AGILE: GIScience Series, 7, 44, 2026. https://doi.org/10.5194/agile-giss-7-44-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Uncertainty in Binned Building Construction Year Data: Comparing EPC and Crowdsourced Datasets

Sophie Teichmann[1,2], Polly Hudson[3], Mihyun Kim[4], Hendrik Herold[1,2], and Robert Hecht[1,2]

[1]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, TU Dresden, Germany
[2]Research Group Advanced Environmental Risk and Sustainability Modelling of Cities and Regions Using AI (SITES.AI), Leibniz Institute of Ecological Urban and Regional Development, Germany
[3]University of Cambridge (Visiting academic), UK
[4]Loughborough University, UK

Correspondence: Sophie Teichmann (sophie.teichmann@tu-dresden.de)

Abstract. The building's construction year indicates energy performance and retrofit potential, and is used in energy performance ratings. However, in many countries, comprehensive datasets providing the construction date of buildings are unavailable or difficult to access. In the UK, Energy Performance Certificates (EPC) provide the only open dataset on building construction years, at national scale. However, the data are binned, and minimal information is available on data quality. For specific areas of England, construction year data have also been contributed by historians to the Colouring Britain data platform. This study investigates systematic differences between EPC and historian-crowdsourced construction year data for 4,849 buildings in Loughborough, covering 14 EPC bands. To increase comparability of datasets with differing age bands, we test a random forest method to resolve the binning using two feature sets: urban form features, and urban form features plus EPC bands. The median of each EPC band acts as the baseline. The results show that combining urban form features and EPC bands delivers the highest accuracy.

Submission Type. analysis, case study

BoK Concepts. [IP3-4-7-1] Random forest (RF), [GD6] Data Quality, Metadata and Data Infrastructure

Keywords. data quality, urban form, construction years

1 Introduction

Buildings are responsible for around 30% of global final energy consumption and 26% of global energy-related emissions according to the (International Energy Agency, 2022). Over the past decade, efforts to capture and generate open microspatial building attribute data at scale have increased across non-profit, academic, government and commercial sectors (Oostwegel et al., 2025; Touzani and Granderson, 2021; Milojevic-Dupont et al., 2023; OpenStreetMap contributors, 2017; Overture Maps Foundation). Included within these sought after building attribute datasets are construction year datasets. Construction years are now considered essential, in many areas of urban analysis including retrofit planning, material flow analysis (Aksoezen et al., 2015), and energy performance ratings (UK Office for National Statistics, 2021),

A growing number of official datasets now include building construction years, with varying scale, coverage, and quality (Milojevic-Dupont et al., 2023; Martínez et al., 2025; Dionelis et al., 2025). In the UK these include Energy Performance Certificates (EPC), English Housing Survey (EHS) data, and restricted Valuation Office Agency (VOA) property tax data. The EHS provides a confidence interval for its aggregated age bands, with uncertainties largely unscrutinised. EPC, VOA and EHS datasets are binned into age bands with different definitions, which reduces comparability within institutions and, even more so, between countries (Hawas et al., 2025). The year 1981, for example, would fall into the following age bands respectively for: tax, VOA, 1973-1982; housing, EHS, 1981 - 1990; and energy, EPC, 1976-1982. There is a lack of consistency in naming of construction year information, we choose to refer to the binned construction years of the EPC as EPC bands.

In 2014 a 'Mystery Shopper' experiment was performed (Pyle, 2014), to evaluate the accuracy of the EPC process, but not the EPC bands specifically. In 8% of inspections, the assessor did not ask about the age. The study identified

variations in the EPC ratings, which is supported by results of Hardy and Glew (2019) and Crawley et al. (2019).

However, where data can be crowdsourced systematically, with historical sources, it becomes possible to obtain verified individual construction year estimates. Crowdsourcing of construction year data from historians has been tested in the UK since 2016, using the Colouring Britain open mapping platform, which is part of the Colouring Cities Research Programme (CCRP) (Hudson, 2018; Hecht et al., 2025, 2023). Since 2024 local historians in Loughborough have used the platform to create a comprehensive construction year map of the town.

In this paper, we build on research into the standardization of EPC in Europe and initial comparisons of UK EPC and crowdsourced construction years (Hawas et al., 2025) collected through CCRP platforms. We focus on data uncertainty through systematic comparison of these datasets, using Loughborough, Leicestershire, England as a test case to understand measurable differences.

