018.

[23] N. Milojevic-Dupont *et al.*, "Learning from urban form to predict building heights," *Plos one*, vol. 15, no. 12, e0242010, 2020.

[24] M. Wurm, A. Schmitt, and H. Taubenböck, "Building types' classification using shape-based features and linear discriminant functions," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 5, pp. 1901–1912, 2015.

[25] Z. Lu, J. Im, J. Rhee, and M. Hodgson, "Building type classification using spatial and landscape attributes derived from lidar remote sensing data," *Landscape and Urban Planning*, vol. 130, pp. 134–148, 2014.

[26] A. Droin, M. Wurm, and W. Sulzer, "Semantic labelling of building types. a comparison of two approaches using random forest and deep learning," *Publikationen der DGPF*, 2020.

[27] F. Biljecki, H. Ledoux, and J. Stoter, "Generating 3d city models without elevation data," *Computers, Environment and Urban Systems*, vol. 64, pp. 1–18, 2017.

[28] J. F. Rosser, D. S. Boyd, G. Long, S. Zakhary, Y. Mao, and D. Robinson, "Predicting residential building age from map data," *Computers, Environment and Urban Systems*, vol. 73, pp. 56–67, 2019.

[29] O. M. Garbasevschi *et al.*, "Spatial factors influencing building age prediction and implications for urban residential energy modelling," *Computers, Environment and Urban Systems*, vol. 88, p. 101 637, 2021.

[30] F. Biljecki and M. Sindram, "Estimating building age with 3d gis," in *Proceedings of the 12th International 3D GeoInfo Conference 2017*, 2017, pp. 17–24.

[31] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller, "Automatic prediction of building age from photographs," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 126–134.

[32] Y. Li, Y. Chen, A. Rajabifard, K. Khoshelham, and M. Aleksandrov, "Estimating building age from google street view images using deep learning (short paper)," in *10th international conference on geographic information science (GIScience 2018)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[33] Milojevic-Dupont, Nikola and Wagner, Felix *et al.*, *EUBUCCO v0.1*, version v0.1, Zenodo, May 2022. DOI: 10.5281/zenodo.6524781. [Online]. Available: https://doi.org/10.5281/zenodo.6524781.

[34] OpenStreetMap contributors, *Planet dump retrieved from https://planet.osm.org*, `https://www.openstreetmap.org`, 2017.

[35] R. Hijmans, N. Garcia, and J. Wieczorek, "Gadm: Database of global administrative areas, version 3.6," *GADM Maps and Data*, 2018.

[36] L. Dijkstra and H. Poelman, "A harmonised definition of cities and rural areas: The new degree of urbanisation," *WP*, vol. 1, p. 2014, 2014.

[37] Eurostat, *Local Administrative Units (LAU) - NUTS - Nomenclature of territorial units for statistics - Eurostat*, 2019. [Online]. Available: `https://ec.europa.eu/eurostat/web/nuts/local-administrative-units` (visited on 05/11/2022).

[38] ——, *Regions in the european union. nomenclature of territorial units for statistics*, 2007.

[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *arXiv preprint arXiv:2110.01889*, 2021.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.

[44] K. Le Rest, D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle, "Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation," *Global ecology and biogeography*, vol. 23, no. 7, pp. 811–820, 2014.

[45] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," *International Journal of Geographical Information Science*, vol. 31, no. 10, pp. 2001–2019, 2017.

[46] A. M.-C. Wadoux, G. B. Heuvelink, S. De Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological Modelling*, vol. 457, p. 109 692, 2021.

[47] *Tabula webtool*. [Online]. Available: `https://webtool.building-typology.eu/#bm`.

[48] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[49] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[50] T. Loga, K. Müller, K. Reifschläger, and B. Stein, "Evaluation of the tabula database–comparison of typical buildings and heat supply systems from 20 european countries," *TABULA Work Report*, 2015.

[51] G. Mai, K. Janowicz, B. Yan, R. Zhu, L. Cai, and N. Lao, "Multi-scale representation learning for spatial feature distributions using grid cells," *arXiv preprint arXiv:2003.00824*, 2020.

[52] M. D. Islam, B. Li, C. Lee, and X. Wang, "Incorporating spatial information in machine learning: The moran eigenvector spatial filter approach," *Transactions in GIS*,

[53] F. Filippidou and J. Jimenez Navarro, "Achieving the cost-effective energy transformation of europe's buildings," *Publications Office of the European Union: Luxembourg*, 2019.

[54] H. de Coninck *et al.*, "Strengthening and implementing the global response," in *Global warming of 1.5 C: Summary for policy makers*, IPCC-The Intergovernmental Panel on Climate Change, 2018, pp. 313–443.

[55] E. Commission, *A renovation wave for europe—greening our buildings, creating jobs, improving lives*, 2020.

[56] W. Solecki, K. C. Seto, and P. J. Marcotullio, "It's time for an urbanization science," *Environment: science and policy for sustainable development*, vol. 55, no. 1, pp. 12–17, 2013.

[57] T. Elmqvist, X. Bai, N. Frantzeskaki, and D. Maddox, *The urban planet: knowledge towards sustainable cities*. Cambridge University Press, 2018.

[58] M. Acuto, S. Parnell, and K. C. Seto, "Building a global urban science," *Nature Sustainability*, vol. 1, no. 1, pp. 2–4, 2018.

[59] R. Louf and M. Barthelemy, "A typology of street patterns," *Journal of The Royal Society Interface*, vol. 11, no. 101, p. 20 140 924, 2014.

[60] M. Fleischmann, "Momepy: Urban morphology measuring toolkit," *Journal of Open Source Software*, vol. 4, no. 43, p. 1807, 2019.

[61] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[62] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.

[63] A. D. Cliff and J. K. Ord, *Spatial processes: models & applications*. Taylor & Francis, 1981.

[64] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of mcc and cen error measures in multi-class prediction," 2012.

[65] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[66] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, IEEE, 2008, pp. 1322–1328.

# Appendix

## A Complete feature list

The following list contains all features used in this study with their variable name, unit, short description and, when relevant, their definition and source. For a more contextualized description and rationale for their selection, refer to [23].

### Building features

Building geometry

1. Area of the building's footprint. Unit: squared meter. Variable name: `FootprintArea`
2. Perimeter of the building's footprint. Unit: meter. Variable name: `Perimeter`
3. Anisotropy index. Definition: the ratio between the area of the building footprint and the area of the circumscribed circle. Unit: $x \in [0, 1]$. Source: [59]. Variable name: `Phi`
4. Length of the longest axis of the building footprint. Definition: Axis is defined as a diameter of minimal circumscribed circle around the convex hull. Unit: meter. Source: [60]. Variable name: `LongestAxisLength`
5. Elongation of the minimum bounding box around the building footprint. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `Elongation`
6. Convexity of the footprint. Definition: Area of the footprint divided by the area of the convex hull around the footprint. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `Convexity`
7. Orientation of the footprint. Definition: orientation of the longext axis of bounding rectangle in range 0 - 45. It captures the deviation of orientation from cardinal directions. Unit: degree. Source: [60]. Variable name: `Orientation`
8. Number of corners of the footprint. Unit: count. Source: [60]. Variable name: `Corners`
9. Number of buildings directly adjacent to the building. Unit: count. Variable name: `CountTouches`

Building location

1. Latitude of building's location ranging from -90 to 90. Unit: degree. Variable name: `lat`
2. Longitude of building's location ranging from -180 to 180. Unit: degree. Variable name: `lon`

Buildings within 100 and 500 m buffer

1. Standard deviation of building footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `std_orientation_within_buffer_{100,500}`
2. Average orientation of building footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `av_orientation_within_buffer_{100,500}`

3. Standard deviation of the convexity of building footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `std_convexity_within_buffer_{100,500}`

4. Average convexity of building footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `av_convexity_within_buffer_{100,500}`

5. Standard deviation of the elongation of buildings footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `std_elongation_within_buffer_{100,500}`

6. Average elongation of buildings footprints within 100 & 500 m around the building. Unit: $x \in [0, 1]$. Variable name: `av_elongation_within_buffer_{100,500}`

7. Standard deviation of building footprints area within 100 & 500 m around the building. Unit: squared meters. Variable name: `std_footprint_area_within_buffer_{100,500}`

8. Average building footprints area within 100 & 500 m around the building. Unit: squared meters. Variable name: `av_footprint_area_within_buffer_{100,500}`

9. Total building footprints area within 100 & 500 m around the building. Unit: squared meters. Variable name: `total_ft_area_within_buffer_{100,500}`

10. Number of buildings within 100 & 500 m around the building. Unit: counts. Variable name: `buildings_within_buffer_{100,500}`

**Block features**

Block geometry

1. Average footprint area of buildings in the block. Unit: squared meter. Variable name: `AvBlockFootprintArea`

2. Standard deviation of footprint areas of buildings in the block. Unit: squared meter. Variable name: `StdBlockFootprintArea`

3. Total footprint of the block. Unit: squared meters. Variable name: `BlockTotalFootprintArea`

4. Number of buildings in the block. Unit: count. Variable name: `BlockLength`

5. Total perimeter of the block. Unit: meters. Variable name: `BlockPerimeter`

6. Length of the longest axis of whole block footprint. Unit: meters. Variable name: `BlockLongestAxisLength`

7. Elongation of the minimum bounding box around the whole block footprint. Unit: $x \in [0, 1]$. Variable name: `BlockElongation`

8. Convexity of the whole block footprint. Unit: $x \in [0, 1]$. Variable name: `BlockConvexity`

9. Orientation of the whole block footprint. Unit: degree. Variable name: `BlockOrientation`

10. Number of corners of the whole block footprint. Unit: count. Variable name: `BlockCorners`

Blocks within 100 and 500 m buffer

1. Standard deviation blocks' orientation within 100 & 500 m around the building. Unit: degrees. Variable name: `std_block orientation_within_buffer_{100,500}`,

2. Average blocks' orientation within 100 & 500 m around the building. Unit: degrees. Variable name: `av_block_orientation_within_buffer_{100,500}`

3. Average individual building footprint area in blocks within 100 & 500 m around the building. Unit: squared meters. Variable name: `av_block_av_footprint_area_within_buffer_{100,500}`

