Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Leakage-aware multimodal GeoAI for residential rent valuation: evidence from Arabic property listings in four Saudi cities

Othman Almashaqba

othmancs@hotmail.com

Applied Science Private University

Research Article

Keywords: Automated valuation model, GeoAI, Spatial cross-validation, Data leakage, Arabic text mining, Conformal prediction, Saudi Arabia

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11580507/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Automated valuation models (AVMs) increasingly combine structured property attributes with listing text and location, yet their reported accuracy depends on how the data are cleaned and partitioned. This study evaluates a multimodal GeoAI workflow for annual villa rents in Riyadh, Jeddah, Dammam and Al-Khobar using public Aqar listings that pair structured attributes with Arabic descriptions. Listings were linked to a national district gazetteer (98.5% of the analysis sample matched), from which district centroids, district area, distance to the city centre and an inverse-distance spatial lag of neighbouring district rents were derived. A data audit showed that 2,197 of the 3,718 published records (59.1%) were exact duplicates, rising to 86–92% in Dammam and Al-Khobar. Gradient boosting with structured, text and geospatial inputs was compared with hedonic and random-forest baselines under four validation designs. With duplicates retained, random cross-validation gave a median absolute percentage error (MdAPE) of 1.5% and an R^2 of 0.82 on log rent; after de-duplication these became 15.4% and 0.60, and under district-grouped validation 19.3% and 0.55. Under district-grouped validation, geospatial context reduced mean absolute error by 20.6% (95% district-bootstrap interval 14.6–26.1%) relative to structured attributes alone and removed significant residual autocorrelation in Riyadh (Moran's I from 0.40 to − 0.01). Text reduced error by 6.2% on its own but added no significant gain once geospatial context was present. Split-conformal intervals reached 91.6% coverage at the nominal 90% level, while leave-one-city-out transfer degraded all models (MdAPE 26–34%). De-duplication and spatially grouped validation should accompany reported AVM accuracy.

1 Introduction

Residential valuation underpins mortgage lending, taxation, portfolio monitoring and household decisions about where to live. Hedonic theory treats a dwelling's price or rent as the sum of implicit prices for its structural and locational attributes (Rosen 1974), and decades of hedonic studies have catalogued which of those attributes matter (Sirmans et al. 2005). Automated valuation models (AVMs) operationalise this idea at market scale. Comparative studies generally find that tree ensembles such as random forests and gradient boosting are more accurate than linear hedonic regressions (Mayer et al. 2019; Ho et al. 2021), although reviews also stress the interpretability and stability concerns that come with black-box learners (Valier 2020; Wang and Li 2019). Rental markets have received less attention than sales, even though rent estimates are needed for lease pricing, affordability monitoring and housing policy (Füss and Koller 2016; Hu et al. 2019).

Recent work extends AVMs beyond tabular attributes. Photographs, street-level imagery and satellite images have been fused with structured data (Ahmed and Moustafa 2016; Poursaeed et al. 2018; Law et al. 2019; Yousif et al. 2023), and the free text of listings has been shown to carry information on condition, finishing and amenities that fixed fields omit (Nowak and Smith 2017; Shen and Ross 2021; Baur et al. 2023). A recent survey groups these developments under multimodal real-estate appraisal (Huang et al. 2025). In parallel, GeoAI research argues that spatially explicit representations, rather than raw coordinates alone, are needed when learning from geographic data (Janowicz et al. 2020; Kopczewska 2022).

The evaluation of such models has not kept pace with their complexity. House prices and rents are spatially autocorrelated (Basu and Thibodeau 1998), so a random train–test split places near-identical neighbours on both sides of the partition. In ecology and remote sensing, random cross-validation has been shown to overstate predictive skill substantially compared with spatially blocked validation (Roberts et al. 2017; Meyer et al. 2019; Ploton et al. 2020). Listing data add further leakage channels: the same property can be posted repeatedly, and descriptions may state the asking price. More generally, data leakage has been identified as a pervasive source of irreproducible results in machine-learning-based science (Kaufman et al. 2012; Kapoor and Narayanan 2023). A multimodal model evaluated with a leaky design may appear to benefit from extra modalities simply because they make memorisation easier.

Saudi Arabia is a relevant setting. Housing supply and affordability are policy priorities under Vision 2030 (Alqahtany 2021), and preferences for housing attributes have been compared between consumers and property practitioners

(Mulliner and Algrnas 2018). Machine-learning and spatial studies of Saudi property prices have examined land parcels in Riyadh (Louati et al. 2022), apartment attributes in Riyadh (Chikhmous and Rahman 2024) and spatial determinants of housing prices in Dammam (Dano 2025). Arabic listing text poses additional challenges, including dialectal vocabulary and orthographic variation (Darwish et al. 2021). A widely used public dataset of Aqar listings for Riyadh, Jeddah, Dammam and Al-Khobar (Alharbi 2021) has previously been modelled with a neural network evaluated on a single 70/30 split, with near-perfect correlations reported for some cities (Alzain et al. 2022). Whether such accuracy holds under stricter evaluation has not been examined.

This study evaluates an explainable multimodal GeoAI workflow for villa rents using that dataset, with the evaluation design treated as a primary object of study. It contributes (i) a reproducible audit and district-level geocoding of the public Aqar data; (ii) a comparison of four validation designs that isolates the effects of duplicate records and spatial dependence on reported accuracy; (iii) estimates of the marginal contribution of Arabic text and district-level geospatial context, with district-block bootstrap intervals; and (iv) spatial residual diagnostics, conformal prediction intervals and SHAP explanations. Four research questions guide the analysis:

RQ1. How much do duplicate records and random partitioning inflate the apparent accuracy of rent AVMs on listing data?

RQ2. What is the marginal contribution of Arabic listing text and of district-level geospatial context when neighbourhoods are held out?

RQ3. Do the resulting models leave spatially structured residuals, and are their prediction intervals calibrated across cities?

RQ4. How well do models trained in three cities transfer to the fourth?

2 Related work

2.1 Machine-learning valuation and rent models

Early comparisons of machine-learning algorithms for house prices showed gains over linear models in single markets (Park and Bae 2015). On about 40,000 Hong Kong transactions, random forests and gradient-boosting machines outperformed support vector machines (Ho et al. 2021), and on Swiss transactions gradient boosting was the most accurate method while robust regression was the most stable (Mayer et al. 2019). For rents, spatial and temporal structure affects forecast accuracy (Füss and Koller 2016); online rental listings have been used to monitor rents with ensemble learners that reveal district-level clustering (Hu et al. 2019); and large-sample comparisons have weighed kriging against deep neural networks (Seya and Shiroi 2022) and regression-based against machine-learning spatial predictors (Yoshida et al. 2024). Interpretable machine learning applied to rents finds size and age to be leading drivers, with distance-decay effects for locational variables (Lorenz et al. 2023). Gradient-boosted trees remain strong defaults for tabular data of this kind (Friedman 2001; Ke et al. 2017; Grinsztajn et al. 2022).

2.2 Multimodal and text-based valuation

Visual features have improved price estimates for small samples of houses (Ahmed and Moustafa 2016), luxury-level classification from photographs has been combined with attributes (Poursaeed et al. 2018), and street-view and satellite imagery have improved estimates in London, including in boroughs not seen during training (Law et al. 2019). Listing text has been exploited through penalised regression on word counts, which reduced pricing error by more than a quarter in one market (Nowak and Smith 2017), through measures of description uniqueness (Shen and Ross 2021), and through deep text models for Atlanta rents, where combining text with location and attributes outperformed text alone (Zhou et al. 2019). Using descriptions of Berlin rentals and Los Angeles sales, Baur et al. (2023) reported mean absolute error reductions of up to 17%, with smaller gains for low-priced rentals. For Arabic, character n-gram TF-IDF features perform

well on dialectal text (Salameh et al. 2018), and transformer models such as AraBERT are available (Antoun et al. 2020). Many of these studies evaluate with random splits, which leaves open how much of the gain survives when locations are held out.

2.3 Spatial dependence, GeoAI and validation leakage

Spatial autocorrelation in house prices and hedonic residuals is well documented (Basu and Thibodeau 1998) and is commonly diagnosed with Moran's I and its local variants (Moran 1950; Anselin 1995). Machine-learning models can absorb part of this structure through coordinates or engineered neighbourhood features, and local explanation methods can recover spatial effects comparable to those of spatial econometric models (Li 2022). Validation is the weak point. Roberts et al. (2017) showed that blocked cross-validation is almost always more appropriate than random cross-validation for dependent data, Ploton et al. (2020) found that spatial validation turned apparently strong mapping models into ones with little predictive power, and Meyer et al. (2019) showed that highly autocorrelated predictors such as coordinates can produce overfitting that random validation conceals. Wadoux et al. (2021) caution that spatial cross-validation is not a universal answer for map accuracy, and Meyer and Pebesma (2021) argue that error estimates apply only within a model's area of applicability. Different validation designs therefore answer different questions, which is the stance taken in this study.