Biljecki et al. (2023) and Dorn et al. (2015) show that crowdsourced data often exhibit a high degree of agreement with real reference data or can serve as a valid approximation of ground truth. Following this argument, we quantify the uncertainties in the EPC bands. To mitigate the disadvantages of binning, we propose a random forest approach to 'translate' the EPC bands into single years. For this, we derive features developed by Milojevic-Dupont et al. (2020); Nachtigall et al. (2023) using official building footprints from UK Ordnance Survey Master Map (OSMM).

2 Dataset and Methods

In the following section, we introduce the datasets used and the necessary preprocessing for our analysis.

2.1 Data and Software Availability Section

Crowdsourced data is downloadable from Colouring Britain (extract from 2025-11-17). The dataset includes source type and source links. Data for Loughborough is contributed by local historians from the Loughborough Library Local Studies Volunteer Group. This involves visual in-situ or Google Street View inspection, and historical source assessment using town directories, historical publication and official conservation area appraisals, etc. Where no source exists, construction years are estimated by eye. Personal knowledge may also be used. As yet, no prioritisation of specific historical sources has been implemented.

Domestic EPC data are available for England at EPC open data, and include certificate inspection dates from 2007-07-10 to 2025-11-30. EPCs are issued by relevant local authorities for all rented, sold and new properties, and are required to be updated every 10 years. The 4,849 entries for Loughborough with both EPC data and crowdsourced data are binned within construction age bands covering specific year intervals (Tab. 1).

OSMM provides the highest detailed building footprints for the UK. The footprints are used for urban form feature derivation. Due to licensing restrictions relating to OSMM data, we cannot publish the building footprint data itself. However, we provide a different set of footprints, to test the pipeline. The full code and documentation are available at GitHub.

2.2 Data Preprocessing

Both EPC data and crowdsourced data are preprocessed to ensure consistency and omit invalid entries. All processing is done using Python 3.12.3 and *geopandas* (Jordahl et al., 2020) and the visualization executed in *matplotlib* (Hunter, 2007).

Crowdsourced Dataset

The Colouring Britain dataset (under Open Data Commons Open Database Licence (ODbL) v1.0) is a mixture of crowdsourced entries and bulk-uploaded records (e.g., from EPCs — uploaded by the Colouring Britain EPC bot), so crowdsourced records must be identified and separated before analysis. The dataset contains construction years including source type and source links. To distinguish crowdsourced entries from EPC-derived records for Loughborough, we use the platform's edit history provided with the data download. If the most recent edit is not made by the Colouring Britain EPC bot it is presumed to be crowdsourced, and is included in the analysis. If the entry by the bot has mistakenly overwritten an existing crowdsourced construction year entry, but then reversed it, we also include this entry. Additionally, we exclude entries where the upper and lower limit does not agree with the provided construction years. The crowdsourced data are merged with OSMM footprints. The remaining entries are visualized in Fig. 1 (green and orange).

EPC Dataset

For each EPC Building ID we use the most recent EPC (under open government licence). We exclude invalid entries, such as 'INVALID!' or 'NO DATA', and remove duplicates. Where there are multiple EPCs for a single ID with contradictory bands, we remove the entry. In early 2012, the EPC band titled 'England and Wales: 2007 onwards' was removed and replaced by two new bands: 'England and Wales: 2007-2011' and 'England and Wales: 2012 onwards'. Subsequent updating introduced 'England and Wales: 2012-2021' and 'England and Wales: 2022 onwards' bands. Despite this, the older band definitions still occur in newer certificates. In the cases where a more