4. Standard deviation block total footprint area within 100 & 500 m around the building. Unit: squared meters. Unit: Variable name: `std_block_footprint_area_within_buffer_{100,500}`

5. Standard deviation block average footprint area within 100 & 500 m around the building. Unit: squared meters. Unit: Variable name: `std_block_av_footprint_area_within_buffer_{100,500}`

6. Average of the block total footprint area within 100 & 500 m around the building. Unit: squared meters. Variable name: `av_block_footprint_area_within_buffer_{100,500}`

7. Standard deviation of building count in blocks within 100 & 500 m around the building. Unit: count. Variable name: `std_block_length_within_buffer_{100,500}`

8. Average building count in blocks within 100 & 500 m around the building. Unit: count. Variable name: `av_block_length_within_buffer_{100,500}`

9. Number of blocks within 100 & 500 m around the building. Unit: count. Variable name: `blocks_within_buffer_{100,500}`

**Street features**

Closest street & intersection

1. Openness of the closest street to building. Definition: proportion of the street where buildings are or not present on the sides of the street. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_openness_closest_road`

2. Standard deviation of the width of the closest street to the building. Definition: Width is defined here as the average distance between buildings on both sides of the street. Unit: meters Source: [60]. Variable name: `street_width_std_closest_road`

3. Average width of the closest street to the building. Unit: meters Source: [60]. Variable name: `street_width_av_closest_road`

4. Length of the closest street to the building. Unit: meters. Variable name: `street_length_closest_road`

5. Distance between the building and the closest street. Unit: meters. Variable name: `distance_to_closest_road`

6. Distance between the building and the closest intersection. Unit: meters. Variable name: `distance_to_closest_intersection`

Street centrality

1. Local closeness centrality for the closest street to the building. Definition: Local closeness for a radius of 500 m around each node. Value for one edge/street are averages of the values at the two nodes/intersections. Unit: Variable name: Unit: $x \in [0, 1]$. Source: [60]. `street_closeness_500_closest_street`

2. Betweeness centrality of the closest street to the building. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_betweeness_global_closest_road`

3. Global closeness centrality of the closest street to the building. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_closeness_global_closest_road`

4. Average local closeness centrality of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Note: 500 m radius for the closeness centrality. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_closeness_{100,500}_av_inter_buffer_{100,500}`

5. Largest local closeness centrality of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Note: 500 m radius for the closeness centrality. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_closeness_{100,500}_max_inter_buffer_{100,500}`

6. Average betwenness centrality of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_betweeness_global_av_inter_buffer_{100,500}`

7. Largest betwenness centrality of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Unit: $x \in [0, 1]$. Source: [60]. Variable name: `street_betweeness_global_max_inter_buffer_{100,500}`

Streets & intersections within 100 and 500 m buffer

1. Standard deviation of the width of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Unit: meters Source: [60]. Variable name: `street_width_std_inter_buffer_{100,500}`

2. Average width of the streets intersecting a 100 & 500 m buffer around the centroid of the building. Unit: meters Source: [60]. Variable name: `street_width_av_inter_buffer_{100,500}`

3. Standard deviation length of streets *within* a 100 & 500 m buffer around the centroid of the building. Unit: meters. Variable name: `street_length_std_within_buffer_{100,500}`

4. Average length of streets *within* a 100 & 500 m buffer around the centroid of the building. Unit: meters. Variable name: `street_length_av_within_buffer_{100,500}`

5. Total length of streets *within* a 100 & 500 m buffer around the centroid of the building. Unit: meters. Variable name: `street_length_total_within_buffer_{100,500}`

6. Intersection count *within* a 100 & 500 m buffer around the centroid of the building. Unit: count. Variable name: `intersection_count_within_buffer_{100,500}`

**Street-based block features**

Street-based block, own block

1. Anisotropy index of the street-based block in which the building is. Unit: Unit: $x \in [0, 1]$. Variable name: `street_based_block_phi`

2. Area of the street-based block in which the building is. Unit: squared meters. Variable name: `street_based_block_area`

3. Number of corners of the street-based block in which the building is. Unit: squared meters. Variable name: `street_based_block_corners`

Street-based blocks within 100 and 500 m buffer

1. Standard deviation of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: degrees Variable name: `street_based_block_std_orientation_inter_buffer_{100,500}`

2. Standard deviation of the anisotropy index of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: Unit: $x \in [0, 1]$. Variable name: `street_based_block_std_phi_inter_buffer_{100,500}`

3. Average anisotropy index of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: Unit: $x \in [0, 1]$. Variable name: `street_based_block_av_phi_inter_buffer_{100,500}`

4. Standard deviation of the area of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: squared meters. Variable name: `street_based_block_std_area_inter_buffer_{100,500}`

5. Average area of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: squared meters. Variable name: `street_based_block_av_area_inter_buffer_{100,500}`

6. Number of the street-based blocks intersecting a 100 & 500 m buffer around the centroid of the building. Unit: count Variable name: `street_based_block_number_inter_buffer_{100,500}`

**City level features**

City level

1. Total of building footprint area in the city. Unit: squared meters. Variable name: `total_buildings_footprint_city`

2. Total number of buildings in the city. Unit: count. Variable name: `total_buildings_city`

3. Average building footprint area in the city. Unit: squared meters. Variable name: `av_building_footprint_city`

4. Standard deviation of the building footprints area in the city. Unit: squared meters. Variable name: `std_building_footprint_city`

5. Number of detached buildings in the city. Unit: count. Variable name: `n_detached_buildings`

6. Number of blocks from 2 to 4 buildings in the city. Unit: count. Variable name: `block_2_to_4`

7. Number of blocks from 5 to 9 buildings in the city. Unit: count. Variable name: `block_5_to_9`

8. Number of blocks from 10 to 19 buildings in the city. Unit: count. Variable name: `block_10_to_19`

9. Number of blocks of 20 or more buildings in the city. Unit: count. Variable name: `block_20_to_inf`

10. Total intersections count in the city. Unit: count. Variable name: `intersections_count`
11. Total street length in the city. Unit: meters. Variable name: `total_length_street_city`
12. Average street length in city. Unit: meters. Variable name: `av_length_street_city`
13. Number of street based blocks in the city. Unit: count. Variable name: `total_number_block_city`
14. Average area street-based blocks in the city. Unit: squared meters. Variable name: `av_area_block_city`
15. Standard deviation of the street-based blocks in the city. Unit: squared meters. Variable name: `std_area_block_city`

## B Model selection

While [23] found that XGBoost [39] models yield the smallest prediction error when inferring building attributes, the majority of prior studies [11, 28–30] used Random Forest models [40]. Therefore, we conduct preliminary experiments to compare the prediction performance of different ensemble learning methods: XGBoost, Random Forest, and AdaBoost [61]. We utilize a throw-away set of 10% of the data (similarly to hyperparameter tuning (see section 2.2) to avoid data leakage. We define a search space for each of the three algorithms and perform nested 5-fold nested cross-validation with a random search across 20 hyperparameter combinations. The outer cross-validation loop ensures that we evaluate the tuned model on all samples of the data by performing a random hyperparameter search for each fold. This may yield different optimal hyperparameters for each fold. For each of the 20 combinations of the random search, the inner cross-validation loop reports an averaged error across all inner folds, which is used to robustly compare the hyperparamter combinations. Finally, we report the averaged prediction error across all outer folds and use it to compare XGBoost, Random Forest, and AdaBoost.

The results are depicted in table 6. We find that XGBoost outperforms Random Forest by 2 percentage points in terms of $R^2$ for regression the construction year in the Netherlands. Further, the fit time of the optimal Random Forest model is more than 2 times slower than the optimal XGBoost model. Therefore, we utilize XGBoost for all further experiments.

| Model | MAE | RMSE | R2 | Fit time [s] |
|---|---|---|---|---|
| AdaBoost | 18.5 | 24.0 | 0.211 | 409 |
| RandomForest | 9.2 | 17.0 | 0.602 | 722 |
| XGBoost | 8.9 | 16.7 | 0.620 | 281 |

**Table 6:** ***Model comparison between Random Forest, AdaBoost, and XGBoost for regression.*** *Summary of the mean absolute error (MAE), the root mean squared error (RMSE) in years and the coefficient of determination ($R^2$) of different models. Further, the training time in seconds is reported. 5-fold nested cross-validation with a random search over 20 hyperparameter combinations was performed for all models.*

## C Hyperparameter tuning

It is difficult to make an educated guess for the optimal hyperparameters based on hyperparameter tuning on a subsample of data because the optimal model complexity varies with the amount of training data. More training data can lead to further

insights, as the model can detect correlations that were indistinguishable from noise with less training data. Yet, a more complex model might be required to capture those correlations.

Therefore, the best practice is to perform nested cross-validation, which allows to tune the hyperparameters in the inner loop on almost the full dataset without the risk of introducing data leakage and at the same time, robustly assess the generalization capacity in the outer loop. However, to due computational limitations it was not suitable to perform nested cross-validation for every experiment. Since the training time of the top performing hyperparameter combinations is around 1000 seconds for regression and 2,500 seconds for classification with 10 year bins, conducting 5-fold nested cross-validation with a random search across 20 hyperparameter combinations, each experiment iteration would take more than 5 days. For classification each iteration would even take up to 15 days.

To retrospectively evaluate if a more complex model could have significantly improved the prediction performance, we tested the impact of different maximal tree depths on the model performance using all data from the Netherlands. We reused the other hyperparameters from the initial hyperparameter tuning (`n_estimators`=1000, `learning_rate`=0.025, `colsample_bytree`=0.9 and `colsample_bylevel`=0.5) and kept them stable across all iterations.

The experiment shows that the optimal `max_depth` for Netherlands is 16 (see Appendix figure 16). Yet, the decrease in RMSE compared to a `max_depth` of 13 is below 0.2 years. We conclude that nested cross-validation can potentially improve the prediction results across all experiments which utilize more than the 1 million buildings we used in our preliminary experiments. Though, the trade-off between prediction performance and training time must be considered, since increasing model complexity by raising the maximum tree depth is accompanied by an exponential increase in training time. Analyzing the bias-variance trade-off shows that the model starts to overfit above a `max_depth` of 16 with the variance term dominating the test error.

