Inferring Property Typology in Real Estate Tax Records via Domain-Aware Feature Alignment

Luís Henrique Alves Rosa[1], **Gustavo Luiz Bueno Pereira**[1], **Lucas Araújo Pereira**[1], **Luiz Mário Lustosa Pascoal**[1], **Sávio Salvarino Teles de Oliveira**[1]

[1]Universidade Federal de Goiás (UFG)

{luis_alves,gustavobueno}@discente.ufg.br, apereiral@ufg.br

luizmlpascoal@gmail.com, savioteles@ufg.br

Abstract. *Real estate tax records in Brazil typically contain transaction values and built areas but lack structural attributes such as bedroom counts, limiting comparative market analyses. We propose a source-domain feature engineering approach to infer bedroom counts in unlabelled tax records using labelled listing portal data. To address the domain shift caused by the gap between announced and declared prices (66% median difference), we introduce an intra-neighbourhood relative price metric that corrects the shift without cross-dataset parameter estimation. Evaluated on a held-out listing test set, k-Nearest Neighbours (k=15) achieves 93.4% accuracy, while a density-based pipeline (UMAP and HDBSCAN) achieves 80.9%, providing interpretable clusters and per-record uncertainty estimates (ECE = 0.026). A zero-shot cross-portal experiment on 6,882 external listings yields 84.4% accuracy, validating the portability of the relative price feature. The method is evaluated on 1,542 tax transactions across 39 neighbourhoods in Goiânia, Brazil.*

Keywords: *ITBI, domain shift, property typology, UMAP, HDBSCAN, KNN, urban informatics.*

1. Introduction

The *Imposto sobre Transmissão de Bens Imóveis* (ITBI) is a municipal tax levied on real estate transfers in Brazil [Brasil 1988]. Ensuring declared values reflect actual market conditions is essential for fair taxation: under-declaration erodes the municipal tax base, shifts the fiscal burden onto compliant taxpayers, and distorts the property-value data used for urban planning [Carvalho Jr 2014]. Detecting such inconsistencies requires comparing transactions against genuinely comparable properties, which relies on property typology—most notably the number of bedrooms—since units of different typologies occupy distinct price segments within the same neighbourhood, a core consensus in the hedonic pricing literature [Sirmans et al. 2005].

The primary challenge is the complete absence of property typology in ITBI records: the municipal database stores only the declared value, cadastral identifier, and registered built area. Listing portals hold these structural attributes, but exact cross-referencing via addresses is notoriously unreliable due to formatting and encoding inconsistencies across administrative databases [Elmagarmid et al. 2007b], and supervised models cannot be trained directly on ITBI records due to the absence of target labels. Moreover, listing portals report announced prices while ITBI records reflect declared

transaction values—a 66% median gap in price-per-m^2 that causes listing-trained models to misclassify most tax records as smaller unit types. Inferring typology is therefore a cross-domain inference problem under heavy covariate shift.

This work infers bedroom counts for ITBI records from labelled listing portal data by framing the transfer as a source-domain feature engineering problem. Rather than aligning datasets through cross-dataset optimisation, raw prices are replaced with an intra-neighbourhood relative price—the ratio of a record's price-per-m^2 to its local median, computed independently within each dataset—which corrects the domain shift via a closed-form transformation while preserving the typological signal.

The contributions of this work are fourfold: (1) we quantify the domain shift, demonstrating that raw price is inadequate for cross-domain inference; (2) we introduce a neighbourhood-normalised relative price feature that aligns the distributions independently of target labels; (3) we evaluate two contrasting approaches, k-Nearest Neighbours (k=15) for maximising accuracy and a density-based pipeline (UMAP and HDBSCAN) for interpretable clusters with per-record uncertainty; and (4) we validate robustness through a cross-portal transferability experiment on OLX listings and an empirical application on municipal transactions from Goiânia, Brazil.