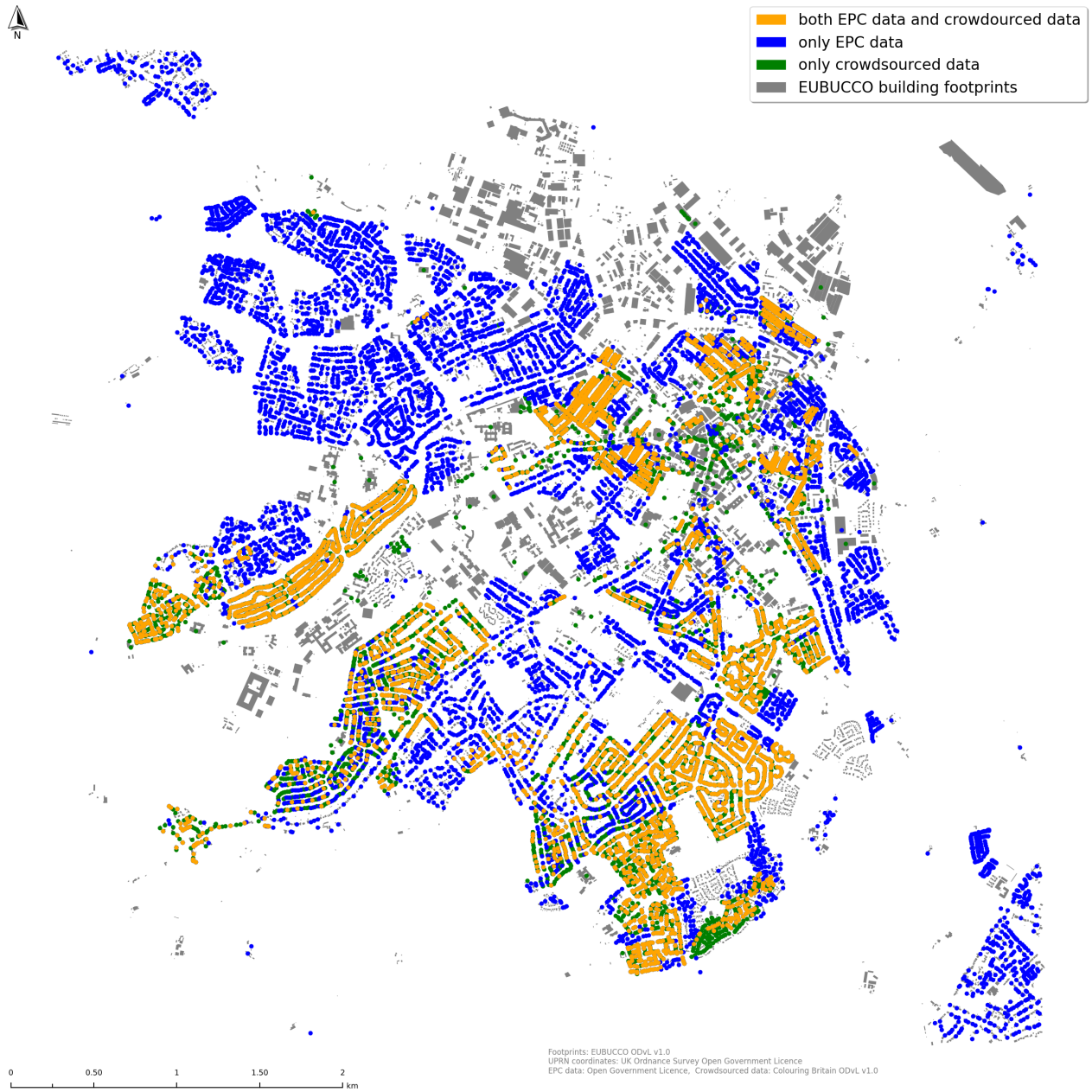

Figure 1. Visualisation of the spatial distribution of the building age data, displayed on the UPRN coordinates under an open government licence. EPC-only data (open government licence) is shown in blue; crowdsourced-only data (Colouring Britain; Open Data Commons Open Database Licenc (ODbL) v1.0) is shown in green. UPRN coordinates with both EPC and crowdsourced data available are shown in orange on top. The underlying map shows the EUBUCCO v0.2 building footprints (Milojevic-Dupont, Nikola and Wagner, Felix et al., 2023), visualized in grey and licensed under ODbL v1.0.

Table 1. EPC bands and their occurrences, median and 68% (1σ) confidence interval and 95% (2σ) confidence interval of the difference (diff) between crowdsourced and EPC data.