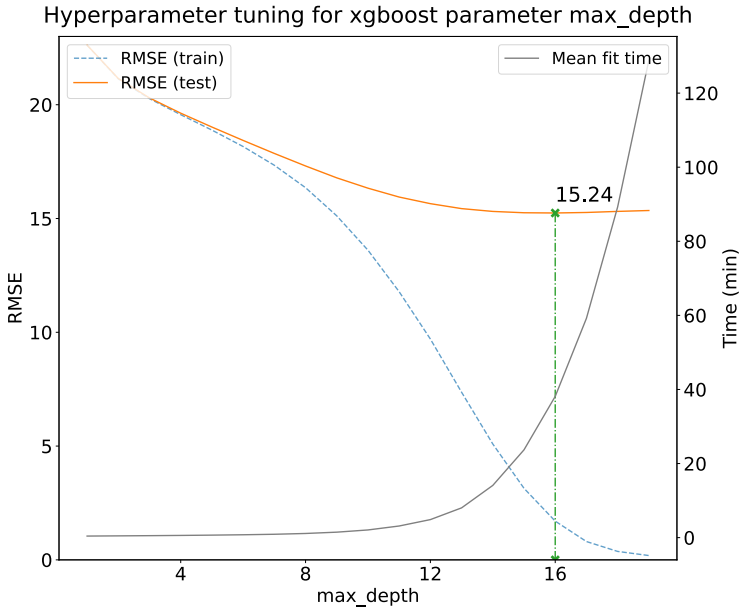


**Figure 16:** ***Impact of tree depth on prediction performance*** *Visualization of hyperparameter tuning results for the xgboost parameter* `max_depth` *using all building data from the Netherlands. A* `max_depth` *of 16 yields the lowest prediction error (RMSE=15.24). Yet, there is a trade-off between prediction performance and training time, which is determined by the complexity of the model.*

## D Spatial autocorrelation

To assess the level of spatial autocorrelation present in the data, we use Moran's I [62], a measure of spatial autocorrelation ranging from -1 to 1, where 0 represents a perfectly random spatial distribution with no spatial autocorrelation and values close to 1 or -1 respectively represent strong spatial clustering or spatial dispersion of similar samples [63]. We find that building age exhibits a high level of spatial clustering; Moran's I for immediate neighbors (k=4) is around 0.6 and for all close-by buildings (distance $<$ 100 m) between 0.3 in the Netherlands and 0.4 in France. Moran's I decays over distance as depicted in figure 17, but for many cities some residual remains even over 1 km distance. Yet, it is comparatively small with a Moran's I below 0.1 in France and Spain and below 0.05 in the Netherlands in the vast majority of cities. We argue that this sufficiently mitigates overoptimistic prediction estimates caused by spatial autocorrelation between nearby buildings. Similarly, to the phasing out of spatial autocorrelation after 1km in most cities, we find that the decrease in prediction performance levels off when increasing the distance threshold of neighborhood clustering to 1km.

Other geometric building attributes like height, perimeter or footprint area also exhibit patterns of spatial autocorrelation but to a lesser degree (Moran's I between 0.1 and 0.3 for a distance of 100 m).

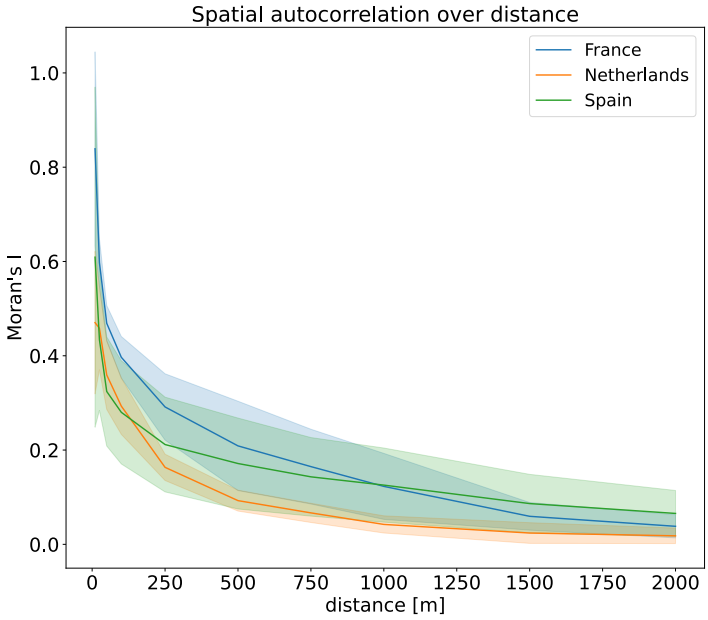


**Figure 17:** ***Spatial autocorrelation of building age decreases over distance.*** *Correlogram illustrating the decline in spatial autocorrelation of building age over distance for the Netherlands, France and Spain. The spatial autocorrelation is measured by Moran's I and is calculated city-wise. The line depicts the average Moran's I for a random selection of 25 cities per country with at least 5,000 buildings. The shaded area depicts the standard deviation around the mean. We determine the spatial autocorrelation for 10, 25, 50, 100, 250, 500, 750, 1000, 1500, and 2000 m bins. We always consider buildings located between two distances in one iteration, e.g., we consider buildings between 50 and 100 m distance and afterwards buildings between 100 and 250 m distance.*

## E Experimental setup: Impact of geographical distance on generalization performance

To assess if the geographic distance between buildings in the train and test set impacts the generalization performance, we evaluate the model using city-based cross-validation while enforcing an increasing spatial distance between the train and test set. We select a test region and divide the cities outside the region in up to 9 groups based on their distance from the test region (see figure 4). For each group we train a separate model and compare how the prediction quality changes over spatial distance. The closest group of cities is at least 50km and at most 100km away from the test region. For every group the distance increases by 50km, so that cities from the ninth group are between 450km and 500km away from the test set. Due to the size of the Netherlands the most distant group is 250km to 300km away. We consistently utilize around 100,000 buildings from randomly selected cities within each group for training. We don't split up cities, but always use all buildings from a city to allow for learning city structures.

We repeat this experiment for every province in the country which has at least 100,000 buildings by using it once as the test region. For each distance group, we report the average over all provinces to get a robust estimate of the impact of spatial distance on generalization performance that is not biased by the geography of the test region.

## F Classification approach

**Motivation** Motivated by the finding that the regression residuals are heavily biased towards the mean when prediction building age, we redefine the task as a multi-class classification problem as previously done by [28, 29]. We group the construction years into construction periods (e.g., 1950–1965) to test whether buildings at the periphery of the distribution are indeed more difficult to predict or whether this is primarily a methodological problem with the regression approach.

**Methods** Analog to our regression approach, we utilize XGBoost's implementation of a gradient boosted tree ensemble classifier and perform neighborhood-based cross-validation to control for spatial autocorrelation effects as described in section 2.2. As our primary evaluation metric, we use the Matthews Correlation Coefficient (MCC), as it is considered a good compromise among discriminancy, consistency, and coherent behaviors with varying number of classes and imbalanced datasets [64]. An MCC of $+1$ represents a perfect prediction, 0 a random prediction, and $-1$ indicates complete disagreement between predictions and true values. To compare the predictive performance of different classes and identify construction period specific challenges, we further report the class sensitivities. Since different use cases require different granularities of construction periods, we test three sets of different construction periods: uniformly distributed classes spanning 5 years, 10 years, and 20 years.

Due to the imbalanced, skewed distribution of our target variable, the construction year, discretizing it into construction periods as a preprocessing step for classification yields an imbalanced class distribution. To mitigate the negative effect of class imbalance on the predictive accuracy of minority classes and thus make class sensitivities more comparable, we pass sample weights to the classifier which are inversely proportional to the class size. We also test Synthetic Minority Oversampling (SMOTE) [65], Adaptive Synthetic Sampling (ADASYN) [66], random majority class under-, and minority class oversampling, yet sample weights yield slightly more accurate results in terms of the MCC (see Appendix table 7).

**Results** The best classification results are achieved when performing random cross-validation, especially in the Netherlands with a MCC between 0.73 and 0.76 depending on the bin size used for classification (see Appendix table 8). Using larger bins

| Strategy | MCC | F1 | Recall <1960 | Recall >1990 |
|---|---|---|---|---|
| undersampling | 0.653 | 0.614 | 0.592 | 0.663 |
| oversampling | 0.714 | 0.664 | 0.610 | 0.719 |
| SMOTE | 0.715 | 0.665 | 0.582 | 0.726 |
| ADASYN | 0.715 | 0.663 | 0.576 | 0.734 |
| baseline | 0.715 | 0.665 | 0.548 | 0.726 |
| sample weights | 0.716 | 0.667 | 0.616 | 0.722 |

**Table 7:** ***Comparison of resampling strategies for imbalanced classification.*** *Summary of classification results using Matthews Correlation Coefficient (MCC), F1 score, and the average class recall across multiple minority classes, i.e. classes containing buildings constructed before 1960 and classes containing buildings constructed after 1990. 5-fold cross-validation is performed. As resampling strategies, Synthetic Minority Oversampling (SMOTE) [65], Adaptive Synthetic Sampling (ADASYN) [66], random majority class undersampling, and minority class oversampling, and sample weights, with weights inversely proportional to the class size, are compared. The baseline strategy refers to a classifier, which does not mitigate class imbalance.*

increases the MCC at the expense of lower granularity. Analog to the regression approach, the prediction performance decreases when performing spatial cross-validation and thereby reducing the exploitation of spatial autocorrelation. In contrast to regression, the prediction quality does not deteriorate for old and new buildings (see Appendix figure 18). Specifically, the detection of new buildings benefits from the classification approach. Also, historic events such as World War I and II are more visible.

## G Determination of residential building types

We classify residential buildings into subtypes using reference attributes from TABULA [47]. We employ a simple decision tree (see Appendix figure 19) which classifies the buildings into the following types:

- Single-Family House (SFH)
- Multi-Family House (MFH)
- Terraced House (TH)
- Apartment Block (AB)

The decision splits are determined using reference values for building attributes from TABULA for the four buildings types. For example, the reference footprint area is significantly larger than $300\text{m}^2$ for almost all apartment blocks and multi-family houses in Spain and France (besides MFH from France constructed between 1975 and 1981 and before 1914) and smaller for single-family and terraced houses (besides SFH from Spain constructed between 1937 and 1959).