2.4 Uncertainty and explanation

Conformal prediction provides distribution-free prediction intervals with finite-sample marginal coverage under exchangeability (Vovk et al. 2005; Lei et al. 2018; Angelopoulos and Bates 2023), with quantile-based variants for heteroscedastic targets (Romano et al. 2019) and spatial variants for dependent data (Mao et al. 2024). SHAP values attribute predictions additively to features (Lundberg and Lee 2017), with an exact and fast algorithm for tree ensembles (Lundberg et al. 2020), and have been used to interpret mass-appraisal models (Iban 2022). Attributions describe the fitted model, not causal effects.

3 Study area and data

3.1 Listing data

The listing data are version 4 of the public "Saudi Arabia Real Estate (AQAR)" dataset (Alharbi 2021), which contains 3,718 villa rental advertisements scraped from the Aqar platform for Riyadh, Jeddah, Dammam and Al-Khobar and was collected in late January 2021 (Alzain et al. 2022). Each record has 24 fields: city, district, street frontage, land or building size (m²), property age, counts of bedrooms, bathrooms and living rooms, 14 binary amenity indicators (kitchen, garage, driver's room, maid's room, furnished, air conditioning, roof access, swimming pool, front yard, basement, duplex, internal stairs, elevator and fireplace), the advertised annual rent in Saudi riyals (SAR), and a free-text Arabic description. In the published file the descriptions are truncated at 253 characters; 37.6% of descriptions in the analysis sample are at or within three characters of that limit. The target is the advertised (asking) annual rent, not a contracted rent. Because the records come from a single collection window, no temporal validation is possible.

3.2 District gazetteer and geospatial features

District geometry comes from an openly licensed gazetteer of Saudi regions, cities and districts (Homaily n.d.; GPL-2.0), which the repository documents as compiled from public Saudi National Address map data and which provides district polygons, Arabic and English names and a centre point for each city. District names in the listings were matched to the gazetteer within each city after normalising Arabic orthography (unifying alef variants, ta marbuta and alef maqsura, removing diacritics and tatweel, and dropping the prefix *hay*, "district"). Of the 207 districts in the analysis sample, 201 were matched, covering 1,463 of 1,485 listings (98.5%). The unmatched districts include a generic name in Riyadh (*al-*

Sharq, “the East”, with 11 listings) and the King Khalid International Airport district; their listings retain missing geospatial values and an availability flag rather than being imputed.

For each matched district, the polygon centroid (or an interior point for non-convex shapes), the district area in km² and the great-circle distance from the centroid to the gazetteer city centre were computed. Listings carry only a district name, so all geospatial information is at district resolution. Figure 1 maps the study districts.

3.3 Data audit and cleaning

The audit was carried out before any modelling. Exact duplicates, records identical on all 24 published fields, were removed first. Two pre-specified validity rules were then applied: annual rent between 10,000 and 1,000,000 SAR, which excludes values that are almost certainly monthly rents or entry errors, and size between 100 and 2,000 m². The rules removed 26 and 10 records respectively, leaving an analysis sample of 1,485 listings in 207 districts (Table 1). Applying the same validity rules without de-duplication yields 3,682 records, which are used only to quantify the effect of duplicates. The 1,485 retained listings form 1,447 distinct combinations of city, district, size, age, room counts and rent; the 38 surplus listings are probable reposts with edited text, which district-grouped validation keeps in the same fold but random validation may split.

Table 1
Records, analysis sample and descriptive statistics by city

City	Raw records	Unique records	Analysis sample	Districts	Median rent (IQR), SAR	Median size, m²	Median rent per m², SAR	Description present, %	District matched, %
Riyadh	960	904	888	107	80,000 (50,000–120,000)	340	217	96.6	97.9
Jeddah	888	411	394	49	100,000 (65,375–150,000)	350	267	98.0	99.7
Dammam	894	123	121	26	60,000 (36,000–75,000)	300	173	94.2	98.3
Al-Khobar	976	83	82	25	65,000 (50,000–85,000)	349	190	93.9	100.0
All	3,718	1,521	1,485	207	80,000 (55,000–120,000)	350	220	96.6	98.5

Raw and unique records are before the validity rules. Rents are advertised annual rents for villas. IQR, interquartile range.

4 Methods

Figure 2 Workflow. Listings and the district gazetteer are harmonised, three feature modalities are constructed, and models are assessed with an audit layer of validation designs, prediction intervals and diagnostics.

4.1 Target and feature modalities

The target is the natural logarithm of annual rent. Predictions are back-transformed by exponentiation, which yields a median-type estimate on the rent scale. Three feature modalities were defined.

Structured (S): log size, property age, numbers of bedrooms, bathrooms and living rooms, the 14 amenity indicators, ten one-hot street-frontage categories and four city indicators. City indicators are dropped in the leave-one-city-out design, where the held-out city is never seen in training.

Text (T): descriptions were normalised (Arabic-Indic digits converted, Latin text lower-cased, diacritics and tatweel removed, alef, ya and ta marbuta variants unified, punctuation runs removed). To prevent the advertised price from leaking through the text, every numeral and every currency or number word (for example the Arabic words for riyal, thousand and million, and the written tens) was deleted, which also removes telephone numbers. The cleaned text was represented by TF-IDF word unigrams and bigrams together with character 2–5-grams within word boundaries (minimum document frequency 2, up to 60,000 character features, sublinear term frequency), following evidence that character n-grams suit dialectal Arabic (Salameh et al. 2018). A ridge regression on these features, with its penalty chosen from {0.3, 1, 3, 10, 30} by efficient leave-one-out validation inside the training data, produced a text score. The text modality comprises this score, the description length and a description-present indicator; listings without a description receive a missing score rather than a neutral value.

Geospatial (G): district centroid latitude and longitude, district area, distance to the city centre, a geocoding-availability flag, and a spatial lag defined as the inverse-distance-weighted mean of log rent per m² over the five nearest districts in the same city that contain training listings (distances floored at 0.5 km). In the leave-one-city-out design only the transferable features (distance to centre, district area and the availability flag) are used, because coordinates and the lag carry no information about an unseen city.

The text score and the spatial lag are themselves learned from rents, so they were constructed by stacking. For test rows they are computed from the full training fold; for training rows they are computed out-of-fold through an inner five-fold split that mirrors the outer design. In district-grouped validation, therefore, a listing's own district never contributes to its own spatial lag or text score, matching the conditions faced by a genuinely unseen neighbourhood.

4.2 Models

Three model families were compared: a hedonic ridge regression on standardised structured features (penalty 1.0) as the linear benchmark; a random forest (500 trees, minimum leaf size 2, half of the features per split); and LightGBM gradient boosting (Ke et al. 2017) with 500 trees, learning rate 0.03, 15 leaves, minimum child size 10, row and column subsampling of 0.8 and an L2 penalty of 1. Hyperparameters were fixed before evaluation and not tuned on any test fold, which avoids optimistic selection at the cost of possibly sub-optimal settings. Gradient boosting handles missing values natively. Modality ablations used the same boosting configuration with the feature sets S, S + T, S + G and S + T+G, and two single-modality references were added: the text-only ridge model and boosting on geospatial features with city indicators (G).

4.3 Validation designs

Four designs were compared. (a) Random five-fold cross-validation on the 3,682 records with duplicates retained, reproducing common practice with this dataset. (b) Random five-fold cross-validation on the 1,485 de-duplicated listings, which estimates accuracy for new listings in districts already represented in the training data. (c) District-grouped five-fold cross-validation, in which all listings of a district fall in the same fold, which estimates accuracy for neighbourhoods without training listings. (d) Leave-one-city-out validation, which estimates transfer to a new city. Designs (a)–(c) were repeated with three random seeds; results are reported as mean ± standard deviation across seeds. Design (d) has a single deterministic partition.

4.4 Metrics and statistical comparison

Accuracy is reported as mean absolute error (MAE) and root mean squared error (RMSE) in SAR, median and mean absolute percentage error (MdAPE, MAPE), the share of predictions within ± 10% and ± 20% of the observed rent (PE10,

PE20), and R^2 on the log scale. To compare models, out-of-fold log predictions were averaged across the three seeds and a district-block bootstrap was applied: districts were resampled with replacement 2,000 times, and the percentage change in MAE and the change in MdAPE (percentage points) between two models were recomputed on each resample. Percentile intervals at 95% are reported. Resampling districts rather than listings respects within-district dependence.