2. Related Work

Hedonic pricing models form the foundation for property valuation by capturing the contribution of structural, locational, and neighbourhood characteristics to market prices [Malpezzi 2003]. Machine learning has improved predictive accuracy in this domain, with Random Forests outperforming classical hedonic regression [Antipov and Pokryshevskaya 2012] and gradient-boosted trees serving as the standard for automated valuation models [Levantesi and Piscopo 2020]; recent variants incorporate spatial embeddings and image features [You et al. 2017]. These methods assume structural attributes are observed during training. In our setting, structural attributes are available in the source domain (listings) but absent from the target domain (tax records).

Dataset shift occurs when the joint distribution of features and labels differs between training and deployment [Quiñonero-Candela et al. 2009]. In covariate shift, the marginal distribution of inputs changes while $P(y \mid \mathbf{x})$ remains constant—the case here, since price-per-m^2 carries consistent typological information but differs in absolute scale due to the listing-versus-transaction gap. Theoretical bounds show target-domain error depends on source error and a divergence term between distributions [Ben-David et al. 2010]; our normalisation aims to minimise this divergence. When target labels are unavailable, adaptation relies on source-side feature engineering [Pan and Yang 2010], whereas deep adaptation methods such as feature alignment [Long et al. 2015], adversarial domain confusion [Ganin and Lempitsky 2015], and optimal transport [Courty et al. 2017] require explicit cross-dataset optimisation or differentiable architectures incompatible with an unlabelled, low-dimensional fiscal setting requiring interpretability. The proposed intra-neighbourhood relative price instead achieves covariate-shift correction via a closed-form, dataset-independent transformation.

Machine learning applied to real estate and municipal tax data has recently gained traction in Brazilian scientific venues: studies at ENIAC and BRACIS have applied Random Forests and deep learning ensembles to property price prediction from on-

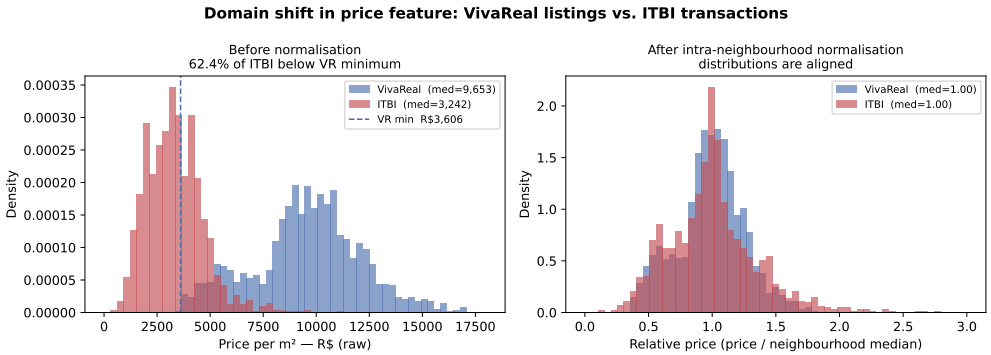


Figure 1. Domain shift in price-per-m^2 between VivaReal listings (blue) and ITBI transactions (red). *Left:* raw distributions – 62.4% of ITBI values fall below the VR minimum (dashed line). *Right:* after intra-neighbourhood normalisation, the distributions align.

line listings [Afonso 2019, Azevedo and Santos Filho 2025], while public-administration work has profiled IPTU taxpayers [Alves et al. 2023] and automated cadastral updating [Carneiro and Centeno 2003]. These approaches either assume structural attributes are fully observed during training or focus on taxpayer behaviour rather than property characteristics. The specific challenge of inferring property typology for ITBI records under severe domain shift remains unexplored, motivating the unlabelled transfer framework proposed here.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{L} = \{(\mathbf{x}_i^L, y_i)\}_{i=1}^{N_L}$ be the set of real estate listings, where $\mathbf{x}_i^L \in \mathbb{R}^d$ is a feature vector and $y_i \in \{1, 2, 3, 4+\}$ is the bedroom count. Let $\mathcal{T} = \{\mathbf{x}_j^T\}_{j=1}^{N_T}$ be the set of tax records, containing the same d features but *without* bedroom labels. The goal is to compute, for each tax record j, a probability distribution $P(y = c \mid \mathbf{x}_j^T)$ over bedroom categories $c \in \{1, 2, 3, 4+\}$.