EPC band	start	end	median	data points	median (diff)	68% confidence interval (diff)	95% confidence interval (diff)
before 1900	0	1899	1899	962	0.0	[0.0, 4.0]	[0.0, 15.0]
1900-1929	1900	1929	1914.5	834	0.0	[-15.0, 0.0]	[-40.0, 1.0]
1930-1949	1930	1949	1939.5	810	0.0	[-4.0, 0.0]	[-15.0, 1.0]
1950-1966	1950	1966	1958	1209	0.0	[0.0, 0.0]	[-24.0, 1.0]
1967-1975	1967	1975	1971	157	-1.0	[-6.0, 0.0]	[-41.0, 1.9]
1976-1982	1976	1982	1979	167	0.0	[-9.0, 0.0]	[-13.5, 3.2]
1983-1990	1983	1990	1986.5	148	-1.0	[-5.0, 1.0]	[-20.9, 1.0]
1991-1995	1991	1995	1993	214	-0.0	[0.0, 0.0]	[-13.0, 8.0]
1996-2002	1996	2002	1999	171	0.0	[-1.0, 0.0]	[-70.4, 3.0]
2003-2006	2003	2006	2004.5	155	0.0	[-8.0, 0.0]	[-54.5, 1.0]
2007-2011	2007	2011	2009	39	0.0	[0.0, 0.0]	[-126.7, 0.1]
2012-2021	2012	2021	2016	7	0.0	[0.0, 0.0]	[0.0, 0.0]
2007 onwards	2007	2025	2011.5	22	0.0	[0.0, 0.0]	[-111.1, 0.0]
2012 onwards	2012	2025	2018.5	17	0.0	[0.0, 0.0]	[0.0, 0.0]
single years				10	-75.0	[-131.3, -17.6]	[-257.1,-2.8]

recent definition of an EPC band is used in building with an existing certificate, the more precise EPC band of the previous certificate is used, for example where 'England and Wales: 2007 onwards' is replaced with 'England and Wales: 2007-2011'.

Using the Unique Property Reference Numbers (UPRN), we merge UPRN coordinates with the EPC dataset. In cases of multi-property buildings, multiple Building IDs with connected EPCs will map to the same UPRN. We repeat the process undertaken with Building IDs, and in cases where EPC bands contradict each other, we omit the property from the analysis. EPC data is then spatially joined to the OSMM footprints (visualized in Fig. 1 blue and orange).

2.3 Comparison of EPC and Crowdsourced Construction Years

The dataset used for the comparison is highlighted in orange in Fig. 1. The temporal distance between the crowdsourced construction year and the EPC is calculated in cases where a single year has been entered for the crowdsourced data (i.e. without an upper and lower date range). For the rest, the temporal distance between bands is calculated by taking the closest limit.

2.4 Random Forest to Resolve Binning

To estimate a single construction year from the EPC bands in combination with urban form features, we use random forest regression (Liu et al., 2012) implemented in Pedregosa et al. (2011) as the machine learning approach. Entries with single-year bands are omitted, as are years before 1850 because the sample is very small.

We employ two models: one using urban form features, and the other using urban form features and EPC age bands. The shared urban form features were developed by ufo-map and used in Milojevic-Dupont et al. (2020); Nachtigall et al. (2023). The features are separated into four spatial scales: building (total 30, e.g. area and orientation of footprint), building block (total 27, e.g. length and perimeter of block), tessellation based block (total 15, e.g. total block area, block corners), and streets (total 25, e.g. length of closest street, average width of closest street). Due to the smaller area covered by this study, we only include 100 m and 200 m radii in the neighbourhood features, instead of 100 m and 500 m, and omit the city scale features. To derive these features, we use building footprints from OSMM. We omit the spatial proxies 'latitude' and 'longitude' from the feature set, as we are interested in an approach transferable to other regions (Milà et al., 2024). For the model that includes the EPCs, we encode the construction age band as 'start' and 'end'(see Tab. 1).

As a baseline, we use the median of the EPC bands. For the 'Before 1900' band, we use the upper bound, 1899, as it would be an unfair comparison if the whole range was included, or if additional information on the range was provided.

We carried out a spatial autocorrelation-aware test-train-split by calculating tessellation based blocks (Fleischmann et al., 2020) and splitting the data accordingly. Due to the small study area and the spatial split, the test and train sets are not representative of all possible cases. To mitigate this effect, we use a five-fold cross-validation taking spatial autocorrelation into account. This is performed using *sklearns StratifiedGroupKFold* (Pedregosa et al., 2011) to ensure that the non-matching entries between EPC and crowdsourced construction years are evenly distributed.

2.5 Evaluation Methods

As evaluation metrics we use the root-mean-square-error (RMSE), which is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{1}$$

where N is the total number of predictions, $\hat{y}_i$ is the i-th prediction and y_i is the corresponding ground truth. Another metric, denoted as PS4, is the percentage of samples whose predicted year lies within ± 4 years of the target. The last metric is the coefficient of determination (R^2), which is the amount of explained variance by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \sum_{i=1}^{N} y_i)^2} \tag{2}$$

For each of the metrics, we report the median across the five folds.