4.5 Prediction intervals

Split-conformal intervals (Vovk et al. 2005; Lei et al. 2018) were evaluated within the district-grouped design. In each outer fold, 25% of the training districts were held out as a calibration set, the full S + T+G pipeline was fitted on the remaining districts, and the absolute log residuals of the calibration listings defined the interval half-width as their empirical quantile at level $\lceil (n+1)(1-\alpha) \rceil / n$. Intervals on the rent scale are $\exp(\hat{y} \pm q)$. Two variants were compared: a marginal version with one quantile for all cities, and a city-Mondrian version with a separate quantile per city, falling back to the marginal quantile when a city had fewer than ten calibration listings. Coverage and interval width were assessed for $\alpha = 0.10$ and $\alpha = 0.20$.

4.6 Explanations and spatial residual diagnostics

TreeSHAP (Lundberg and Lee 2017; Lundberg et al. 2020) was applied to an S + T+G model fitted to the full sample, with the text score and spatial lag computed out-of-fold by district. Mean absolute SHAP values were reported per feature and summed by modality. Residual spatial structure was tested with Moran's I (Moran 1950) on district-mean residuals from the seed-averaged district-grouped predictions, using row-standardised five-nearest-neighbour weights within each city and 9,999 random permutations for pseudo p-values.

4.7 Software and reproducibility

The analysis was carried out in Python with scikit-learn 1.9.1 (Pedregosa et al. 2011), LightGBM 4.7.0, SHAP 0.54.1, Shapely 2.2.0 and pandas 3.0.5. Random seeds were fixed throughout. The scripts that reproduce every table and figure from the two public sources are provided as Online Resource 1.

5 Results

5.1 Data audit

Duplication was severe and uneven (Table 1). Of the 3,718 published records, 2,197 (59.1%) were exact duplicates. The share was 5.8% in Riyadh, 53.7% in Jeddah, 86.2% in Dammam and 91.5% in Al-Khobar, so the published file suggests a roughly balanced sample of about 900–980 records per city, whereas the de-duplicated sample is dominated by Riyadh (888 listings) and Jeddah (394), with only 121 listings in Dammam and 82 in Al-Khobar. Descriptions frequently contained figures: 72.2% included at least one numeral and 16.4% included an explicit price, currency or thousand word before masking.

5.2 Effect of the validation design

The validation design changed the apparent accuracy far more than the choice of model or modality (Table 2, Fig. 3). With duplicates retained and random partitioning, the full boosting model reached an MdAPE of 1.5% and an R^2 of 0.82, with 83.3% of predictions within ± 20%. In Dammam and Al-Khobar its MdAPE fell below 1%, because most test records had identical copies in the training folds. After de-duplication, random cross-validation gave an MdAPE of 15.4% and an R^2 of 0.60 for the same model; holding out whole districts raised the MdAPE to 19.3% and lowered R^2 to 0.55; and transfer to an unseen city gave 26.3% and 0.31. Duplicates mainly flattered the flexible learners: the MdAPE of the hedonic ridge model was 15.1% with duplicates and 24.6% without them, whereas that of structured-only boosting rose from 2.9% to 23.2%.

Under district-grouped validation the three structured-only models were nearly indistinguishable (MdAPE 26.6–27.0%, R^2 0.34–0.36). Boosting did not improve on the hedonic ridge model (MAE change + 1.7%, 95% interval − 0.7% to + 4.4%; Table 3). Gains came from additional information rather than from a more flexible learner. Notably, boosting on geospatial features and city indicators alone (MdAPE 24.9%) was at least as accurate as boosting on structured attributes alone (26.8%).

Table 2
Out-of-sample accuracy by validation design and model

Model	MAE, SAR	RMSE, SAR	MdAPE, %	PE20, %	R^2 (log)
Random 5-fold CV, duplicates retained (n = 3,682)					
Hedonic ridge [S]	23,017 ± 31	47,228	15.1 ± 0.2	54.0	0.54 ± 0.00
GBM [S]	13,661 ± 48	41,481	2.9 ± 0.1	78.1	0.73 ± 0.01
GBM [S + T+G]	10,368 ± 5	37,661	1.5 ± 0.1	83.3	0.82 ± 0.00
Random 5-fold CV, deduplicated (n = 1,485)					
Hedonic ridge [S]	32,245 ± 129	66,340	24.6 ± 0.1	41.8	0.41 ± 0.00
Random forest [S]	31,288 ± 130	65,420	22.0 ± 0.0	46.1	0.44 ± 0.00
GBM [S]	31,810 ± 155	65,633	23.2 ± 0.3	44.4	0.43 ± 0.00
Text-only ridge [T]	33,576 ± 40	67,977	25.4 ± 0.4	41.3	0.37 ± 0.00
GBM [G]	32,005 ± 250	68,501	21.3 ± 0.7	47.4	0.41 ± 0.00
GBM [S + T]	29,670 ± 567	63,659	21.2 ± 0.6	47.7	0.49 ± 0.01
GBM [S + G]	25,111 ± 172	60,514	15.6 ± 0.4	59.3	0.59 ± 0.01
GBM [S + T+G]	24,578 ± 416	59,836	15.4 ± 0.8	60.0	0.60 ± 0.01
District-grouped 5-fold CV (n = 1,485)					
Hedonic ridge [S]	33,811 ± 828	67,231	26.6 ± 1.0	38.8	0.36 ± 0.03
Random forest [S]	34,191 ± 542	67,084	27.0 ± 0.7	38.0	0.36 ± 0.03
GBM [S]	34,764 ± 414	67,423	26.8 ± 0.7	38.5	0.34 ± 0.02
Text-only ridge [T]	37,056 ± 317	70,555	29.7 ± 0.8	34.4	0.26 ± 0.01
GBM [G]	34,400 ± 716	70,791	24.9 ± 1.8	41.7	0.34 ± 0.02
GBM [S + T]	32,926 ± 370	65,866	25.3 ± 0.7	39.9	0.41 ± 0.02
GBM [S + G]	27,988 ± 370	62,304	19.5 ± 0.8	50.9	0.54 ± 0.01
GBM [S + T+G]	27,655 ± 732	62,002	19.3 ± 1.0	51.0	0.55 ± 0.01
Leave-one-city-out (n = 1,485)					
Hedonic ridge [S]	34,758	67,498	27.1	36.5	0.35
Random forest [S]	34,603	67,839	27.6	36.8	0.34
GBM [S]	36,283	69,019	28.4	37.9	0.27
Text-only ridge [T]	40,228	73,266	33.8	29.1	0.15
GBM [S + T]	35,406	68,520	27.2	36.6	0.31
GBM [S + G]	36,624	69,248	29.4	36.0	0.24
GBM [S + T+G]	34,747	68,320	26.3	39.8	0.31

Values are means ± standard deviations over three seeds; leave-one-city-out has a single partition. S, structured; T, text; G, geospatial; GBM, LightGBM gradient boosting. MdAPE, median absolute percentage error; PE20, share of predictions

within ± 20%.

5.3 Contribution of text and geospatial context

Holding districts out, adding geospatial context to structured attributes reduced MAE by 20.6% (95% interval 14.6–26.1%) and MdAPE by 6.8 percentage points (Table 3, Fig. 4). Adding text reduced MAE by 6.2% (3.3–9.2%), but its MdAPE change was not significant (− 1.4 points, interval − 3.7 to + 0.3). Once geospatial context was included, text added a further 1.6% reduction in MAE whose interval included zero (− 3.3% to + 0.3%). Under random cross-validation the same increment was small but significant (2.8%, interval 0.4–5.1%), which suggests that part of the information in the descriptions is locational and is used most when listings from the same district appear in training. Geospatial context remained valuable after text was included (MAE − 16.7%, interval − 21.8% to − 11.1%).

Table 3
Marginal contributions of modalities: seed-averaged differences with 95% district-block bootstrap intervals

Comparison	District-grouped CV: ΔMAE, % [95% CI]	District-grouped CV: ΔMdAPE, pp [95% CI]	Random CV: ΔMAE, % [95% CI]
GBM [S] vs hedonic ridge [S]	+ 1.7 [− 0.7, + 4.4]	−0.4 [− 2.5, + 1.1]	−2.3 [− 5.2, + 0.5]
Add text: [S + T] vs [S]	−6.2 [− 9.2, − 3.3]	−1.4 [− 3.7, + 0.3]	−8.6 [− 11.4, − 5.6]
Add geospatial: [S + G] vs [S]	−20.6 [− 26.1, − 14.6]	−6.8 [− 11.5, − 3.6]	−22.0 [− 26.4, − 17.4]
Add both: [S + T+G] vs [S]	−21.8 [− 27.3, − 15.8]	−7.1 [− 11.9, − 3.6]	−24.2 [− 28.5, − 19.6]
Text given geospatial: [S + T+G] vs [S + G]	−1.6 [− 3.3, + 0.3]	−0.2 [− 1.5, + 1.0]	−2.8 [− 5.1, − 0.4]
Geospatial given text: [S + T+G] vs [S + T]	−16.7 [− 21.8, − 11.1]	−5.7 [− 9.2, − 2.5]	−17.1 [− 21.3, − 13.1]

Negative values indicate lower error for the first model. ΔMAE is the percentage change relative to the second model; ΔMdAPE is in percentage points. Intervals are from 2,000 district-level bootstrap resamples.