3.2. Domain Shift Analysis

A fundamental challenge in applying listing-trained models to ITBI records is the systematic difference in price levels: listing portals report *announced* market prices, while ITBI records reflect *declared transaction* values, typically lower due to negotiated discounts and tax minimisation incentives. Figure 1 quantifies this shift: the VivaReal median is R$ 9,653/m^2 against R$ 3,242/m^2 for ITBI (a 66% gap), and 62.4% of ITBI records fall entirely below the minimum listing price in the training data. As a result, the raw price-per-m^2 of most ITBI records projects onto the low-price region of the listing feature space, where source-domain labels correspond predominantly to smaller unit types, yielding a distorted bedroom-count distribution inconsistent with the target data's area profile.

3.3. Feature Engineering and Intra-Neighbourhood Normalisation

Both datasets share two comparable structural features: built area in m^2 (a; `area_util` in listings, `QTDE_AREA_EDIFICADA_IPT` in tax records) and relative price (p_{rel}), the price per m^2 normalised by the neighbourhood median within each dataset separately:

$$p_{\text{rel},i} = \frac{p_{m^2,i}}{\text{median}_{b(i)}\{p_{m^2}\}} \tag{1}$$

where $b(i)$ denotes the neighbourhood of record i, drawn from the official cadastral administrative field (`bairro_sefin_nome` in listings and `NOME_BAIRRO_BAI_IPT` in tax records). The median is computed over all records of the same dataset in that neighbourhood, using the city-wide median as a fallback for neighbourhoods with fewer than 10 observations. This mitigates absolute price differences while preserving typological signals: a 1-bedroom unit commands a price premium relative to its local area, regardless of the neighbourhood's overall cost.

Figure 1 (right) shows that this normalisation aligns the distributions (p_{rel} median ≈ 1.0). A two-sample Kolmogorov–Smirnov test confirms this: raw distributions yield $\text{KS} = 0.876$ ($p < 10^{-300}$), while after transformation the statistic drops to $\text{KS} = 0.073$. Two feature configurations are evaluated throughout: the Raw Price vector ($[a,\ p_{m^2}]$) as baseline, and the Neighbourhood-Normalised vector ($[a,\ p_{\text{rel}}]$) with the proposed correction.

Temporal robustness. VivaReal listings were collected in April 2026, while ITBI transactions are from January–February 2025, a 14-month gap. The intra-neighbourhood relative price is invariant to uniform price shifts over this period by construction, since any such shift affects numerator and denominator equally and cancels in the ratio. The transferred signal thus depends on the *relative pricing composition* within each neighbourhood, a quantity empirically more stable over time than absolute price levels.

3.4. Scale Alignment via Quantile Normalisation

After computing p_{rel}, a `QuantileTransformer` is fitted on the *training listings* and all ITBI records (unlabelled target) to produce a uniform $[0, 1]$ output:

$$\tilde{x} = \hat{F}_{X_L^{\text{train}} \cup X_T}(x) \tag{2}$$

where $\hat{F}$ is the empirical CDF of the combined training–target sample, so each feature value is replaced by its percentile rank within that combined distribution, placing area and relative price on the same $[0, 1]$ scale. The held-out test listings are *not* included in the transformer fit; they are transformed using the already-fitted CDF, avoiding any leakage from the test partition.