## H Regression vs. classification for energy modeling

**Motivation** For use cases where primarily the construction period (e.g., 1950–1965) is of interest, classification is a possible alternative to regression. Particularly for energy modeling, where the construction period is usually sufficient, but higher sensitivities for certain construction periods are more important than others, e.g., around the implementation date of energy regulations, evaluating which approach leads to lower errors in energy demand estimates when using the predicted construction period as input for energy modeling is essential. It is also important given that regression predictions are biased towards the mean resulting in unbalanced residuals for early and late construction periods (see figure 6). Yet, previous work did not compare the

| | | Netherlands | | | France | | | Spain | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 5 y. | 10 y. | 20 y. | 5 y. | 10 y. | 20 y. | 5 y. | 10 y. | 20 y. |
| *Random cv* | *class.* | 0.73 | 0.75 | 0.76 | 0.37 | 0.42 | 0.48 | 0.36 | 0.39 | 0.45 |
| | *reg.* | 0.25 | 0.38 | 0.50 | 0.08 | 0.15 | 0.25 | 0.06 | 0.13 | 0.23 |
| *Block cv* | *class.* | 0.44 | 0.52 | 0.58 | 0.22 | 0.30 | 0.39 | 0.20 | 0.26 | 0.35 |
| | *reg.* | 0.16 | 0.27 | 0.38 | 0.07 | 0.13 | 0.22 | 0.05 | 0.10 | 0.19 |
| *Neighborhood cv* | *class.* | 0.21 | 0.33 | 0.44 | 0.17 | 0.26 | 0.35 | 0.13 | 0.18 | 0.27 |
| | *reg.* | 0.09 | 0.18 | 0.28 | 0.06 | 0.12 | 0.21 | 0.04 | 0.09 | 0.16 |
| *City cv* | *class.* | 0.17 | 0.28 | 0.39 | 0.15 | 0.24 | 0.33 | 0.09 | 0.14 | 0.23 |
| | *reg.* | 0.08 | 0.14 | 0.24 | 0.05 | 0.11 | 0.20 | 0.03 | 0.08 | 0.14 |

**Table 8:** ***MCC of classification and discretized regression predictions.*** *Summary of Matthews Correlation Coefficient (MCC) for the Netherlands, France, and Spain and 5, 10, and 20 year (y.) bins. For each cross-validation (cv) strategy, the MCC is calculated for the classification (class.) and discretized regression (reg.) results.*

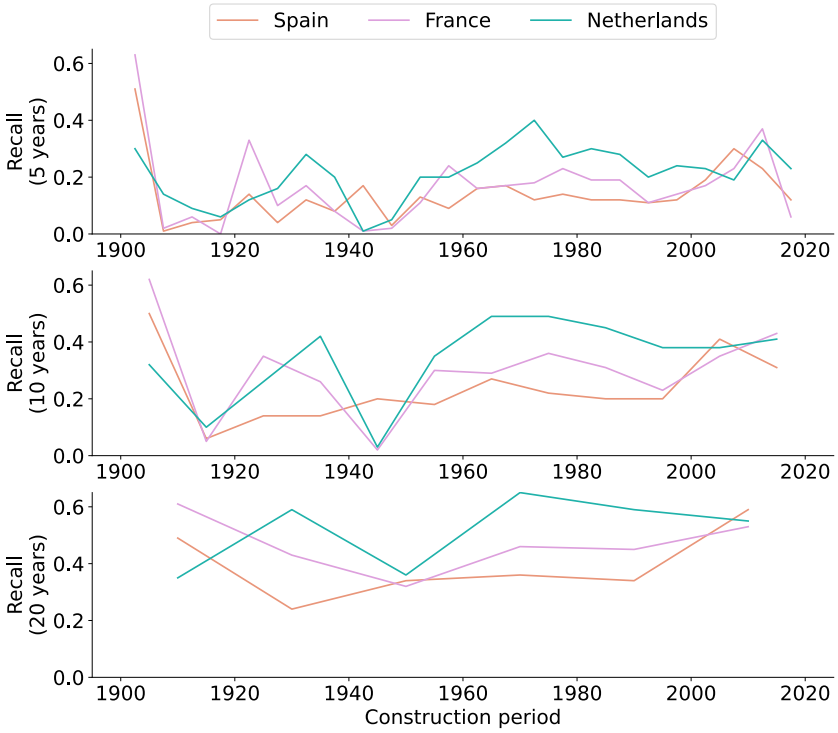


**Figure 18:** ***Classification recall is similar across construction periods.*** *Classifying the construction year using 5 year bins (top), 10 year bin (middle), or 20 year bins (bottom) results in comparable prediction errors over time, except for predictions around 1900, World War I and II. Yet, on average buildings constructed after 1960 are predicted more accurately. The sensitivity of each class is plotted for the mean class year in the graph, e.g., class 1900–1920 is plotted for 1910. 5-fold neighborhood-based cross-validation was performed. Sample weights were passed to the classifier to mitigate the class imbalance.*

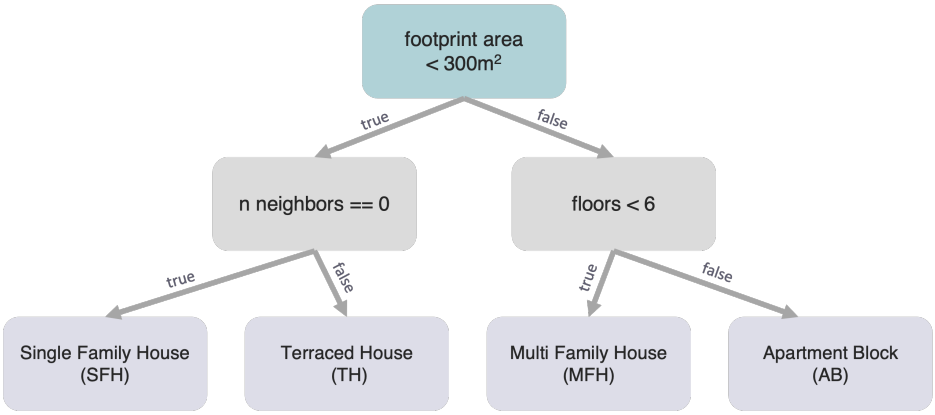


**Figure 19:** ***Decision tree for residential building types*** *The number of neighbors corresponds to the number of adjacent buildings whose geometry touches. If the number of floors was not available in the cadastral data, it was estimated by dividing the height of the building by 2.5m, the minimum height of a floor, and then rounding down.*

predictive performance of regression and multi-class classification but focused either on regressing the specific construction year [11, 30, 32] or classifying the construction period [28, 29, 31].

**Methods** The classification is conducted as described in Appendix section F. Additionally, to compare the predictive performance of regression and classification, we discretize the continuous regression predictions into construction periods and calculate classification metrics for these. To test how the error propagates when the predictions are further used, we perform an application-specific comparison for energy modeling. Instead of using a full energy model, we use precalculated energy estimates for heating for specific residential building types and construction periods from TABULA [47], equivalent to experiment 5 on the applicability for large-scale retrofitting (see section 2.3). The energy model is designed according to EN ISO 13790. We use building data from France, as we have the necessary information to classify buildings into residential types. See Appendix section G for a more detailed explanation of how residential building types are determined, which form the basis for assigning TABULA energy metrics. As classes for classification, we use the predefined construction periods from TABULA for the French building stock. We compare the derived heating demand estimate between the regression and classification model in terms of the relative difference in the MAE of energy demand estimates.

**Results** When comparing classification and regression, we find that the predictive performance of classification is more balanced across construction periods than that of regression (see Appendix figure 20). The overall MCC of classification is also noteworthy larger than that of regression, averaging between 1.7 and 3.7 times across all countries, depending on bin size (see Appendix table 8). For energy modeling based on construction periods, we find that classification slightly outperforms regression. Precisely, classification leads to a 7% smaller mean absolute error in the heating demand estimates when using TABULA construction periods and TABULA building energy information. The fact that the difference between classification and regression is larger according to the MCC than according to the MAE of energy estimates, can be attributed to the fact that the order of classes matters for energy modeling. For example, an energy estimate is likely to be less incorrect when assigning a building constructed in 1985 to the 1990s compared to assigning it to the 1920s. Therefore, when predicting buildings age, it is essential to always compare classification and regression in terms of the final use case metrics, rather than general classification metrics.

In summary, our experimental results indicate that classification should be considered as an alternative to regression when only the construction period and not the specific year is needed. For energy modeling, our findings suggest that it is likely to yield more accurate results. For use cases where the prediction accuracy of buildings at the periphery of the distribution is important, classification outperforms regression.