5.4 City-level accuracy and transfer

The gains were concentrated in the two larger markets (Table 4). Under district-grouped validation, the full model lowered the MdAPE in Riyadh from 29.2% to 18.3% (R^2 from 0.30 to 0.56) and in Jeddah from 25.2% to 21.6%. In Dammam and Al-Khobar, with 121 and 82 listings, it did not improve on simpler models, and the hedonic ridge model had the lowest MdAPE in both (17.2% and 19.8%).

Leave-one-city-out transfer degraded every model (Table 2). Overall MdAPE ranged from 26.3% for the full model to 33.8% for text alone, and R^2 from 0.15 to 0.35, with the hedonic ridge model highest. Transfer to Dammam failed for the full model (MdAPE 50.0%, R^2 −0.14), whereas structured-only boosting reached 23.5%, showing that features learned in other cities can mislead in a new market.

Table 4
City-level accuracy under district-grouped and leave-one-city-out validation

City	Model	District CV: MdAPE, %	District CV: R^2 (log)	City holdout: MdAPE, %	City holdout: R^2 (log)
Riyadh	Hedonic ridge [S]	29.6	0.35	29.2	0.33
Riyadh	GBM [S]	29.2	0.30	31.9	0.22
Riyadh	GBM [S + T+G]	18.3	0.56	26.3	0.28
Jeddah	Hedonic ridge [S]	26.2	0.26	25.0	0.28
Jeddah	GBM [S]	25.2	0.30	23.8	0.24
Jeddah	GBM [S + T+G]	21.6	0.48	23.6	0.31
Dammam	Hedonic ridge [S]	17.2	0.15	27.4	0.07
Dammam	GBM [S]	19.8	0.19	23.5	0.05
Dammam	GBM [S + T+G]	19.8	0.26	50.0	−0.14
Al-Khobar	Hedonic ridge [S]	19.8	0.36	20.1	0.36
Al-Khobar	GBM [S]	20.8	0.31	21.3	0.30
Al-Khobar	GBM [S + T+G]	21.9	0.37	22.6	0.28

District-grouped values are means over three seeds. In the city-holdout column each city is predicted by a model trained on the other three.

5.5 Text leakage test

Keeping numerals and price words in the descriptions improved accuracy only slightly. Under random cross-validation, the MdAPE of boosting on structured and text features was 20.6% with unmasked text and 21.2% with masked text, and that of the text-only model 24.9% versus 25.4%. Under district-grouped validation, the differences were 0.0 and 0.5 percentage points respectively. The price leakage channel in this dataset is therefore small, probably because descriptions are truncated and prices are often written in words, but masking removes it at negligible cost and is retained in all other results.

5.6 Spatial structure of residuals

Models without geospatial context left significant positive spatial autocorrelation in district-mean residuals (Table 5). In Riyadh, Moran's I was 0.47 for the hedonic ridge model and 0.40 for structured-only boosting (both $p < 0.001$); with geospatial context it fell to − 0.01 ($p = 0.51$). Jeddah and Dammam showed the same pattern at smaller magnitudes, while Al-Khobar showed no significant autocorrelation for any model. Figure 5 shows that the structured-only model under-predicted a contiguous belt of north-central districts in Riyadh and over-predicted much of the south and the outer periphery, and that these contiguous patterns largely disappear in the full model.

Table 5
Moran's I of district-mean residuals under district-grouped validation

Model	Riyadh (103)	Jeddah (48)	Dammam (25)	Al-Khobar (25)
Hedonic ridge [S]	0.47 (p < 0.001)	0.12 (p = 0.043)	0.10 (p = 0.091)	0.06 (p = 0.155)
GBM [S]	0.40 (p < 0.001)	0.14 (p = 0.029)	0.16 (p = 0.034)	0.03 (p = 0.211)
GBM [S + G]	−0.01 (p = 0.512)	−0.10 (p = 0.855)	−0.07 (p = 0.562)	−0.18 (p = 0.942)
GBM [S + T+G]	−0.01 (p = 0.507)	−0.08 (p = 0.778)	0.02 (p = 0.257)	−0.19 (p = 0.959)

Numbers of districts in parentheses. Weights: row-standardised five nearest neighbours within each city. Pseudo p-values from 9,999 permutations (one-sided, positive autocorrelation).

5.7 Prediction intervals

At the nominal 90% level, marginal split-conformal intervals covered 91.6% of held-out rents overall but 89.6% in Jeddah; at the 80% level Jeddah coverage fell to 78.2% (Table 6, Fig. 6). City-Mondrian calibration brought every city to at least the nominal level (90.1–95.1% at 90%; 84.3–92.7% at 80%) at the cost of slightly wider intervals. The intervals are wide: at 90% the median width was 103,275 SAR for marginal calibration, about 1.28 times the predicted rent. This reflects the limited information in district-level listing data and is reported rather than hidden.

Table 6
Empirical coverage (%) of split-conformal prediction intervals under district-grouped validation

Nominal level	Method	Riyadh	Jeddah	Dammam	Al-Khobar	All	Median width, SAR (All)
90%	Marginal split-conformal	92.0	89.6	94.2	92.7	91.6	103,275
90%	City-Mondrian split-conformal	92.3	92.4	90.1	95.1	92.3	109,670
80%	Marginal split-conformal	85.2	78.2	85.1	85.4	83.4	73,982
80%	City-Mondrian split-conformal	84.5	84.3	89.3	92.7	85.3	79,306

5.8 Model explanations

The spatial lag was the most influential feature of the full model (mean absolute SHAP 0.218 log units), followed by log size (0.114), the text score (0.070), air conditioning (0.058) and distance to the city centre (0.051) (Fig. 7). Summed by modality, structured features accounted for 47.5% of the total mean absolute attribution, geospatial features for 42.2% and text for 10.3%. The shares were similar in each city (geospatial 41.0–45.2%, text 7.5–12.7%). These attributions describe how the fitted model uses its inputs; they do not identify causal effects of, for example, distance or amenities on rent.

6 Discussion

6.1 Evaluation design dominates reported accuracy

The most consequential decision in this analysis was not the model but the treatment of duplicates and the partitioning scheme (RQ1). Exact duplicates made up nine in ten records in Dammam and Al-Khobar, and random validation on the

published file reduced the full model's MdAPE from 15.4% to 1.5%. Earlier work on the same dataset reported correlations close to 100% for Dammam on a single 70/30 split, without describing de-duplication or grouped validation (Alzain et al. 2022). The present results are consistent with duplication contributing to such figures, although they cannot establish it for that study. More generally, the findings align with evidence from other spatial domains that random validation overstates predictive skill (Roberts et al. 2017; Ploton et al. 2020), and with warnings about leakage in applied machine learning (Kapoor and Narayanan 2023). Multimodal AVM studies that rely on scraped listings should report exact and near-duplicate counts, the grouping used in validation, and city-level results.

The three de-duplicated designs answer different questions rather than ranking one another (Wadoux et al. 2021). Random validation on de-duplicated data approximates in-fill valuation within districts already represented in the data; district-grouped validation approximates valuation in neighbourhoods without comparable listings; and city holdout approximates deployment in a new market. Reporting all three gives users a realistic range of expected error.

6.2 Geospatial context matters more than text

When neighbourhoods were held out, district-level geospatial context was the most valuable addition (RQ2), and the spatial lag of neighbouring district rents was the most influential feature. This agrees with the long-standing importance of location in hedonic models (Sirmans et al. 2005), with evidence that simple coordinate and neighbourhood features capture much of the spatial dependence in rents (Yoshida et al. 2024), and with the use of explanation methods to recover spatial effects from tree ensembles (Li 2022). It also resolves the residual autocorrelation left by structured-only models (RQ3), which is the main geomatics contribution of the workflow: the spatial lag turns district polygons and neighbouring rents into a representation that a learner can use for unseen districts.

The text contribution was smaller than in earlier studies (Nowak and Smith 2017; Baur et al. 2023) and largely redundant once location was modelled. Three features of the data plausibly explain this: descriptions are truncated at 253 characters, many repeat information already in the structured fields, and many mention the district or nearby landmarks, which overlaps with geospatial context. A bag-of-n-grams representation may also under-use the text. Transformer encoders such as AraBERT (Antoun et al. 2020) are a natural next step, but should be evaluated with the same grouped design and masking, because the present results show that random validation inflates the apparent value of text.