3.5. UMAP Embedding

UMAP [McInnes et al. 2018] is fitted on the normalised listing features ($n_neighbors = 80$, $min_dist = 0.033$, Euclidean metric; hyperparameters selected by Bayesian optimisation *within the training listings only*, via an internal 80/20 validation split, Section 4). Although the input space is already two-dimensional ($[a,\ p_{\text{rel}}]$), UMAP learns a non-linear topology-preserving manifold in which bedroom typologies form tighter, more separable clusters than under Euclidean distance in the original axes, and its `transform` operation projects out-of-sample ITBI records into the source-trained embedding without refitting—essential given that no target labels exist for a second fitting pass.

3.6. Density-Based Clustering and Cluster Labelling

HDBSCAN [Campello et al. 2013] is applied to the listing embedding. For each cluster γ, the empirical bedroom distribution is computed from its member listings:

$$P(y = c \mid \gamma) = \frac{\sum_{i:\ell_i=\gamma} \mathbf{1}[y_i = c]}{|\gamma|} \quad c \in \{1, 2, 3, 4+\} \tag{3}$$

High-purity clusters (one category dominating) yield confident predictions; mixed clusters reflect genuine ambiguity in the economic signature of that latent region.

3.7. Probabilistic Typology Inference

For each tax record j, `approximate_predict` returns a predicted cluster $\hat{\gamma}_j$, and the typology probability is read from the cluster profile. Records assigned to noise ($\hat{\gamma}_j = -1$) receive a uniform prior over the four bedroom categories. The final prediction is $\hat{y}_j = \arg\max_c P(y = c \mid \hat{\gamma}_j)$, and the predictive entropy

$$H_j = -\sum_c P(y = c \mid \hat{\gamma}_j) \log P(y = c \mid \hat{\gamma}_j) \tag{4}$$

serves as an uncertainty signal, with high entropy indicating a cluster where no single bedroom category dominates.

4. Experimental Setup

4.1. Datasets

A total of 8,515 apartment listings were collected from the VivaReal portal for Goiânia, GO, in April 2026. After filtering for records with valid area, price, and bedroom count (1–5), and removing outliers at the 1st and 99th percentiles of area and price per m^2, 7,495 listings remained, with bedroom distribution: 1-bedroom 11%, 2-bedroom 23%, 3-bedroom 50%, 4-bedroom or more 16%.

A total of 3,761 ITBI transaction records were obtained from the municipal tax authority of Goiânia for January and February 2025. After filtering for apartment-type transactions with positive built area and declared value of at least R$ 50,000, 1,542 records remained, spanning 39 distinct neighbourhoods (Table 1).

To assess portability across listing portals, 6,882 apartment listings were collected from OLX Imóveis for Goiânia, carrying the same structural attributes. These records concentrate in Setor Bueno (4,329) and Setor Marista (2,109), and are used exclusively for out-of-domain evaluation (Section 5.4), playing no role in training or hyperparameter selection. A two-sample KS test on p_{rel} yields $\text{KS}_{\text{VR–OLX}} = 0.091$, versus $\text{KS}_{\text{VR–ITBI}} = 0.876$, confirming that listing portals share similar relative-price distributions and that the VR–ITBI shift is not an artefact of the VivaReal portal itself.

4.2. Evaluation Protocol

Since no bedroom labels exist on the ITBI side, all methods are evaluated via a self-supervised protocol on the source domain: VivaReal is partitioned into 80% training and 20% test, stratified by bedroom count (1,499 test records). The test split

Table 1. Dataset summary.