3 Results and Discussion

In this section, we first present comparison between EPC and crowdsourced data, and then discuss the results of the machine learning approach to resolve the binning.

3.1 Comparison of EPC and Crowdsourced

We obtain a joint (EPC and crowdsourced) datasets containing 4,849 property entries for Loughborough. 33.4% of the EPC bands do not align with the crowdsourced construction years. To account for minor errors in both sources, we apply a tolerance of four years to the temporal distances. This tolerance roughly corresponds to the width of the narrower EPC bands and is small enough to preserve meaningful temporal distinctions while accommodating typical source and recording errors. Including this tolerance, the mismatched entries decrease to 16.9% as shown in Fig. 2. Overall, the EPCs tend to estimate buildings to be more recent than the crowdsourced data. The highest number of counts is found in the two oldest EPC bands (shown in yellow and green in Fig. 2). Buildings in the EPC band as '1900-1929' tend to be recorded as older in the crowdsourced dataset, while buildings in the band 'before 1900' tend to be younger. 24% of the mismatched points of the crowdsourced data fall in the category 'before 1900', but also include buildings in very recent EPC bands.

Considering the distribution of crowdsourced entries by underlying source (Fig. 3), most matching construction years come from entries sourced as 'Historical map' and 'Other database or gazetteer'. The latter also supplies the most mismatching construction years, followed by 'Other website'. However, there is no clear pattern in which sources mismatches occur.

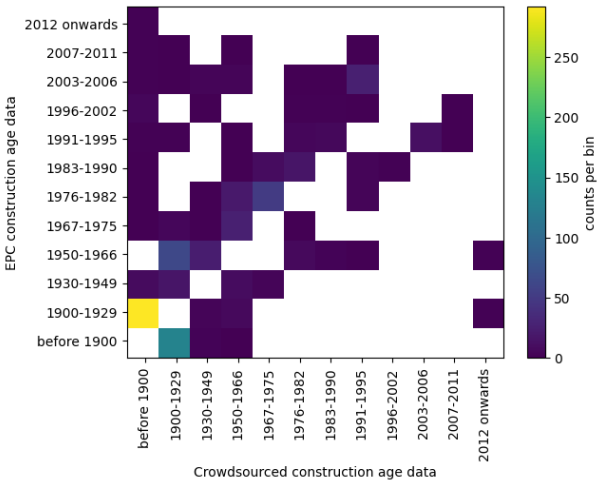

Figure 2. 2D histogram of the overlap of the crowdsourced and EPC data points, while excluding the data points where the labels match. The colour map indicates the number counts, dark blue indicated few entries, yellow indicates the most entries and white indicated 0 entries.

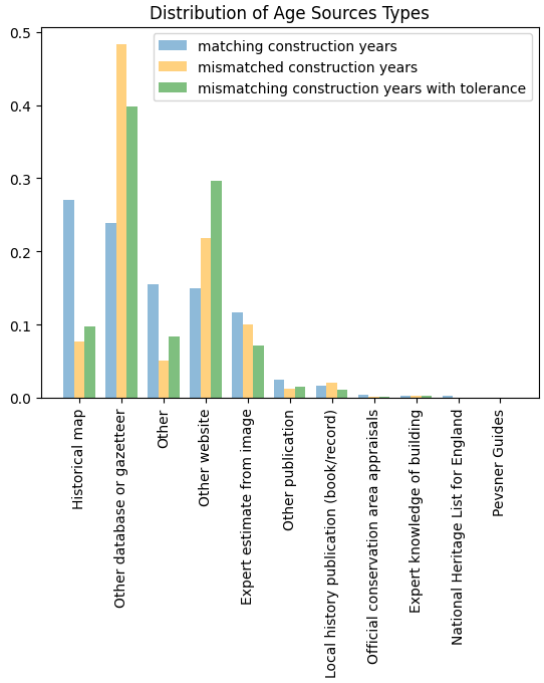

Figure 3. Distribution of the source type for the crowdsourced data. This distribution is shown for the matching construction years (blue), the mismatching construction years (orange) and the mismatching year with a tolerance of four years (green).The distributions are normalized so all individual distribution sum up to one to increase comparability.