6.3 Transferability and uncertainty

No model transferred well between cities (RQ4). The failure of the full model in Dammam illustrates the area-of-applicability problem (Meyer and Pebesma 2021): relationships between rent, distance to centre and district size learned in Riyadh and Jeddah do not hold in Dammam, a smaller market that forms part of a multi-city metropolitan area with Al-Khobar and Dhahran. In practice, a model for a new city needs at least some local listings for recalibration. The conformal analysis shows that marginal calibration can hide under-coverage in a particular city, here Jeddah, and that city-conditional calibration corrects this at modest cost. Because conformal guarantees assume exchangeability, which spatial dependence weakens, spatially adaptive variants (Mao et al. 2024) deserve evaluation in future work.

6.4 Limitations

The study has several limitations. First, the data are a single snapshot from January 2021; rents in Saudi cities have changed since, so the models are not valid for current valuation and temporal validation was impossible. Second, the target is the asking rent of advertised villas, which may differ systematically from contracted rents, and a few very high values (nine listings at or above 500,000 SAR) may be mislabelled sale prices that the validity rules did not remove. Third, geolocation is at district level only; within-district variation, street-level accessibility and points of interest could not be represented. Fourth, the dataset contains no images, so the visual modality common in multimodal AVMs could not be tested. Fifth, Dammam and Al-Khobar contribute few listings after de-duplication, which limits city-specific conclusions. Sixth, hyperparameters were fixed rather than tuned in nested loops, so absolute accuracy may be slightly understated for

the flexible models. Finally, the dataset licence does not permit redistribution of the raw records, so reproduction requires downloading them from the original source.

7 Conclusion

This study examined an explainable multimodal GeoAI workflow for villa rents in four Saudi cities, with the evaluation design treated as part of the research question. Three conclusions follow. First, duplicate listings and random partitioning can turn a model with roughly 19% median error on unseen districts into one that appears to have under 2% error, so de-duplication and grouped validation are prerequisites for credible AVM accuracy claims. Second, district-level geospatial context, in particular a spatial lag of neighbouring district rents, reduced error by about one fifth for unseen neighbourhoods and removed the spatial residual structure left by structured attributes, whereas truncated Arabic descriptions added little once location was modelled. Third, models did not transfer reliably between cities, and city-conditional conformal intervals were needed for calibrated uncertainty in every city. Future work should use transaction-linked rents with dates, finer geolocation and imagery, evaluated with the same leakage controls.

Statements and declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests: The author has no relevant financial or non-financial interests to disclose.

Data availability: The listing data are publicly available from Kaggle (Alharbi 2021; dataset version 4). Their licence (“Data files © Original Authors”) does not permit redistribution, so the raw records are not included with this article. The district gazetteer is available under GPL-2.0 (Homaily n.d.). Derived district-level summaries and all result tables are included in Online Resource 1.

Code availability: The Python scripts that reproduce the data audit, experiments, tables and figures are provided as Online Resource 1.

Ethics approval: Not applicable. The study uses publicly available secondary data on advertised properties and contains no information on human participants. Numerals, including telephone numbers, were removed from listing descriptions before analysis.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

Clinical trial number: Not applicable.

Author contributions: Othman Almashaqba: conceptualisation, methodology, data curation, formal analysis, visualisation, writing – original draft, writing – review and editing.

Use of generative AI: During the preparation of this work the author used Claude (Anthropic) to assist with writing analysis code and with drafting and language editing of the manuscript. The author reviewed and edited the code, results and text and takes full responsibility for the content of the publication.

References

1. Ahmed EH, Moustafa M (2016) House price estimation from visual and textual features. In: Proceedings of the 8th International Joint Conference on Computational Intelligence (IJCCI 2016), vol 3. SciTePress, pp 62–68. https://doi.org/10.5220/0006040700620068
2. Alharbi L (2021) Saudi Arabia Real Estate (AQAR) [dataset], version 4. Kaggle. https://www.kaggle.com/datasets/lama122/saudi-arabia-real-estate-aqar. Accessed 20 October 2026
3. Alqahtany A (2021) Affordable housing in Saudi Arabia's vision 2030: new developments and new challenges. International Journal of Housing Markets and Analysis 14(1):243–256. https://doi.org/10.1108/IJHMA-04-2020-0035
4. Alzain E, Alshebami AS, Aldhyani THH, Alsubari SN (2022) Application of artificial intelligence for predicting real estate prices: the case of Saudi Arabia. Electronics 11(21):3448. https://doi.org/10.3390/electronics11213448
5. Angelopoulos AN, Bates S (2023) Conformal prediction: a gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101
6. Anselin L (1995) Local indicators of spatial association—LISA. Geographical Analysis 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
7. Antoun W, Baly F, Hajj H (2020) AraBERT: transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4). European Language Resources Association, Marseille, pp 9–15
8. Basu S, Thibodeau TG (1998) Analysis of spatial autocorrelation in house prices. Journal of Real Estate Finance and Economics 17(1):61–85. https://doi.org/10.1023/A:1007703229507
9. Baur K, Rosenfelder M, Lutz B (2023) Automated real estate valuation with machine learning models using property descriptions. Expert Systems with Applications 213:119147. https://doi.org/10.1016/j.eswa.2022.119147
10. Chikhmous A, Rahman MT (2024) Examining the effect of apartment attributes on their sale prices in Riyadh, Saudi Arabia. Spatial Information Research 32(4):411–424. https://doi.org/10.1007/s41324-023-00565-7
11. Dano UL (2025) Analyzing the spatial determinants of housing prices in Dammam, Saudi Arabia: an AHP approach. International Journal of Housing Markets and Analysis 18(2):354–377. https://doi.org/10.1108/IJHMA-07-2023-0101
12. Darwish K, Habash N, Abbas M, Al-Khalifa H, Al-Natsheh HT, El-Beltagy SR, Bouamor H, Bouzoubaa K, Cavalli-Sforza V, El-Hajj W, Jarrar M, Mubarak H (2021) A panoramic survey of natural language processing in the Arab world. Communications of the ACM 64(4):72–81. https://doi.org/10.1145/3447735
13. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
14. Füss R, Koller JA (2016) The role of spatial and temporal structure for residential rent predictions. International Journal of Forecasting 32(4):1352–1368. https://doi.org/10.1016/j.ijforecast.2016.06.001
15. Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track, pp 507–520
16. Ho WKO, Tang B-S, Wong SW (2021) Predicting property prices with machine learning algorithms. Journal of Property Research 38(1):48–70. https://doi.org/10.1080/09599916.2020.1832558
17. Homaily A (n.d.) Saudi Arabia regions, cities and districts [GitHub repository, GPL-2.0]. https://github.com/homaily/Saudi-Arabia-Regions-Cities-and-Districts. Accessed 20 October 2026
18. Hu L, He S, Han Z, Xiao H, Su S, Weng M, Cai Z (2019) Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land Use Policy 82:657–673. https://doi.org/10.1016/j.landusepol.2018.12.030
19. Huang C, Li Z, Chen F, Liang B (2025) Multimodal machine learning for real estate appraisal: a comprehensive survey. arXiv:2503.22119. https://doi.org/10.48550/arXiv.2503.22119

20. Iban MC (2022) An explainable model for the mass appraisal of residences: the application of tree-based machine learning algorithms and interpretation of value determinants. Habitat International 128:102660. https://doi.org/10.1016/j.habitatint.2022.102660
21. Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B (2020) GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science 34(4):625–636. https://doi.org/10.1080/13658816.2019.1684500
22. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4(9):100804. https://doi.org/10.1016/j.patter.2023.100804
23. Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data 6(4):15. https://doi.org/10.1145/2382577.2382579
24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp 3146–3154
25. Kopczewska K (2022) Spatial machine learning: new opportunities for regional science. Annals of Regional Science 68(3):713–755. https://doi.org/10.1007/s00168-021-01101-x
26. Law S, Paige B, Russell C (2019) Take a look around: using street view and satellite images to estimate house prices. ACM Transactions on Intelligent Systems and Technology 10(5):1–19. https://doi.org/10.1145/3342240
27. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. Journal of the American Statistical Association 113(523):1094–1111. https://doi.org/10.1080/01621459.2017.1307116
28. Li Z (2022) Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. Computers, Environment and Urban Systems 96:101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845
29. Lorenz F, Willwersch J, Cajias M, Fuerst F (2023) Interpretable machine learning for real estate market analysis. Real Estate Economics 51(5):1178–1208. https://doi.org/10.1111/1540-6229.12397
30. Louati A, Lahyani R, Aldaej A, Aldumaykhi A, Otai S (2022) Price forecasting for real estate using machine learning: a case study on Riyadh city. Concurrency and Computation: Practice and Experience 34(6):e6748. https://doi.org/10.1002/cpe.6748
31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9
32. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp 4765–4774
33. Mao H, Martin R, Reich BJ (2024) Valid model-free spatial prediction. Journal of the American Statistical Association 119(546):904–914. https://doi.org/10.1080/01621459.2022.2147531
34. Mayer M, Bourassa SC, Hoesli M, Scognamiglio D (2019) Estimation and updating methods for hedonic valuation. Journal of European Real Estate Research 12(1):134–150. https://doi.org/10.1108/JERER-08-2018-0035
35. Meyer H, Pebesma E (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12:1620–1633. https://doi.org/10.1111/2041-210X.13650
36. Meyer H, Reudenbach C, Wöllauer S, Nauss T (2019) Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. Ecological Modelling 411:108815. https://doi.org/10.1016/j.ecolmodel.2019.108815
37. Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1–2):17–23. https://doi.org/10.1093/biomet/37.1-2.17