	VivaReal	ITBI
Total records (after filtering)	7,495	1,542
Bedroom labels available	Yes	No
Neighbourhoods covered	8†	39
Features used	a, p_{rel}	a, p_{rel}
Median area (m^2)	107	126
Median price/m^2 (R$)	9,653	3,242
Median p_{rel}	1.00	1.01

†VivaReal listings concentrate in 8 high-activity neighbourhoods (Setor Bueno, Marista, etc.). For the 31 ITBI neighbourhoods with no VivaReal listings, p_{rel} uses the city-wide median as fallback. The per-neighbourhood KS analysis (21 neighbourhoods) covers only those with ≥ 20 ITBI records.

is held out entirely, with its labels used exclusively for final performance reporting. UMAP and HDBSCAN hyperparameters are selected by Bayesian optimisation (Optuna [Akiba et al. 2019], TPE sampler, 300 trials) using an internal 80/20 split *within the training listings only*; the held-out test partition plays no role in tuning, so reported test accuracies are not optimistically biased. Balanced accuracy is used as the optimisation objective to avoid bias towards the dominant 3-bedroom class.

Following standard practice [Quiñonero-Candela et al. 2009], the `QuantileTransformer` is fit on the union of *training* listings and all unlabelled ITBI records so the empirical CDF covers the full feature range of both domains; since target labels are never observed, this does not constitute label leakage. KNN accuracy is identical (93.4%) in a purely inductive ablation (transformer fit on training listings only), confirming no evaluation bias; HDBSCAN drops 4.3 pp inductively, reflecting that the UMAP embedding benefits from a joint distribution covering both domains.

4.3. Methods Compared

Six configurations isolate the effects of the domain shift correction and the choice of algorithm:

- **Majority Class:** Heuristic baseline always predicting the modal 3-bedroom category.
- **Area Threshold:** Deterministic baseline using built-area quantile rules ($<60\text{m}^2 \rightarrow 1$, 60–90 $\rightarrow 2$, 90–140 $\rightarrow 3$, $\geq 140 \rightarrow 4+$).
- **KNN (Raw Price):** Distance-weighted k-Nearest Neighbours (k=15) on quantile-normalised raw features; the uncorrected algorithmic baseline.
- **HDBSCAN (Raw Price):** UMAP + HDBSCAN on raw features without hyperparameter tuning, capturing the baseline latent topology.
- **KNN (Neighbourhood-Normalised):** The proposed supervised approach on the domain-corrected feature space; k=15 verified via sensitivity analysis (variance <1 pp for $k \in \{5, 10, 15, 20, 30\}$).
- **HDBSCAN (Neighbourhood-Normalised + Optuna):** The complete proposed unsupervised pipeline, with domain-corrected features and tuned hyperparameters.

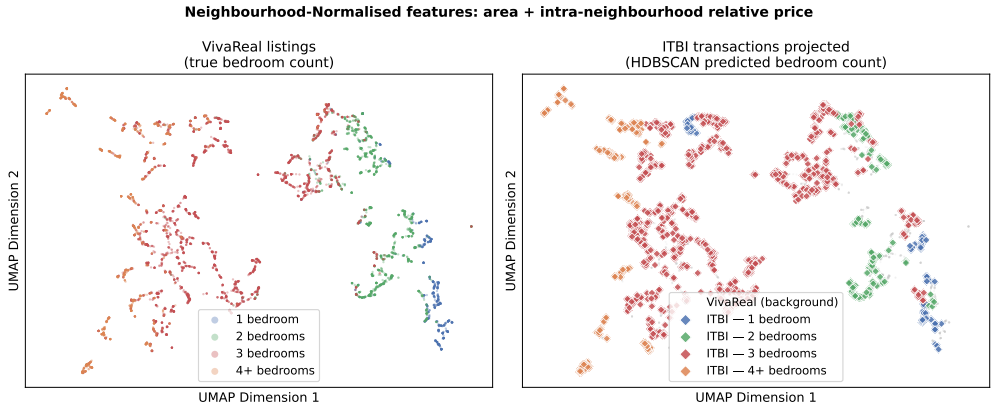


Figure 2. UMAP latent space with Neighbourhood-Normalised features. *Left:* VivaReal listings coloured by true bedroom count. *Right:* ITBI transactions (diamonds) projected into the same space, coloured by HDBSCAN predicted bedroom count; VivaReal shown in grey for reference.