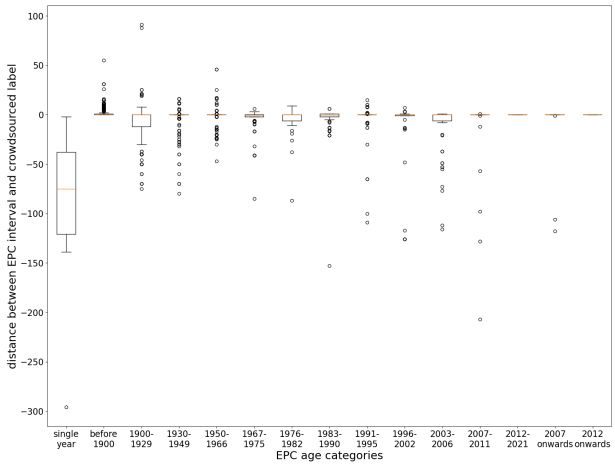

Figure 4. Box plot of the temporal distances between crowdsourced and EPC data.

We assume that the crowdsourced data is a more reliable source than the EPC for building age, since 97.8 % of the entries provide a source (Fig. 3), and the edit history is traceable. In comparison, the EPC bands are determined by energy assessors who are not necessarily trained in building age. Additionally, in Department for Communities and Local Government (2017) it is recommended for occupants (regardless of expertise), to provide a construction year, to increase EPC accuracy. The inconsistencies discovered in the pre-processing of the Loughborough EPC data align with the discovered inconsistency in the variance of the EPC point ratings established by Hardy and Glew (2019); Crawley et al. (2019); Pyle (2014).

Following this argument, we use the crowdsourced data to quantify the uncertainty of each EPC band. Due to the sampling bias towards older buildings introduced with crowdsourced data, the entries with bands younger than '2007-2011' have low significance. The rest of the confidence intervals for temporal distance between EPC bands and crowdsourced, reported in Tab. 1 can be used as guide for EPC band uncertainties. For all ten reported cases with a single year EPC band, the crowdsourced data deviates significantly (Fig. 4). Due to the small sample size, the informative value is low. As the official guide provides no documentation on single-year EPC bands, these entries should be treated with caution or excluded.

3.2 Resolving the Binning

To increase comparability between different datasets, we build a machine learning approach to map the EPC bins to single years (see Sec. 2.4). The crowdsourced data is set as ground truth, following the argument in Sec. 3.1. To evaluate the success of the combination of urban form and EPC data, the median of the EPC bands is used as the baseline. The baseline differs on average by 15.7 years from the crowdsourced data (RMSE = 15.7 years). Only taking into account the 16.9% of mismatched cases, the RMSE rises to 32.0 years. This baseline outperforms the median of the crowdsourced construction years (R^2 = 0.78). Introducing our approach to predict the construction years with both urban form features and EPC band, we obtain a mean RMSE of 11.2 years in the test sets. This is an improvement by four years towards the baseline. 53.0% of the predicted construction years lie within a four-year tolerance of the target. The R^2 is 0.89. There is a notable difference between the test and train performance, which indicates overfitting. This is most likely due to the small size of the dataset, which we attempted to mitigate though hyperparameter optimisation.

For the second reference model utilising random forest regression with only urban form features, an improvement towards the baseline by over one year in the RMSE can be observed (see Tab. 2). In most metrics, the combined model with both urban form and EPC features outperforms the urban form model alone (see Tab. 2). The exception relates to metrics focussing on the mismatched data. These can be explained due to the high importance of EPC features in the prediction and thus confusion in the model. The overfitting is worse for the model where only urban form features are used. This is most likely caused by lack of additional information in the EPC dataset.

Depending on the use case, taking the baseline model may or may not be sufficient. However, we achieve higher accuracies using prediction based on urban form, as well as our new 'translating' approach which can be transferred to any binned data at building level.

3.3 Limitations

The major limitation in this work is the size and balancing of the dataset. Since we cover a period of over 150 years, not every construction year is sufficiently represented. The EPC data also only contains construction year information for domestic properties. Additionally, uncertainties for each EPC band are limited to the Loughborough sample and, as noted by Hardy and Glew (2019), general EPC quality will vary across geographies. This effects the transferability of the obtained confidence intervals. The confidence intervals for bands after 2007 are of limited informative value due to the small sample size.