38. Mulliner E, Algrnas M (2018) Preferences for housing attributes in Saudi Arabia: a comparison between consumers' and property practitioners' views. Cities 83:152–164. https://doi.org/10.1016/j.cities.2018.06.018
39. Nowak A, Smith P (2017) Textual analysis in real estate. Journal of Applied Econometrics 32(4):896–918. https://doi.org/10.1002/jae.2550
40. Park B, Bae JK (2015) Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data. Expert Systems with Applications 42(6):2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
41. 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, Duchesnay É (2011) Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12:2825–2830
42. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet-Fleury S, Pélissier R (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11:4540. https://doi.org/10.1038/s41467-020-18321-y
43. Poursaeed O, Matera T, Belongie S (2018) Vision-based real estate price estimation. Machine Vision and Applications 29(4):667–676. https://doi.org/10.1007/s00138-018-0922-2
44. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881
45. Romano Y, Patterson E, Candès E (2019) Conformalized quantile regression. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), pp 3543–3553
46. Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy 82(1):34–55. https://doi.org/10.1086/260169
47. Salameh M, Bouamor H, Habash N (2018) Fine-grained Arabic dialect identification. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018). Association for Computational Linguistics, Santa Fe, pp 1332–1344
48. Seya H, Shiroi D (2022) A comparison of residential apartment rent price predictions using a large data set: kriging versus deep neural network. Geographical Analysis 54(2):239–260. https://doi.org/10.1111/gean.12283
49. Shen L, Ross S (2021) Information value of property description: a machine learning approach. Journal of Urban Economics 121:103299. https://doi.org/10.1016/j.jue.2020.103299
50. Sirmans S, Macpherson D, Zietz E (2005) The composition of hedonic pricing models. Journal of Real Estate Literature 13(1):1–44. https://doi.org/10.1080/10835547.2005.12090154
51. Valier A (2020) Who performs better? AVMs vs hedonic models. Journal of Property Investment & Finance 38(3):213–225. https://doi.org/10.1108/JPIF-12-2019-0157
52. Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer, New York. https://doi.org/10.1007/b106715
53. Wadoux AMJ-C, Heuvelink GBM, de Bruin S, Brus DJ (2021) Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling 457:109692. https://doi.org/10.1016/j.ecolmodel.2021.109692
54. Wang D, Li VJ (2019) Mass appraisal models of real estate in the 21st century: a systematic literature review. Sustainability 11(24):7006. https://doi.org/10.3390/su11247006
55. Yoshida T, Murakami D, Seya H (2024) Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. Journal of Real Estate Finance and Economics 69(1):1–28. https://doi.org/10.1007/s11146-022-09929-6
56. Yousif A, Baraheem S, Vaddi SS, Patel VS, Shen J, Nguyen TV (2023) Real estate pricing prediction via textual and visual features. Machine Vision and Applications 34(6):126. https://doi.org/10.1007/s00138-023-01464-5

57. Zhou X, Tong W, Li D (2019) Modeling housing rent in the Atlanta metropolitan area using textual information and deep learning. ISPRS International Journal of Geo-Information 8(8):349. https://doi.org/10.3390/ijgi8080349

Figures

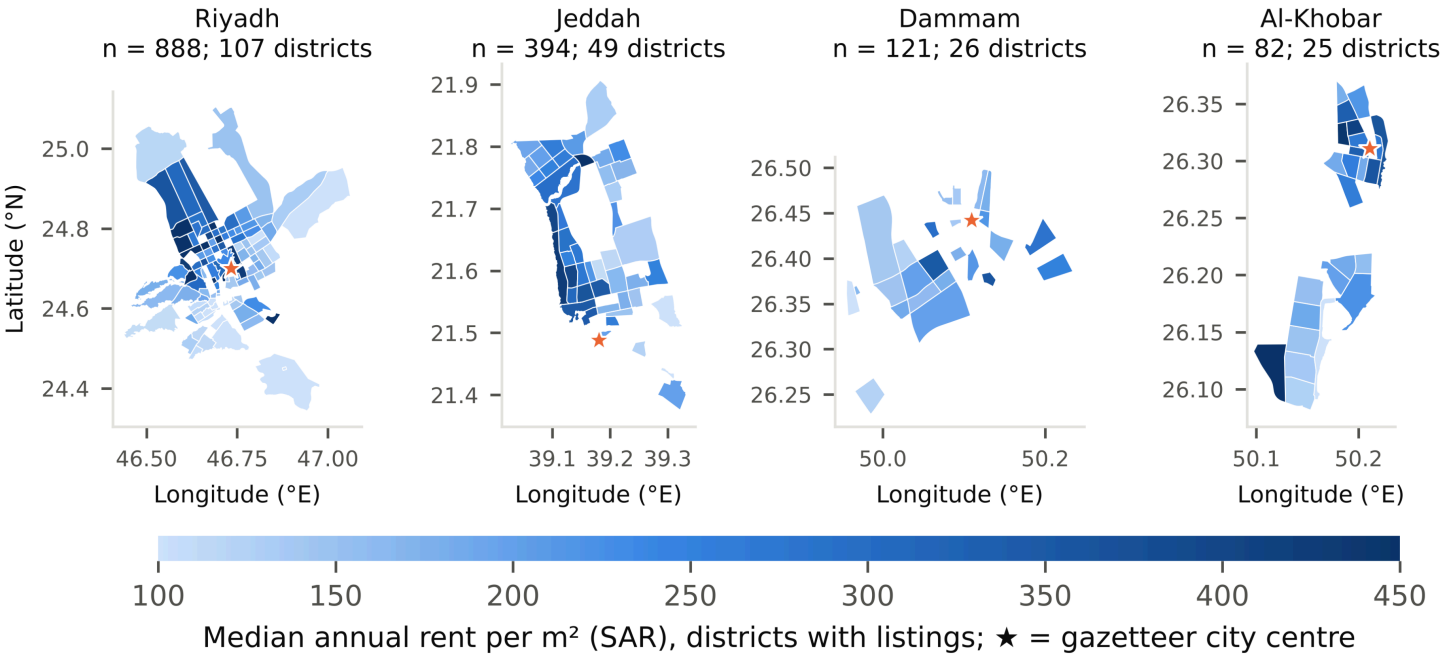
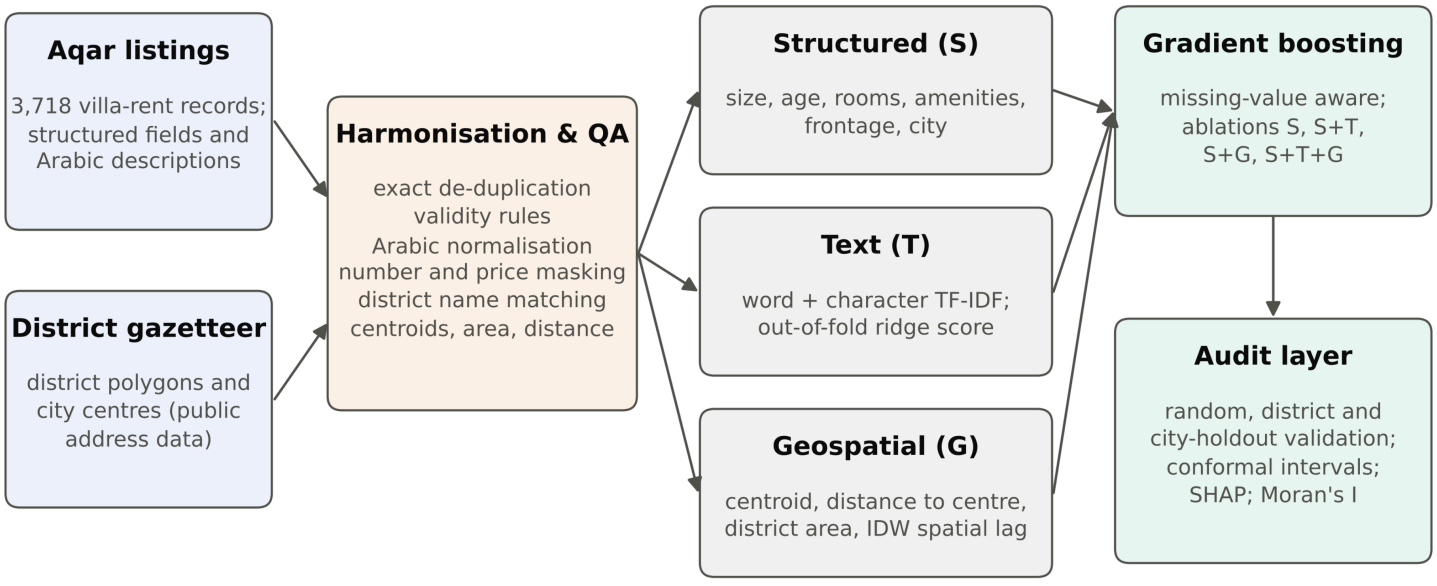


Figure 1

Study area. Districts with at least one listing in the analysis sample, shaded by median annual rent per m². The star marks the city-centre point given in the gazetteer. District polygons: Homaily (n.d.).