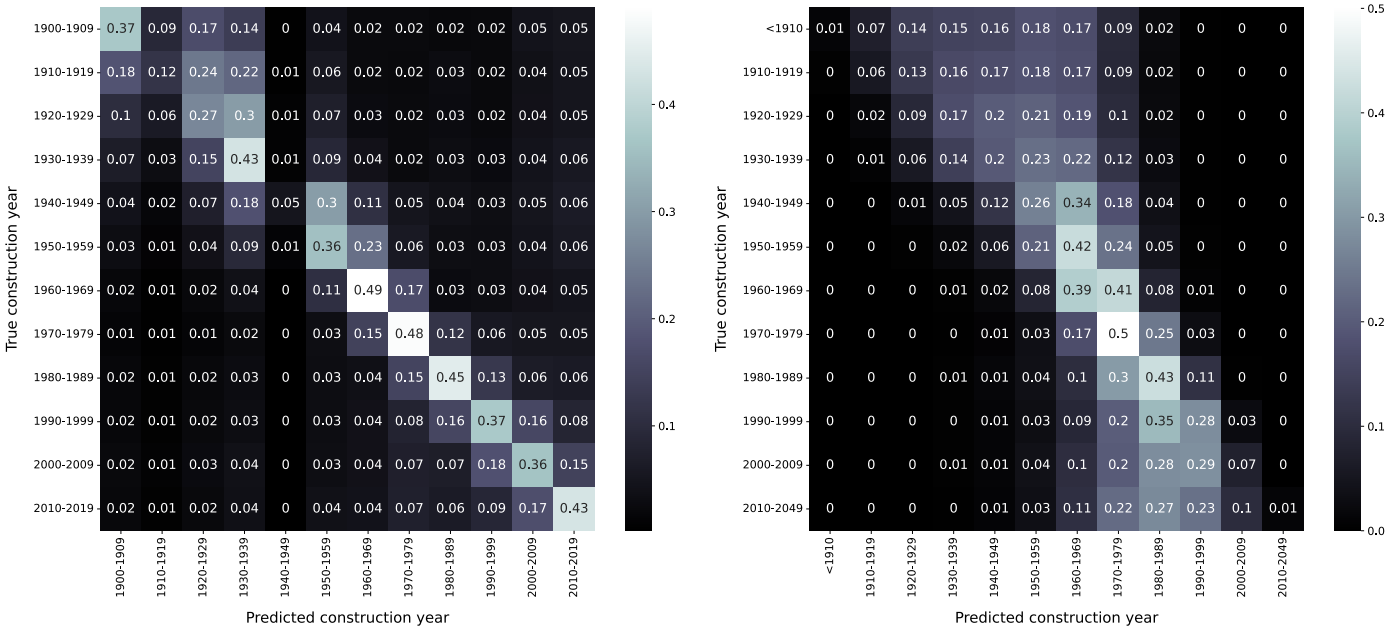


**Figure 20:** ***Classification yields more balanced prediction errors over the construction periods*** *Confusion matrix of model sensitivity for classification (left) and discretized regression (right) results for equally-sized 10 year bins in the Netherlands. Performing classification results in balanced prediction residuals, whereas the regression residuals are biased towards the mean. Thus, classification particularly outperforms regression for buildings constructed before 1940 and after 2000. Overall, the MCC of the classification results (*0.33*) is larger than the MCC of the discretized regression results (*0.18*). 5-fold neighborhood-based cross-validation was performed. Sample weights were passed to the classifier to mitigate the class imbalance.*

# I Prediction of other building attributes

To assess if urban form characteristics are a robust source of information for predicting buildings attributes in general, we repeat the experiments on local inference (Exp. 1) and regional generalization (Exp. 2) of building construction year, as described in section 2.3), for building height and type.

**Building type prediction** We chose Matthews Correlation Coefficient (MCC) as our primary evaluation metric instead of the overall accuracy, which can be misleading for multi-class classification, because it poorly copes with imbalanced classes and cannot distinguish among different misclassification distributions. The MCC is considered a good compromise among discriminancy, consistency, and coherent behaviors with varying number of classes, imbalanced datasets, and randomization [64]. An MCC of $+1$ represents a perfect prediction, 0 a random prediction, and $-1$ indicates complete disagreement between predictions and true values. To assess the prediction performance of different classes, identify construction period specific challenges, and

| | France | | | Spain | | |
|---|---|---|---|---|---|---|
| Cross-validation | MCC | F1 | Recall residential | MCC | F1 | Recall residential |
| Random cv | 0.605 | 0.800 | 0.892 | 0.623 | 0.808 | 0.883 |
| Block cv | 0.595 | 0.794 | 0.887 | 0.581 | 0.787 | 0.873 |
| Neighborhood cv | 0.593 | 0.793 | 0.888 | 0.564 | 0.780 | 0.875 |
| City cv | 0.579 | 0.785 | 0.880 | 0.507 | 0.751 | 0.861 |

**Table 9:** ***Binary building type classification error.*** *Summary of Matthews Correlation Coefficient (MCC), the F1 score, and majority class recall for the classification of building type into residential and non-residential houses. The results are reported for different cross-validation (cv) strategies in the Netherlands, France, and Spain.*

| Country | Cross-validation | MCC | F1 | Recall | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | residential | commercial | agricultural | industrial | others |
| France | Random cv | 0.564 | 0.572 | 0.851 | 0.598 | 0.242 | 0.612 | 0.779 |
| | Block cv | 0.557 | 0.524 | 0.852 | 0.555 | 0.147 | 0.473 | 0.777 |
| | Neighborhood cv | 0.555 | 0.516 | 0.852 | 0.538 | 0.146 | 0.441 | 0.776 |
| | City cv | 0.553 | 0.504 | 0.853 | 0.531 | 0.113 | 0.409 | 0.773 |
| Spain | Random cv | 0.569 | 0.650 | 0.807 | 0.583 | 0.923 | 0.705 | – |
| | Block cv | 0.519 | 0.600 | 0.807 | 0.429 | 0.869 | 0.631 | – |
| | Neighborhood cv | 0.506 | 0.591 | 0.815 | 0.400 | 0.858 | 0.583 | – |
| | City cv | 0.448 | 0.549 | 0.815 | 0.339 | 0.717 | 0.543 | – |

**Table 10:** ***Fine-granular building type classification error.*** *Summary of Matthews Correlation Coefficient (MCC), the F1 score, and class recalls for the multiclass classification of building type. The results are reported for different cross-validation (cv) strategies in the Netherlands, France, and Spain. Buildings are classified into residential, commercial, agricultural, industrial, and others (France only). For the class distribution please refer to figure 12.*

compare our results to previous studies, we further calculate the model's recall and F1 score.

First large-scale experiments demonstrate that urban form is also predictive for the building type in several countries. Buildings can be classified into residential and non-residential buildings with an F1 score between 0.78 and 0.79 and an MCC between 0.56 and 0.59 (see Appendix table 9) and into more fine-granular types, namely residential, commercial, agricultural, industrial and others, with an MCC of 0.51 and 0.56 (see Appendix table 10) when performing neighborhood-based cross-validation. Compared to the prediction of the construction year, we find that the difference between random and neighborhood-based cross-validation is relatively small (in terms of MCC, below 0.01 in France and 0.06 in Spain). We conclude that the model doesn't exploit the spatial autocorrelation as much, resulting in type predictions that are less overoptimistic when relying only on local data and not performing spatial cross-validation. The high MCC for cross-city predictions also shows that the building type can estimated comparatively well in unseen regions.

**Building height prediction** Our experiments show that urban form characteristics are also highly informative for building height across the Netherlands, France and Spain. Building height can be predicted with an MAE between 0.71 and 1.36 meters and an $R^2$ of up to 0.77 when performing neighborhood-based cross-validation and (see Appendix table 11). Access to local information improves the height prediction in all three countries with a similar magnitude as for the prediction of the construction year. In the Netherlands, the MAE can even be reduced to 0.56 meters and the $R^2$ increases to 0.84. Overall, the predictive performance in terms of $R^2$ is higher than for the prediction of the construction year in all three countries. This makes it a promising tool to infer missing buildings heights in available administrative data and

| | Netherlands | | | France | | | Spain | | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cross-validation | MAE | RMSE | R2 | MAE | RMSE | R2 | MAE | RMSE | R2 | MAE | RMSE | R2 |
| Random cv | 0.56 | 1.10 | 0.84 | 1.17 | 1.66 | 0.56 | 1.20 | 2.02 | 0.71 | 1.08 | 1.73 | 0.67 |
| Block cv | 0.66 | 1.25 | 0.79 | 1.23 | 1.75 | 0.51 | 1.28 | 2.17 | 0.66 | 1.13 | 1.82 | 0.64 |
| Neighborhood cv | 0.71 | 1.30 | 0.77 | 1.26 | 1.80 | 0.48 | 1.36 | 2.30 | 0.62 | 1.17 | 1.90 | 0.61 |
| City cv | 0.73 | 1.34 | 0.76 | 1.31 | 1.86 | 0.45 | 1.43 | 2.45 | 0.57 | 1.23 | 1.99 | 0.57 |

**Table 11:** ***Building height prediction error.*** *Summary of the mean absolute error (MAE), the root mean squared error (RMSE) in meters and the coefficient of determination ($R^2$) of the different cross-validation strategies. For random, urban block-based, neighborhood-based and city-based cross-validation (cv), training and prediction is conducted in the same country or in all countries at once. For country-based cross-validation the model was trained on two countries to predict in the third country. The table shows the prediction result for the country used as test set and for all countries, the average over all validation folds.*

more broadly, help to upscale geographically differentiated solutions for low-carbon urban planning.

## J Additional figures


**Figure 21:** ***Exemplary illustration of in-fill housing for Vichy, France.*** *Exemplary illustration of in-fill housing from bird's eye view for Vichy, France. In-fill housing refers to new buildings constructed on underused lots in existing, older urban neighborhoods. Specifically, we define in-fill housing as adjacent buildings whose construction year is more than two standard deviations above the mean of all buildings in the same street-based block. All in-fill buildings are highlighted in orange on the map. The mean absolute error (MAE) is substantially larger for in-fill housing in France (43.69 years) than for other types of housing (17.93 years).*