Though, we assume that the crowdsourced data is close to the ground truth, there is still the possibility of outdatedness of information and human error. By the occlusion inconsistent data points, we introduce a bias in the resolve binning result.

4 Conclusion and Future Work

This study highlights the problem of uncertainty within official construction year data. It compares banded EPC data with individual construction year, collected

Table 2. Mean metrics across the five cross folds for the three different models. Baseline is the median of the EPC band. The suffix 'mismatched' denotes metrics taking only the predictions where the EPC data does not align with the crowdsourced data including tolerance into account

model	RMSE in years				PS4 in %		R^2	
	train	train mismatched	test	test mismatched	train	test	train	test
baseline			15.7	32.0		35.2		0.78
urban form	8.5	11.8	14.5	18.2	80.1	46.8	0.94	0.80
urban form + EPC	7.4	15.0	11.2	19.2	81.8	53.0	0.96	0.89

by historians using Colouring Britain. 16.9% of EPC entries contradict the single estimated construction year entered by historians. On average, EPCs tend to underestimate buildings to be younger than the crowdsourced entries. Assuming the crowdsourced data by local historians is close to the ground truth, we can now identify uncertainties for every EPC band before 2007. This underscores the importance of documenting and communicating data provenance, as well as enhancing the credibility of crowdsourced data by providing more detailed documentation of the data collection process. CCRP platforms such as Colouring Britain should in future introduce features to capture, quantify and store uncertainties as metadata.

The study also tests methods of improving accuracy in automated estimation of year of construction within bands, to facilitate national and international comparison between EPC data, and potentially other official, binned datasets. To 'translate' EPC bands into single years, we established a random forest based regression. Our approach, combining urban form features derived from OSMM, using the ufo-map code, with EPC bands outperforms both reference models.

However, the performance of the model is still insufficient to deploy resolve binning of construction year at larger scale. We expect accuracy to improve by better characterising the relationship between urban form and construction year and by developing, testing and using additional predictive features. Accordingly, we plan to extend the analysis to other UK cities and then to European contexts where CCRP platforms exist. Our aim is to support the generation of stable, scalable construction-year estimates in the absence of ground-truthed data and to assess impacts on downstream tasks (e.g. EPC ratings).

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, DeepL and duck.ai were utilized for language/grammar/punctuation checking or improving linguistic expressions but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

Acknowledgements

With thanks to all our CCRP platform partners and contributors, who give so generously their time, data, resources, knowledge and expertise. We are grateful for the building footprints received from the Ordinance Survey UK. This work was carried out as part of the DRESDENconcept research group SITES.AI and gained financially support by the BMFTR-funded research project ScaDS.AI Dresden/Leipzig. The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by 'Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus' in the programme Center of Excellence for AI-research 'Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig', project identification number: ScaDS.AI.

References

Aksoezen, M., Daniel, M., Hassler, U., and Kohler, N.: Building age as an indicator for energy consumption, Energy and Buildings, 87, 74–86, 2015.

Biljecki, F., Chow, Y. S., and Lee, K.: Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes, Building and Environment, 237, 110 295, 2023.

Crawley, J., Biddulph, P., Northrop, P. J., Wingfield, J., Oreszczyn, T., and Elwell, C.: Quantifying the measurement error on England and Wales EPC ratings, Energies, 12, 3523, 2019.

Department for Communities and Local Government: A guide to energy performance certificates for the marketing, sale and let of dwellings, https://assets.publishing.service.gov.uk/media/5a821a74ed915d74e3401be1/A_guide_to_energy_performance_certificates_for_the_marketing__sale_and_let_of_dwellings.pdf, 2017.

Dionelis, N., Longépé, N., Feliciotti, A., Marconcini, M., Peressutti, D., Kadunc, N. O., Park, J., Sinulingga, H. R., Immanuel, S. A., Tran, B., et al.: Building Age Estimation: A New Multi-Modal Benchmark Dataset and Community Challenge, arXiv preprint arXiv:2502.13818, 2025.

Dorn, H., Törnros, T., and Zipf, A.: Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany, ISPRS International Journal of Geo-Information, 4, 1657–1671, 2015.

Fleischmann, M., Feliciotti, A., Romice, O., and Porta, S.: Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale, Computers, Environment and Urban Systems, 80, 101 441, 2020.

Hardy, A. and Glew, D.: An analysis of errors in the Energy Performance certificate database, Energy policy, 129, 1168–1178, 2019.