All learned steps (text model, spatial lag, booster, calibration quantile) are fitted on training rows only.

Figure 2

Workflow. Listings and the district gazetteer are harmonised, three feature modalities are constructed, and models are assessed with an audit layer of validation designs, prediction intervals and diagnostics.

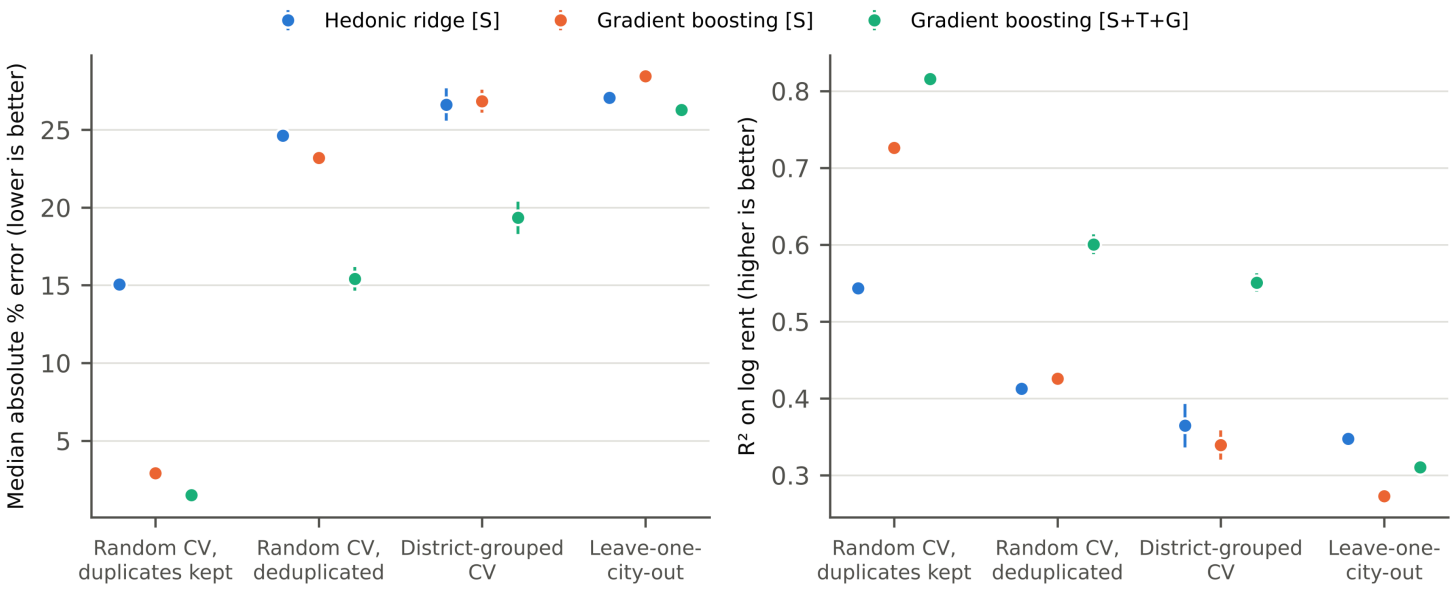


Figure 3

Accuracy of three models under four validation designs. Points are means over three seeds and bars show ±1 standard deviation; leave-one-city-out has a single partition.

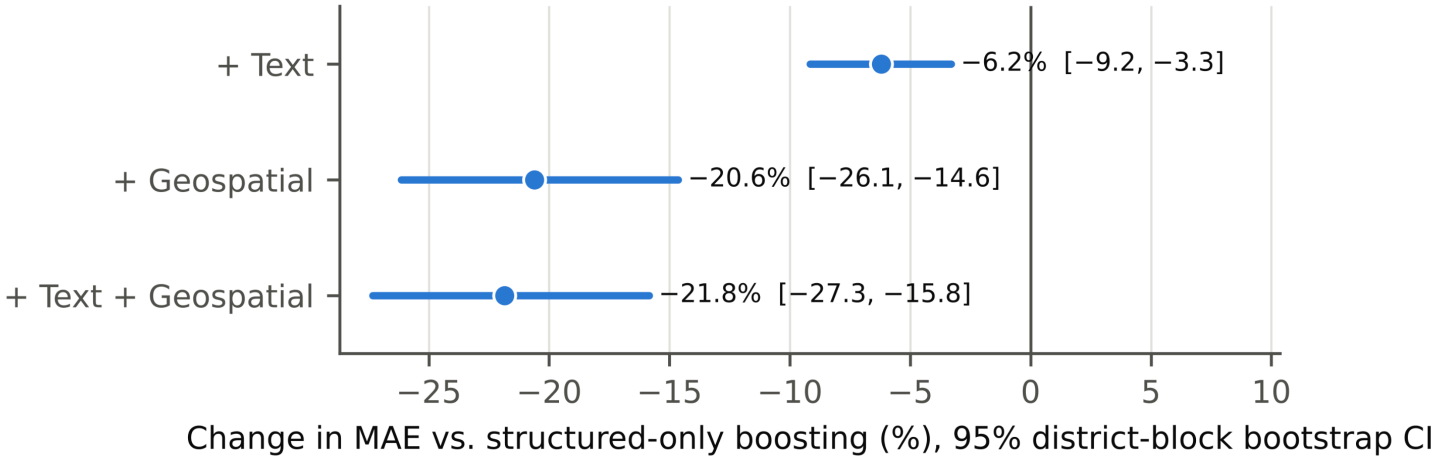


Figure 4

Change in MAE relative to structured-only gradient boosting under district-grouped validation, with 95% district-block bootstrap intervals.