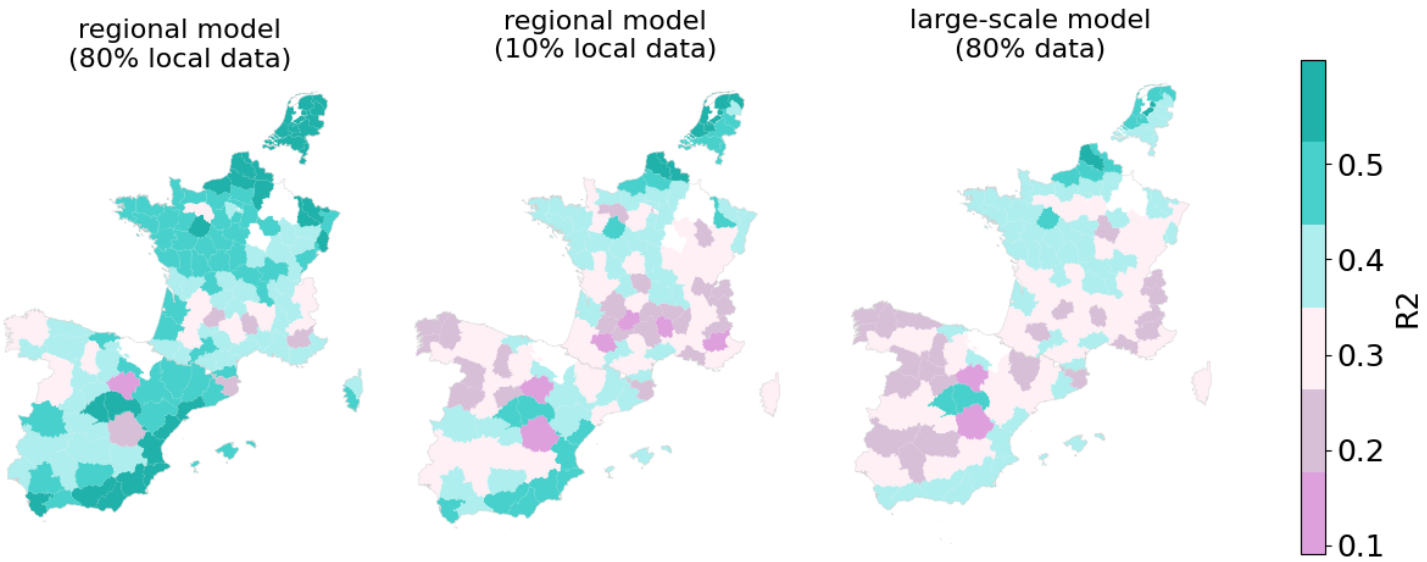


**Figure 22:** ***Regional models outperform national model in local inference.*** *Illustration of the coefficient of determination, $R^2$, of specialized regional models with either 80% or 10% local data availability across provinces in the Netherlands, France, and Spain. For comparison, the prediction quality across regions of the national model evaluated by 5-fold random cross-validation, consequently using 80% of data for training, is depicted on the right. The median number of buildings per region is about 125,000. See Appendix table 14 for the precise $R^2$ and MAE values.*

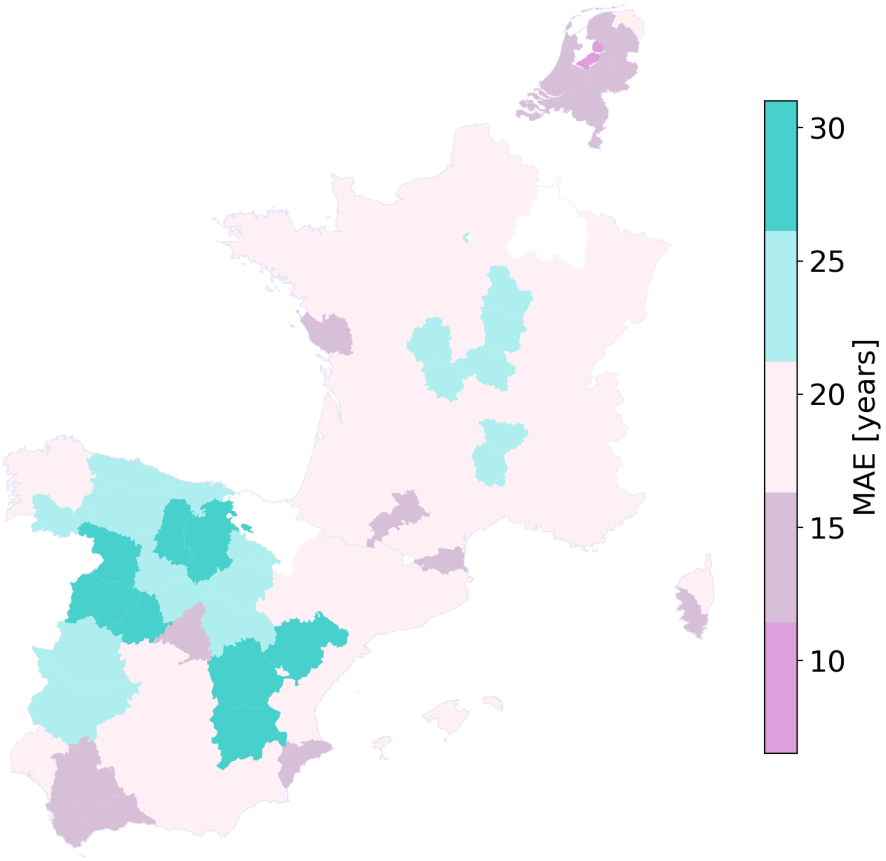


**Figure 23:** ***Prediction error varies across regions.*** *The mean absolute error (MAE) varies across countries and regions. At the cross-country level, buildings in the Netherlands are predicted more accurately than in France and Spain. In France, departments within the center show comparatively high prediction errors. In Spain, the error varies by a factor of 2 between the provinces. Particularly, provinces in the center and north, with the exception of the Madrid provincia, show high prediction errors.*

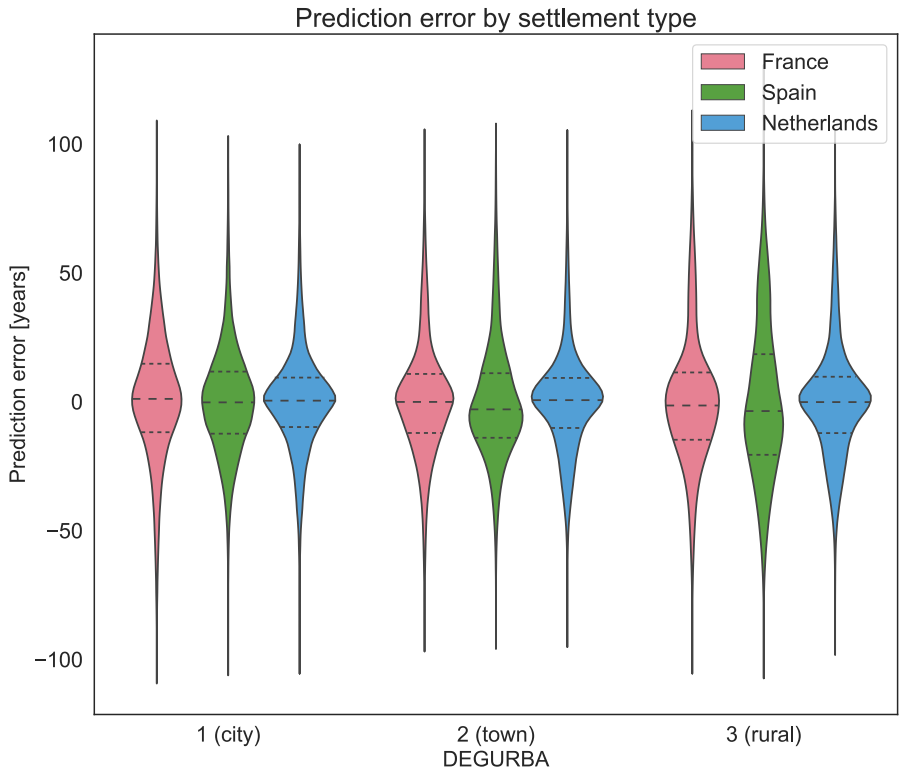


**Figure 24:** ***Prediction error differs between settlement types.*** *Visualization of the error distribution for cities, town, and rural areas according to the EU's DEGURBA classification for the Netherlands, France, and Spain.*


**Figure 25:** ***Value distributions of most predictive features differ between France, Netherlands, and Spain.*** *The feature value distributions of the 8 most predictive features (besides latitude and longitude) differ between France, Netherlands, and Spain. For instance, more buildings with a small footprint area exist in France compared to the Netherlands and Spain. On the other side, the Netherlands are more densely built-up, with the number of buildings and total footprint area in a spatial buffer around each building being larger than in France and Spain.*