Finally, to assess cross-portal generalisation, the VivaReal-trained KNN model is applied without retraining to the held-out OLX dataset under both feature variants.

5. Results

5.1. Clustering Quality

Applied to the 7,495 listing embeddings with Neighbourhood-Normalised features and optimised hyperparameters, HDBSCAN identified 54 clusters with 449 noise points (6.0%). Internal cluster validity metrics on the 7,046 non-noise points confirm good structure: Silhouette coefficient $= 0.546$ and Davies–Bouldin index $= 0.623$. Of the 54 clusters, 30 achieve bedroom-class purity ≥ 0.85, confirming that the normalised features separate typologies effectively in the latent space. Figure 2 shows the embedding: the left panel colours VivaReal listings by true bedroom count, and the right panel overlays ITBI projections coloured by predicted bedroom count.

5.2. Typology Prediction Accuracy

Table 2 reports classification metrics on the 1,499-record VivaReal test set, using Balanced Accuracy and Macro F1 to account for class imbalance (3-bedroom units are half of the source dataset). The results reveal a systematic class collapse when raw-feature models are applied to the target domain: both KNN and HDBSCAN assign near-zero 1-bedroom predictions to ITBI records (0.4% and 0.0%, respectively), since the target's lower median price forces most transactions into the low-price region of the source space, a zone monopolised by larger typologies in training. The Neighbourhood-Normalised transformation resolves this: by converting absolute price into an intra-neighbourhood relative ratio, the models recover the local premium characterising compact units, with KNN and HDBSCAN recovering 8.0% and 6.1% of the 1-bedroom share, aligning with the 10.4% reference from the deterministic Area Threshold. Normalised pricing also improves in-domain generalisation, raising KNN accuracy to 0.934.

The top-performing algorithms, KNN (0.934) and Random Forest (0.932), are local, non-parametric models performing within 0.2 pp of each other, indicating that

Table 2. Classification results on the VivaReal 20% test set (1,499 records). Best values per column in bold.

Method	Acc.	Bal. Acc.	Macro F1	Weighted F1	$\hat{p}_{1\text{BR}}^{\text{ITBI}}$ (%)*
Majority Class	0.496	0.250	0.166	0.329	–
Area Threshold	0.701	0.769	0.711	0.702	10.4
KNN (Raw Price)	0.922	0.906	0.913	0.922	0.4
HDBSCAN (Raw Price)	0.773	0.739	0.757	0.772	0.0
KNN (Neighbourhood-Normalised)	**0.934**	**0.921**	**0.927**	**0.933**	8.0
Logistic Regression (NN)	0.829	0.785	0.806	0.827	5.6
LinearSVC (NN)	0.771	0.667	0.696	0.743	4.5
Random Forest (NN)	0.932	0.921	0.927	0.932	8.9
HDBSCAN (Neighbourhood-Normalised + Optuna)	0.809	0.760	0.785	0.807	6.1

*Predicted 1-bedroom share (%) among the 1,542 ITBI target-domain records; not a classification metric on the source-domain test set.

typology is dictated by the localised joint position of area and relative price. Linear classifiers—LinearSVC (0.771) and Logistic Regression (0.829)—fail because a single hyperplane cannot cleanly separate these interleaved classes: premium 1-bedroom units often show low area but high relative price, while luxury 4-bedroom units show both high area and high relative price. The convergence of KNN and Random Forest near 93% suggests the residual error stems from physically overlapping layouts, not classifier capacity.

The 12.5 pp gap between the supervised ceiling (KNN) and the unsupervised pipeline (HDBSCAN) represents the cost of operating without target label supervision: HDBSCAN clusters by density on the economic manifold rather than optimising explicit classification boundaries, so it cannot artificially sharpen boundaries between adjacent typologies. Its Macro F1 of 0.785 reflects this, with dominant errors between physically adjacent categories (e.g., 18% of true 2-bedroom units predicted as 3-bedroom)—genuine structural ambiguities that no model can resolve without floor-plan attributes.