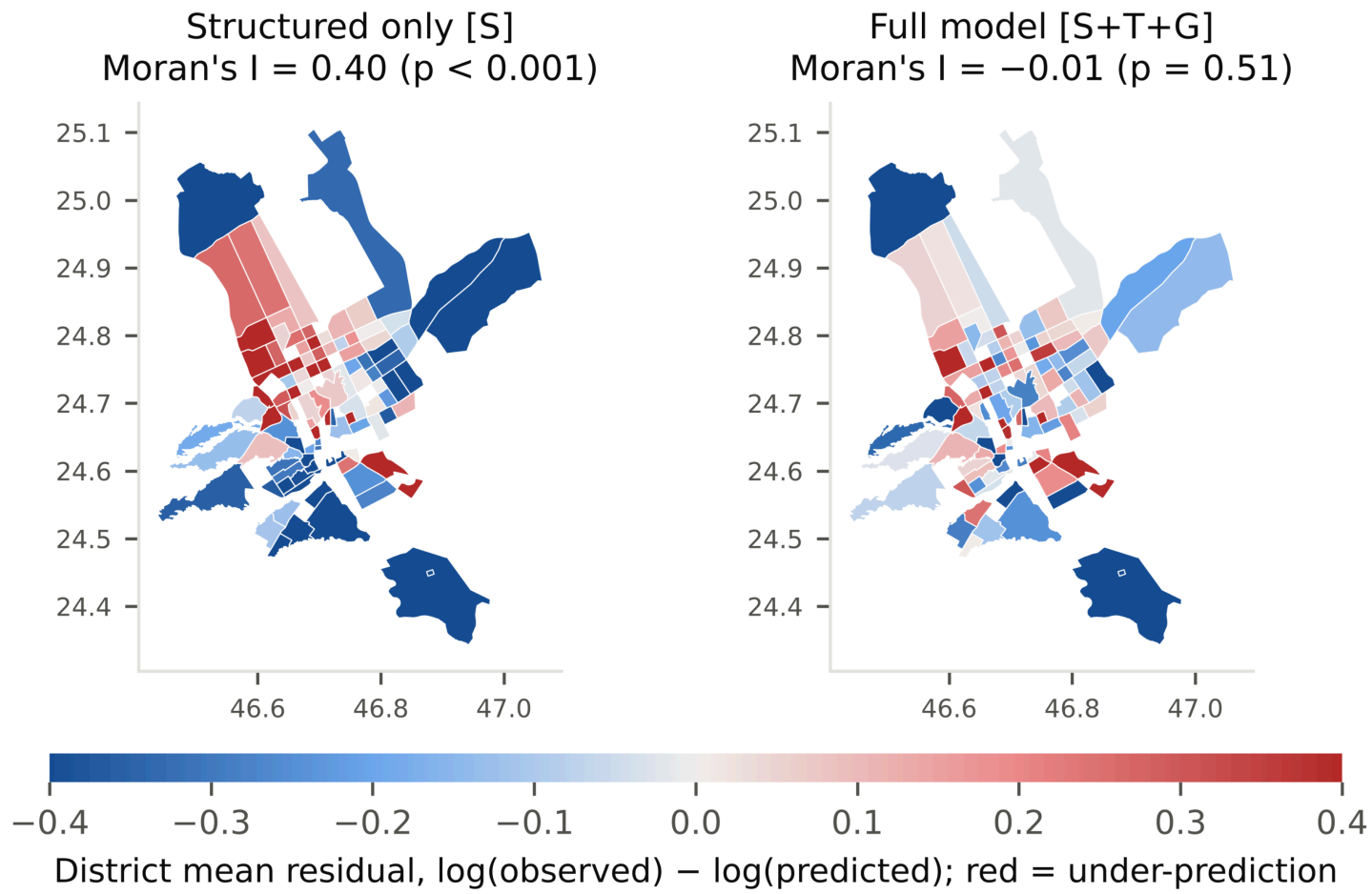


Figure 5

District-mean residuals in Riyadh under district-grouped validation for structured-only boosting (left) and the full model (right). Red indicates under-prediction and blue over-prediction; values are clipped at ±0.4 log units.

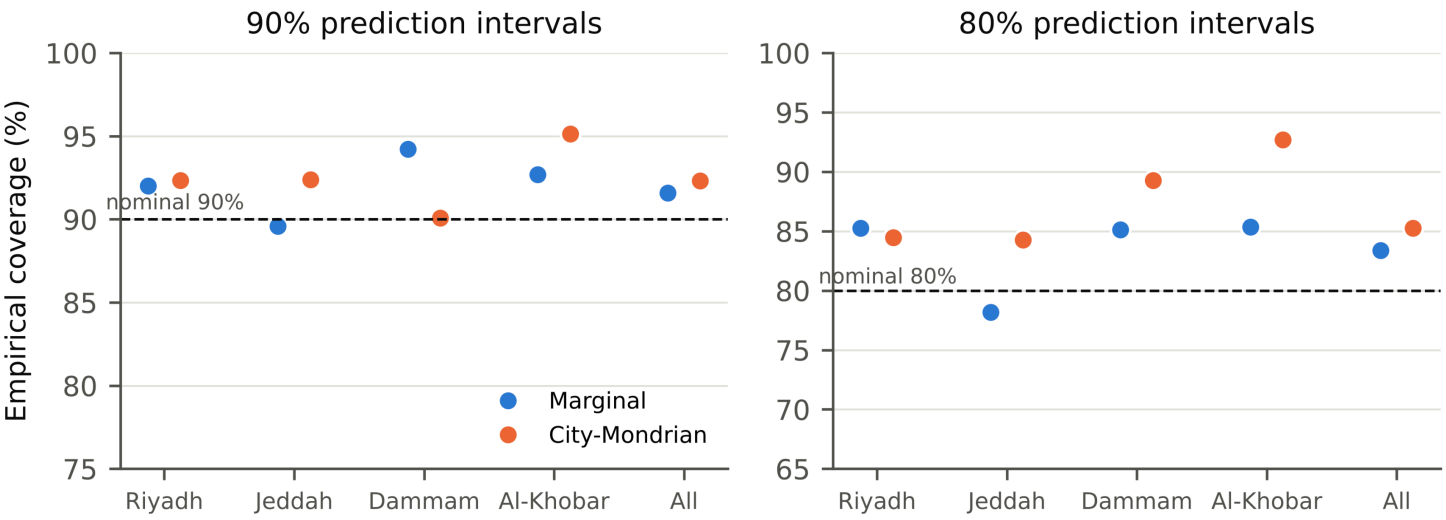


Figure 6

Empirical coverage of marginal and city-Mondrian split-conformal intervals by city at nominal 90% (left) and 80% (right) levels.

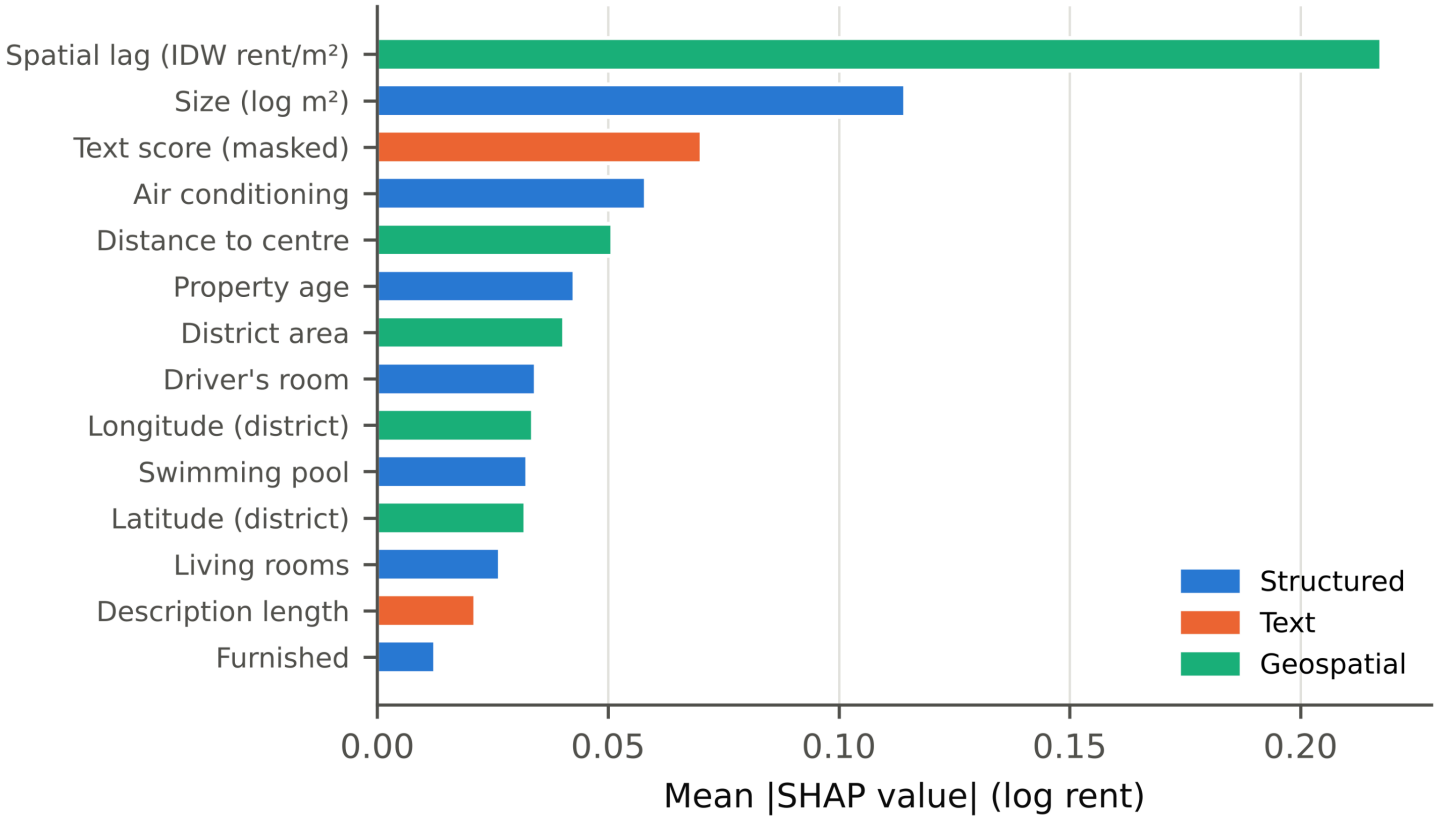


Figure 7

Global SHAP importance of the 14 most influential features of the full model, coloured by modality.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1.zip