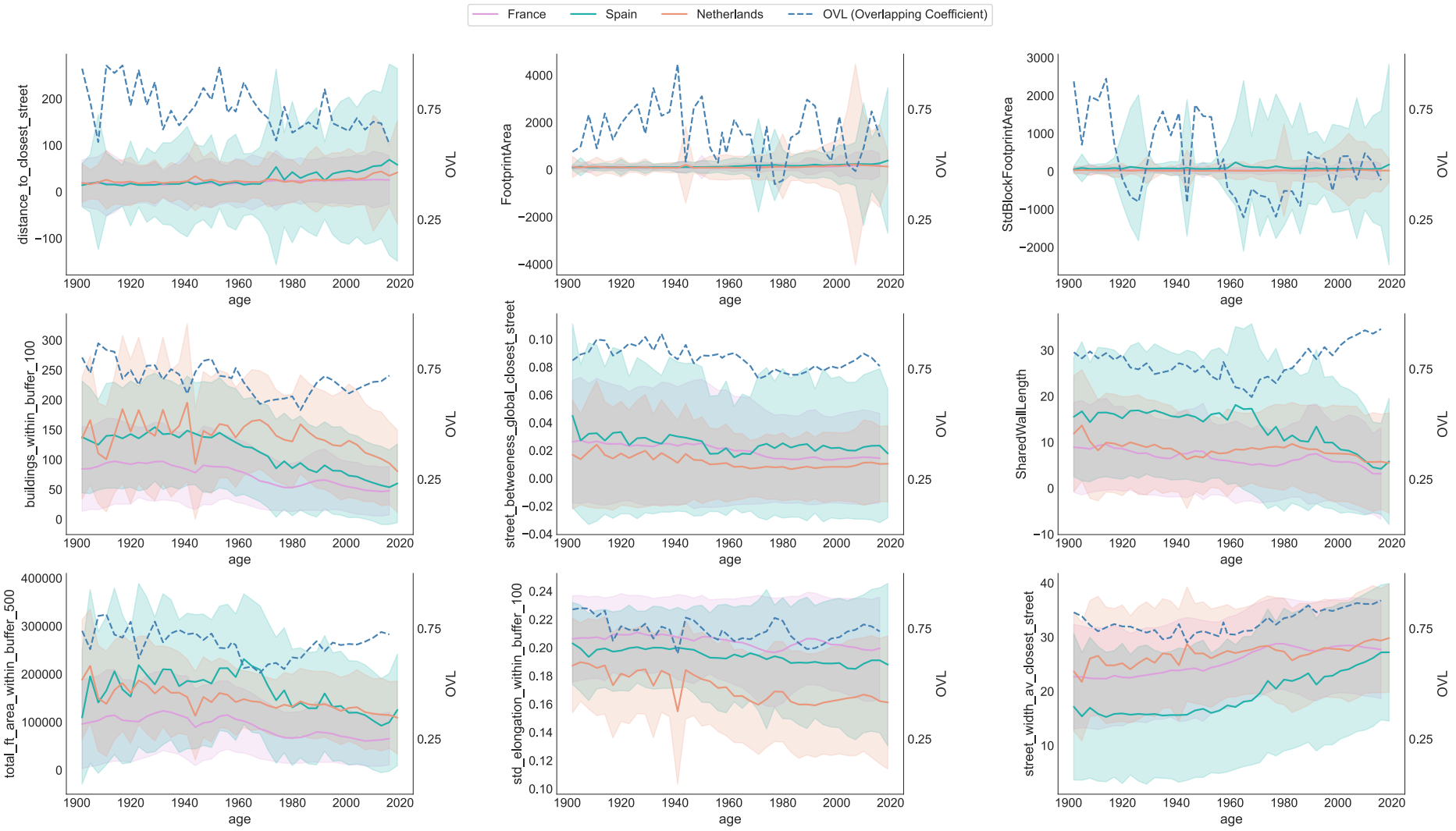


**Figure 26:** ***Feature value distributions of most predictive urban form characteristics over time.*** *The distributions of the 9 most important urban form characteristics differ between the Netherlands, France and Spain over the construction periods. On average, they overlap by 71.4%. The dashed line indicates the overlapping coefficient (OVL) for each feature over time. The shaded area marks the standard deviation around the mean estimate. A moving average of 3 years is applied to smooth the development. The ordering is according to their feature importance, ranging from distance_to_closest_street with 4.4% to street_width_av_closest_street with 1.9% on average across all three countries. Overall, 28.4% of the prediction contribution can be attributed those 9 features according to their SHAP-values.*

## K Additional tables

| Country | Type | Share | Settlement size | construction year mean | construction year std |
|---|---|---|---|---|---|
| Netherlands | 1 (city) | 43.1 | 23800 | 1970.5 | 27.9 |
| | 2 (town) | 41.6 | 9463 | 1974.4 | 24.6 |
| | 3 (rural) | 15.4 | 9462 | 1973.2 | 27.2 |
| France | 1 (city) | 8.0 | 5433 | 1959.1 | 32.1 |
| | 2 (town) | 17.8 | 4760 | 1971.5 | 29.6 |
| | 3 (rural) | 74.3 | 3019 | 1973.7 | 31.9 |
| | non-residential | 17.9 | – | 1966.5 | 34.3 |
| | residential | 82.1 | – | 1973.1 | 31.0 |
| Spain | 1 (city) | 23.4 | 14117 | 1978.6 | 27.0 |
| | 2 (town) | 39.3 | 3819 | 1981.4 | 28.8 |
| | 3 (rural) | 37.3 | 655 | 1968.2 | 37.5 |
| | non-residential | 21.6 | – | 1979.0 | 32.4 |
| | residential | 78.4 | – | 1974.6 | 32.3 |

**Table 12:** ***Settlement and building type statistics of preprocessed dataset.*** *The table indicates the share of buildings located in cities, town, and rural areas according to the EU's DEGURBA classification as well as the share of residential and non-residential buildings. Building type information is not available in the Netherlands. Settlement size refers to the average number of buildings per settlement type. Mean and standard deviation (std) of the construction year is provided for each settlement and building type respectively.*

| | | France | | | Netherlands | | | Spain | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bin size | Cross-validation | <1960 | 1960-1990 | >1990 | <1960 | 1960-1990 | >1990 | <1960 | 1960-1990 | >1990 |
| 5 years | *Random cv* | 33.80 | 40.00 | 37.20 | 67.70 | 79.30 | 69.80 | 43.80 | 36.70 | 37.70 |
| | Block cv | 19.30 | 24.70 | 22.70 | 34.70 | 52.80 | 43.20 | 26.30 | 20.70 | 21.00 |
| | *Neighborhood cv* | 14.80 | 18.70 | 18.00 | 15.10 | 30.30 | 23.70 | 11.80 | 13.80 | 17.80 |
| | City cv | 13.10 | 17.70 | 17.00 | 9.70 | 28.00 | 19.00 | 7.70 | 10.80 | 15.50 |
| 10 years | *Random cv* | 43.50 | 47.70 | 47.70 | 71.70 | 81.30 | 73.30 | 52.50 | 42.30 | 44.00 |
| | *Block cv* | 31.20 | 36.30 | 37.00 | 43.80 | 63.30 | 53.30 | 36.70 | 29.30 | 33.00 |
| | *Neighborhood cv* | 26.70 | 32.00 | 33.70 | 24.70 | 47.70 | 39.00 | 20.30 | 23.00 | 30.70 |
| | *City cv* | 24.20 | 31.00 | 32.30 | 18.50 | 44.70 | 35.30 | 15.20 | 19.70 | 28.70 |
| 20 years | *Random cv* | 59.70 | 55.50 | 62.00 | 78.70 | 82.50 | 79.00 | 63.30 | 49.00 | 62.00 |
| | *Block cv* | 50.00 | 48.50 | 56.00 | 57.70 | 70.50 | 66.00 | 51.70 | 39.50 | 58.00 |
| | *Neighborhood cv* | 45.30 | 45.50 | 53.00 | 43.30 | 62.00 | 55.00 | 35.70 | 35.00 | 59.00 |
| | *City cv* | 42.30 | 45.00 | 52.00 | 35.00 | 58.00 | 54.00 | 30.30 | 31.50 | 59.00 |

**Table 13:** ***Mean classification recall before 1960, after 1990, and in between.*** *Summary of the recall for the classification of construction periods in the Netherlands, France, and Spain. The recall is averaged across classes whose mean construction year is before 1960, in between 1960 and 1990, and after 1990. The averaged recall is reported for different cross-validation (cv) strategies and bin sizes.*

| | Spain | | France | | Netherlands | |
|---|---|---|---|---|---|---|
| Local data availability | MAE | R2 | MAE | R2 | MAE | R2 |
| 10% | 20.0 | 0.33 | 18.4 | 0.33 | 11.0 | 0.56 |
| 20% | 19.4 | 0.34 | 17.6 | 0.36 | 10.4 | 0.60 |
| 50% | 18.5 | 0.40 | 16.7 | 0.41 | 9.7 | 0.63 |
| 80% | 18.1 | 0.42 | 16.2 | 0.43 | 9.5 | 0.65 |
| national model | 19.1 | 0.39 | 17.6 | 0.39 | 10.2 | 0.62 |

**Table 14:** ***Impact of local data availability for region models.*** *Comparison of the prediction performance of regional models that are trained using 10%, 20%, 50%, or 80% of the local data from the region and predict all remaining buildings. For 12 states in the Netherlands, 47 provinces in Spain, and 91 departments in France the averaged results are reported. For comparison, the prediction quality of the national model evaluated by random cross-validation is depicted as well.*

| | | Neighborhood cv | | | Random cv | | | Across cities | | | Across countries | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Country | n cities | R2 | MAE | RMSE | R2 | MAE | RMSE | R2 | MAE | RMSE | R2 | MAE | RMSE |
| France | 1 | -0.07 | 25.49 | 31.97 | 0.30 | 18.53 | 26.19 | -0.86 | 33.74 | 40.02 | -3.28 | 46.92 | 53.17 |
| | 2 | 0.14 | 23.13 | 30.18 | 0.41 | 17.37 | 25.12 | -0.79 | 33.34 | 39.90 | -3.32 | 47.65 | 54.16 |
| | 4 | 0.20 | 21.29 | 27.86 | 0.43 | 15.97 | 23.65 | -0.14 | 25.66 | 31.88 | -2.17 | 39.80 | 46.21 |
| | 8 | 0.25 | 20.02 | 26.44 | 0.43 | 15.65 | 23.09 | 0.02 | 23.82 | 29.68 | -1.37 | 33.97 | 40.25 |
| | 16 | 0.30 | 19.48 | 25.84 | 0.46 | 15.34 | 22.68 | 0.17 | 21.78 | 27.49 | -1.08 | 31.78 | 37.75 |
| | 32 | 0.34 | 19.13 | 25.50 | 0.48 | 15.30 | 22.63 | 0.25 | 20.63 | 26.21 | -0.55 | 26.90 | 32.54 |
| | 64 | 0.35 | 18.72 | 25.16 | 0.47 | 15.39 | 22.61 | 0.28 | 19.98 | 25.65 | -0.41 | 25.57 | 31.13 |
| | 128 | 0.37 | 18.61 | 25.03 | 0.49 | 15.51 | 22.62 | 0.30 | 19.58 | 25.23 | -0.29 | 24.24 | 29.80 |
| | 256 | 0.38 | 18.48 | 24.91 | 0.48 | 15.76 | 22.72 | 0.33 | 19.03 | 24.72 | -0.31 | 24.47 | 30.04 |
| | 512 | 0.39 | 18.36 | 24.79 | 0.49 | 16.19 | 22.99 | 0.32 | 19.28 | 24.88 | -0.28 | 24.28 | 29.76 |
| | 1024 | 0.40 | 18.23 | 24.67 | 0.48 | 16.43 | 23.12 | 0.36 | 18.29 | 24.18 | -0.28 | 24.14 | 29.80 |
| | 2048 | 0.40 | 18.06 | 24.49 | 0.46 | 16.69 | 23.27 | 0.38 | 17.87 | 23.79 | -0.31 | 24.48 | 30.14 |
| | 3483 | 0.41 | 16.83 | 23.02 | 0.49 | 14.99 | 21.49 | 0.38 | 17.99 | 23.85 | 0.39 | 15.26 | 20.58 |
| Netherlands | 1 | -0.15 | 20.30 | 25.71 | 0.56 | 8.60 | 15.83 | -0.44 | 26.76 | 32.38 | -0.12 | 27.01 | 32.60 |
| | 2 | 0.08 | 19.01 | 24.21 | 0.61 | 8.72 | 15.84 | -0.03 | 22.50 | 27.80 | -0.00 | 25.52 | 30.91 |
| | 4 | 0.20 | 16.47 | 21.80 | 0.64 | 7.56 | 14.58 | 0.14 | 20.42 | 25.49 | 0.06 | 24.58 | 29.96 |
| | 8 | 0.27 | 16.19 | 21.48 | 0.66 | 7.73 | 14.63 | 0.18 | 19.83 | 24.78 | 0.07 | 24.48 | 29.86 |
| | 16 | 0.31 | 15.59 | 20.94 | 0.66 | 7.80 | 14.57 | 0.23 | 19.19 | 24.13 | 0.12 | 23.60 | 29.02 |
| | 32 | 0.38 | 15.04 | 20.43 | 0.68 | 7.92 | 14.49 | 0.27 | 18.54 | 23.44 | 0.13 | 23.34 | 28.76 |
| | 64 | 0.40 | 14.55 | 19.96 | 0.67 | 8.28 | 14.68 | 0.32 | 17.74 | 22.70 | 0.17 | 22.76 | 28.18 |
| | 128 | 0.41 | 14.58 | 20.09 | 0.66 | 9.10 | 15.34 | 0.34 | 17.21 | 22.23 | 0.18 | 22.58 | 27.97 |
| | 256 | 0.43 | 14.34 | 19.84 | 0.64 | 9.60 | 15.69 | 0.37 | 16.65 | 21.80 | 0.19 | 22.37 | 27.74 |
| | 424 | 0.44 | 14.31 | 19.87 | 0.63 | 10.09 | 16.09 | 0.39 | 16.28 | 21.46 | 0.20 | 22.25 | 27.66 |
| | 512 | 0.43 | 14.49 | 20.10 | 0.62 | 10.44 | 16.52 | 0.39 | 16.23 | 21.44 | 0.26 | 20.84 | 26.52 |
| | 1024 | 0.42 | 15.34 | 21.21 | 0.56 | 12.09 | 18.43 | 0.39 | 16.27 | 21.43 | 0.33 | 19.63 | 25.34 |