5.3. Inference on ITBI Records

Table 3 presents the predicted bedroom distribution and uncertainty metrics for the ten largest ITBI neighbourhoods, showing clear spatial patterns: Faz. Santa Rita is homogeneous (83.7% 3-bedroom, $H_{med} = 0.42$); Lot. Água Azul is the only neighbourhood with a 1-bedroom plurality (41.2%), consistent with a compact, lower-cost district; and Set. Central shows the highest uncertainty ($U_{>1.2} = 11.1\%$), reflecting mixed-use heterogeneity.

Table 3. Predicted bedroom distribution (HDBSCAN, Neighbourhood-Normalised + Optuna) for selected ITBI neighbourhoods. H_{med}: median predictive entropy. $U_{>1.2}$: share of records with entropy > 1.2.

Neighbourhood	N	1BR%	2BR%	3BR%	4BR+%	H_{med}	$U_{>1.2}$%
SET BUENO	369	1.9	6.5	64.5	27.1	0.53	10.0
FAZ SANTA RITA	49	6.1	8.2	83.7	2.0	0.42	4.1
SET CENTRAL	45	11.1	20.0	57.8	11.1	0.50	11.1
LOT AGUA AZUL	34	41.2	11.8	47.1	0.0	0.69	0.0

Across all 1,542 ITBI records, median predictive entropy is 0.562 and 9.5% exceed 1.2 (near-uniform; $\ln 4 \approx 1.39$). The source-domain median under the same model is 0.522, a gap of only 0.040 nats, confirming that Neighbourhood-Normalised bridges the distributional shift: the model is as confident on ITBI as on its training domain. For tax analysts, entropy < 0.80 indicates reliable predictions and entropy > 1.2 flags records for manual review. The macro ECE [Guo et al. 2017] is 0.026, confirming well-calibrated cluster-profile probabilities without post-hoc adjustment.

5.4. Transferability and Robustness

Table 4 reports KNN accuracy when the VivaReal-trained model is applied to the OLX dataset with no retraining. Raw features achieve 84.4% accuracy (balanced 0.847); Neighbourhood-Normalised yields 81.4% (balanced 0.807). This 3.0 pp drop confirms that normalisation is neutral when the shift is already small ($\mathrm{KS} = 0.091$), unlike the VR–ITBI setting ($\mathrm{KS} = 0.876$), where the same normalisation rescues the model from near-zero minority-class accuracy: *normalisation is beneficial precisely when the domain shift is large.*

Table 4. Cross-portal KNN accuracy: VivaReal (train) → OLX (test, zero-shot). $n = 6{,}882$.

Feature variant	Accuracy	Balanced Acc.
Raw Price	0.844	0.847
Neighbourhood-Normalised	0.814	0.807

Ablation tests confirm the necessity of local neighbourhood structure: replacing per-neighbourhood normalisation with a city-wide global median reduces KNN accuracy from 0.934 to 0.923 (balanced accuracy 0.906 vs 0.921), and a z-score of raw price yields identical results to the city-wide median. Varying the minimum-neighbourhood-size threshold from 5 to 30 observations alters accuracy by at most 0.003. KNN achieves a macro ECE of 0.013, competitive with the HDBSCAN pipeline.

6. Discussion

The intra-neighbourhood relative price is the core mechanism enabling cross-domain transfer. Raw prices collapse the target minority classes because the median price gap forces most tax records into the low-price region of the source space; the relative price metric restores a plausible typology distribution and improves in-domain accuracy, scaling in value with the magnitude of the domain shift while remaining neutral across already-aligned distributions.

A structural accuracy gap exists between supervised models like KNN (93.4%) and unsupervised pipelines like HDBSCAN (80.9%). Because HDBSCAN derives typology post-hoc from cluster composition without label supervision, this gap reflects the cost of obtaining interpretable market segments and uncertainty estimates. Given the restriction to two available features, the performance ceiling is set by feature constraints rather than tuning deficiencies. The methods are complementary: KNN delivers optimal point predictions, while HDBSCAN provides calibrated probabilistic flags for fiscal auditing and manual review triage.