Hawas, A., Hudson, P., Kim, M., Hatvani, L., Konieczny, M., Palaiologou, F., Shi, X., and Larimian, T.: New approaches to comparing and visualising national energy rating data for buildings, Energy Proceedings, 9, 2025.

Hecht, R., Danke, T., Herold, H., Hudson, P., Munke, M., and Rieche, T.: Colouring Cities: A Citizen Science Platform for Knowledge Production on the Building Stock-Potentials for Urban and Architectural History, in: Workshop on Research and Education in Urban History in the Age of Digital Libraries, pp. 145–164, Springer, 2023.

Hecht, R., Hudson, P., Gavaldà, O., Hawas, A., Palaiologou, F., Herold, H., and Larimian, T.: Colouring Cities: A network of open platforms to improve the availability of data on the building stock for research, government and society, 28th AGILE Conference on Geographic Information Science, Dresden, https://doi.org/https://doi.org/10.5281/zenodo.15336081, 2025.

Hudson, P.: Urban Characterisation; Expanding Applications for, and New Approaches to Building Attribute Data Capture, The Historic Environment: Policy & Practice, 9, 306–327, 2018.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

International Energy Agency: https://www.iea.org/energy-system/buildings, 2022.

Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F.: geopandas/geopandas: v0.8.1, https://doi.org/10.5281/zenodo.3946761, 2020.

Liu, Y., Wang, Y., and Zhang, J.: New machine learning algorithm: Random forest, in: International conference on information computing and applications, pp. 246–252, Springer, 2012.

Martínez, A. M., Kakoulaki, G., Florio, P., Politis, P., Anselmo, S., Freire, S., Goch, K., Gounari, O., et al.: DBSM R2025: EU Digital Building Stock Model update including satellite-based attributes, 2025.

Milà, C., Ludwig, M., Pebesma, E., Tonne, C., and Meyer, H.: Random forests with spatial proxies for environmental modelling: opportunities and pitfalls, Geoscientific Model Development, 17, 6007–6033, 2024.

Milojevic-Dupont, N., Hans, N., Kaack, L. H., Zumwald, M., Andrieux, F., de Barros Soares, D., Lohrey, S., Pichler, P.-P., and Creutzig, F.: Learning from urban form to predict building heights, PLOS one, 15, e0242 010, 2020.

Milojevic-Dupont, N., Wagner, F., Nachtigall, F., Hu, J., Brüser, G. B., Zumwald, M., Biljecki, F., Heeren, N., Kaack, L. H., Pichler, P.-P., et al.: EUBUCCO v0. 1: European building stock characteristics in a common and open database for 200+ million individual buildings, Scientific data, 10, 147, 2023.

Nachtigall, F., Milojevic-Dupont, N., Wagner, F., and Creutzig, F.: Predicting building age from urban form at large scale, Computers, Environment and Urban Systems, 105, 102 010, 2023.

Oostwegel, L. J., Schorlemmer, D., and Guéguen, P.: From Footprints to Functions: A Comprehensive Global and Semantic Building Footprint Dataset, Scientific Data, 12, 1699, 2025.

OpenStreetMap contributors: Planet dump retrieved from https://planet.osm.org , https://www.openstreetmap.org, 2017.

Overture Maps Foundation : Home - Overture Maps Foundation — overturemaps.org, https://overturemaps.org/, accessed 17-02-2026.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pyle, J.: Green Deal Assessments, https://assets.publishing.service.gov.uk/media/5a7ebd2740f0b6230268b3b5/Green_Deal_Assessment_Mystery_Shopping_FINAL_PUBLISHED.pdf, 2014.

Touzani, S. and Granderson, J.: Open data and deep semantic segmentation for automated extraction of building footprints, Remote sensing, 13, 2578, 2021.

UK Office for National Statistics: Age of the property is the biggest single factor in energy efficiency of homes, https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/ageofthepropertyisthebiggestsinglefactorinenergyefficiencyofhomes/2021-11-01, 2021.

Milojevic-Dupont, Nikola and Wagner, Felix, Nachtigall, F., Hu, J., Brüser, G. B., Zumwald, M., Biljecki, F., Heeren, N., Kaack, L. H., Pichler, P.-P., and Creutzig, F.: EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings, Scientific Data, 10, 147, https://doi.org/10.1038/s41597-023-02040-2, 2023.