**Table 15:** ***Impact of additional data on prediction performance.*** *Results of the experiment evaluating the impact of additional data on the prediction performance for France and the Netherlands for different use cases. See section 2.3 for more details. The dashed horizontal line indicates when data from other countries were used for model training, because all cities of the respective country were already used.*

| Country | DEGURBA | non-residential | residential |
|---|---|---|---|
| France | 1 (city) | 21.16 | 17.44 |
| | 2 (town) | 20.63 | 15.54 |
| | 3 (rural) | 22.84 | 17.48 |
| | All | 21.54 | 16.82 |
| Spain | 1 (city) | 16.18 | 15.57 |
| | 2 (town) | 15.57 | 17.27 |
| | 3 (rural) | 22.34 | 24.37 |
| | All | 18.03 | 19.07 |

**Table 16:** ***Prediction error (MAE) by building and settlement type in France and Spain.*** *Summary of the mean absolute error (MAE) in meters of residential and non-residential buildings for different settlement types in France and Spain. Settlement types are determined according to the EU's DEGURBA classification. 5-fold neighborhood-based cross-validation was performed.*

| Construction period | France | Netherlands | Spain | All |
|---|---|---|---|---|
| 1900-1944 | 33.82 | 27.84 | 37.99 | 33.22 |
| 1945-1969 | 11.52 | 9.76 | 13.45 | 11.58 |
| 1970-1979 | 9.66 | 6.56 | 11.91 | 9.38 |
| 1980-1989 | 9.91 | 8.28 | 12.14 | 10.11 |
| 1990-1999 | 15.29 | 12.57 | 14.13 | 14.00 |
| 2000-2009 | 20.77 | 21.07 | 18.13 | 19.99 |
| >=2010 | 27.96 | 29.40 | 25.43 | 27.60 |
| All | 18.42 | 16.50 | 19.03 | 17.98 |

**Table 17:** ***Prediction error (MAE) across construction periods.*** *Summary of the mean absolute error (MAE) across construction periods in the Netherlands, France, Spain. All refers to the column and row average, respectively. 5-fold neighborhood-based cross-validation was performed.*

| | France | | Spain | |
|---|---|---|---|---|
| Construction period | non-residential | residential | non-residential | residential |
| 1900-1944 | 34.27 | 34.07 | 41.53 | 36.74 |
| 1945-1959 | 11.28 | 12.15 | 20.77 | 14.28 |
| 1960-1969 | 11.14 | 10.99 | 15.99 | 11.14 |
| 1970-1979 | 10.73 | 9.52 | 12.81 | 11.83 |
| 1980-1989 | 13.51 | 9.35 | 10.73 | 12.81 |
| 1990-1999 | 22.51 | 13.77 | 10.94 | 15.35 |
| 2000-2009 | 27.92 | 19.48 | 15.27 | 19.08 |
| >=2010 | 35.82 | 26.45 | 22.60 | 27.29 |
| All | 22.48 | 17.20 | 18.50 | 19.40 |

**Table 18:** ***Prediction error (MAE) by building type across construction periods in France and Spain.*** *Summary of the mean absolute error (MAE) of residential and non-residential buildings across construction periods in France and Spain. All refers to the column-wise average. 5-fold neighborhood-based cross-validation was performed.*

| Metatype | Type | France MAE | France RMSE | France R2 | Spain MAE | Spain RMSE | Spain R2 |
|---|---|---|---|---|---|---|---|
| residential | TH | 18.09 | 24.67 | 0.43 | 21.28 | 28.61 | 0.28 |
| | SFH | 16.05 | 21.98 | 0.38 | 16.55 | 22.49 | 0.26 |
| | MFH | 20.72 | 27.47 | 0.31 | 16.00 | 22.29 | 0.30 |
| | AB | 15.62 | 20.55 | 0.04 | 10.93 | 14.72 | 0.40 |
| non-residential | others | 21.87 | 28.36 | 0.28 | – | – | – |
| | industrial | 15.76 | 22.35 | 0.14 | 21.09 | 27.47 | 0.39 |
| | commercial | 26.66 | 33.09 | 0.12 | 19.88 | 26.13 | 0.25 |
| | agricultural | 27.75 | 34.21 | 0.15 | 14.75 | 19.56 | 0.18 |

**Table 19:** ***Prediction error by fine-granular building type in France and Spain.*** *Summary of the mean absolute error (MAE), the root mean squared error (RMSE) in meters and the coefficient of determination ($R^2$) of different building types. 5-fold neighborhood-based cross-validation was performed. Building type data is available for Spain and France. Of all buildings 78.7% are residential in Spain and 82.2% in France. The subtypes of residential buildings, single-family houses (SFH), terraced houses (TH), multi-family houses (MFH), and apartment blocks (AB), are estimated using a simple decision tree based on available building attributes such as height, footprint area and the number of adjacent buildings (see Appendix section G). The remaining non-residential building, are labeled as agricultural, industrial or commercial and service buildings based on harmonized cadaster type information.*

| Feature | Netherlands | France | Spain | All |
|---|---|---|---|---|
| distance_to_closest_street | 0.032 | 0.078 | 0.022 | 0.044 |
| lat | 0.022 | 0.043 | 0.062 | 0.043 |
| FootprintArea | 0.011 | 0.057 | 0.050 | 0.039 |
| StdBlockFootprintArea | 0.026 | 0.013 | 0.077 | 0.039 |
| buildings_within_buffer_100 | 0.025 | 0.037 | 0.042 | 0.035 |
| street_betweeness_global_closest_street | 0.031 | 0.050 | 0.018 | 0.033 |
| SharedWallLength | 0.017 | 0.011 | 0.059 | 0.029 |
| total_ft_area_within_buffer_500 | 0.028 | 0.016 | 0.028 | 0.024 |
| lon | 0.010 | 0.017 | 0.036 | 0.021 |
| std_elongation_within_buffer_100 | 0.041 | 0.004 | 0.014 | 0.019 |
| street_width_av_closest_street | 0.007 | 0.022 | 0.027 | 0.019 |
| av_footprint_area_within_buffer_100 | 0.015 | 0.031 | 0.007 | 0.018 |
| Phi | 0.017 | 0.032 | 0.005 | 0.018 |
| Convexity | 0.020 | 0.007 | 0.022 | 0.017 |
| total_ft_area_within_buffer_100 | 0.017 | 0.010 | 0.019 | 0.016 |
| std_convexity_within_buffer_100 | 0.014 | 0.013 | 0.016 | 0.014 |
| av_convexity_within_buffer_100 | 0.011 | 0.021 | 0.011 | 0.014 |
| BlockConvexity | 0.012 | 0.021 | 0.008 | 0.014 |
| BlockTotalFootprintArea | 0.027 | 0.006 | 0.006 | 0.013 |
| AvBlockFootprintArea | 0.010 | 0.011 | 0.014 | 0.012 |
| Total | 0.4020 | 0.5079 | 0.5503 | 0.4867 |

**Table 20:** ***Feature importance based on SHAP-values.*** *Feature importance comparison between the Netherlands, France, and Spain for the 20 most decisive features when performing neighborhood-based cross-validation. Feature importance refers to the normalized, individual contribution of each feature to the prediction according to their SHAP-values. The average feature importance across all countries is depicted in the All column.*

| | Netherlands | | | France | | | Spain | | |
|---|---|---|---|---|---|---|---|---|---|
| Add. fts | MAE | RMSE | R2 | MAE | RMSE | R2 | MAE | RMSE | R2 |
| both | 13.90 | 19.42 | 0.46 | 17.59 | 23.98 | 0.43 | 18.67 | 25.66 | 0.37 |
| height | 13.90 | 19.43 | 0.46 | 17.75 | 24.19 | 0.42 | 18.84 | 25.85 | 0.37 |
| type | 14.30 | 19.86 | 0.44 | 17.82 | 24.22 | 0.42 | 19.09 | 26.06 | 0.36 |
| baseline | 14.31 | 19.87 | 0.44 | 18.00 | 24.44 | 0.41 | 19.29 | 26.29 | 0.34 |

**Table 21:** ***Prediction error for additional features.*** *Impact on the prediction performance when using building height and or type as additional features (add. fts) for model training. We perform neighborhood-based 5-fold cross validation for all experiments. Baseline refers to a model that does not use either feature as defined in the methods section 2.2 and evaluated in section 3.4.*