7. Limitations

The primary limitation is the inability to validate directly on the target domain: due to the absence of target bedroom labels, performance is supported only by indirect consistency checks such as aligned entropy distributions and stable minority-class proportions, so an anomalous sub-population in the tax data could go undetected. Geographic coverage is also restricted by the source domain; neighbourhoods lacking listing data rely on a city-wide median fallback, weakening the local discriminative signal.

The minimal feature space limits separability between adjacent typologies (e.g., 2- versus 3-bedroom). Finally, the relative price assumption relies on uniform under-declaration practices; if under-declaration varies systematically by typology, the relative ordering could be distorted. This study is confined to apartment-type transactions in a single municipality, leaving generalisation to other housing types or cadastral conventions untested.

8. Conclusion

Domain shift is the primary obstacle to transferring listing-trained classifiers to unlabelled municipal tax records. Replacing raw prices with an intra-neighbourhood relative price resolves this shift without cross-dataset calibration: KNN achieves 93.4% accuracy on a held-out test set, correcting the minority-class collapse produced by raw-price models, while HDBSCAN achieves 80.9% with calibrated uncertainty estimates (macro ECE = 0.026). A zero-shot cross-portal experiment on 6,882 external listings confirms the normalisation's benefit under large domain shifts. Applied to 1,542 tax transactions in Goiânia, Brazil, the pipeline successfully infers property typologies and flags uncertain records for manual review using only two universally available features [Elmagarmid et al. 2007a].

Future work will focus on cross-referencing identifiers with official cadastral records to supply verified labels for direct target-domain evaluation, incorporating additional listing portals to extend source coverage, and replicating the methodology in municipalities with different property dynamics.

8.1. Acknowledgments

This work was supported by the Centro de Excelência em Inteligência Artificial (CEIA) and by the Prefeitura de Goiânia. This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR), grant number 408490/2024-1.

References

Afonso, B. K. D. A. (2019). Housing prices prediction with a deep learning and random forest ensemble. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, RS, Brasil. SBC.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Alves, E. L. G., Viana, L. Z. B., Santos, F., Franco, W., and Pinheiro, V. (2023). Leveraging property tax collection through an unsupervised model for taxpayer qualification. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, RS, Brasil. SBC.

Antipov, E. and Pokryshevskaya, E. (2012). Mass appraisal of residential apartments: an application of Random Forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2):2772–2778.

Azevedo, A. C. and Santos Filho, R. C. d. (2025). Predictive modeling of real estate prices using machine learning. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, RS, Brasil. SBC.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. (2010). A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Brasil (1988). Constituição da república federativa do Brasil de 1988, Art. 156, Inciso II. Diário Oficial da União.

Campello, R., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, volume 7819 of *LNAI*, pages 160–172, Heidelberg. Springer.

Carneiro, A. and Centeno, J. (2003). Atualização cadastral multifinalitária utilizando imagens de alta resolução e SIG. *Revista Brasileira de Cartografia*, 55(1):1–10.

Carvalho Jr, P. H. B. d. (2014). Distributive aspects of real estate property and its taxation among brazilian families. Texto para discussão, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília, Brasil.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Elmagarmid, A., Ipeirotis, P., and Verykios, V. (2007a). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007b). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1180–1189.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330. PMLR.

Levantesi, S. and Piscopo, G. (2020). The importance of interest rates in building residential real estate price predictions with machine learning: the case of Italy. *Risks*, 8(3):92.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 97–105.

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, pages 67–89.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Pan, S. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N., editors (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.

Sirmans, G. S., Macpherson, D. A., and Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1):1–44.

You, Q., Pang, R., Cao, L., and Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, 19(12):2751–2759.