

Using machine learning to aggregate apartment prices: comparing the performance of different Luxembourg indices

Bob Kaempff^a; **David Kremer**^a

^aCentral Bank of Luxembourg

International Conference on Real Estate Statistics
Tokyo 20 February 2026

The views expressed are those of the authors and may not be shared by other research staff or policymakers in the BCL or the Eurosystem. Contact: david.kremer@bcl.lu

Data on dwellings and price changes are “critical ingredients for understanding household wealth, its evolution over time, and for the vulnerability of households’ financial position”

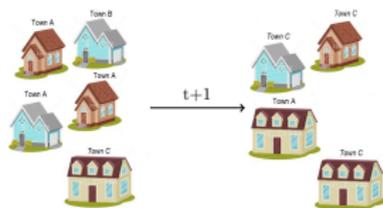
Financial Stability Board (2009) “The financial crisis and information gaps” Report to the G-20 Finance Ministers and Central Bank Governors.

- ▶ Need for reliable real estate data widely recognised.
- ▶ Housing cycle affects economy, price stability and financial stability.

Aims of the paper

- ▶ Compare **characteristics (reliability, volatility and coherence) of several house price indices (HPIs)** in Luxembourg.
- ▶ Apply **artificial intelligence (AI) techniques** to HPIs.

- ▶ Highly heterogeneous and infrequent real estate transactions. \Rightarrow changes in composition and methods complicate price comparisons over time.
- ▶ Main difficulties include:
 - ▶ Choice of method to calculate a HPI
 - ▶ Coverage by region
 - ▶ Coverage by type of housing
 - ▶ Data source
- ▶ Mean/median prices do not adjust for changes in quality.
- ▶ More problematic than homogenous products in HCPI (bananas, potatoes, diesel,...)



- ▶ Individual apartment transactions between 2007Q1 and 2025Q3 extracted from notary deeds collected by the Tax Authority (AED)
- ▶ Complemented with information regarding the living surface and annexes from the Land Registry and Topography Administration
- ▶ Initial data processing to harmonize treatment of new apartment sales and eliminate irrelevant transactions

	Hedonic			Repeat Sales
	<i>All</i>	<i>New</i>	<i>Existing</i>	
Number of transactions	78,684	26,436	52,248	33,284
Average price per m2 (thousand euros)	5,805	6,189	5,610	5,562
Average living surface (m2)	81.1	81.7	81.1	80.8

Figure 1: Summary statistics for the cleaned hedonic and repeat sales samples

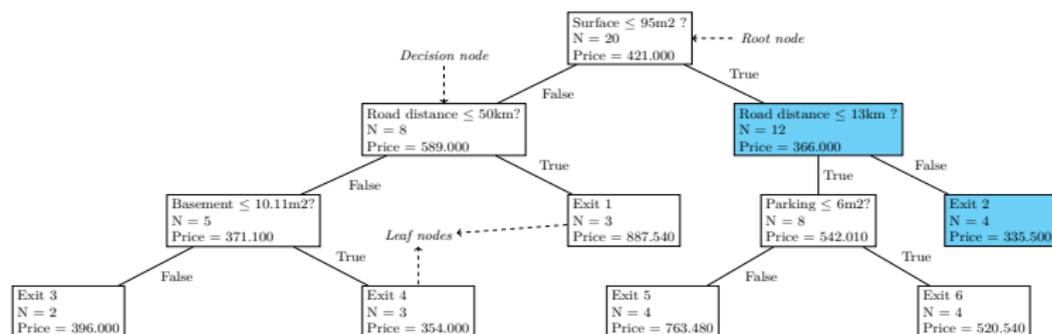
Sources: AED, Own calculations

- ▶ Hedonic indices (e.g. *Eurostat (EA Reference Index)*)
 - ▶ Estimated relationship between the price and characteristics of a dwelling ⇒ **Quality adjusted index.**
- ▶ Repeat sales indices (e.g. *S&P Corelogic Case-Shiller*)
 - ▶ Assume a given dwelling's characteristics unchanged between two different sales. ⇒ **No need for quality adjustment.**
 - ▶ Price changes for a given dwelling reflect general price changes for all dwellings
- ▶ **New approach based on Krause (2019):** Indices based on artificial intelligence
 - ▶ Random Forest (RF) algorithm builds **Residential Property Price Index (RPPI)**.
 - ▶ Popular algorithm in other domains of macroeconomic analysis (e.g. Goulet Coulombe (2024))

Random Forests for RPPI: Motivation & Principles (1)

- ▶ A *random forest* (RF) model is an ensemble of regression trees¹
- ▶ Individual regression trees:
 - ▶ Capture nonlinearities and interactions with no parametric assumptions
 - ▶ Repeatedly split data into smaller, similar groups (*node splitting*)
 - ▶ Produce predictions by following the path to a final (leaf) node
 - ▶ Example: Apartment with 80m² of living surface, located 50km from the capital ⇒ 335.500€

Figure 2: Decision tree - Illustration



(Displayed prices correspond to average prices in each node.)

¹We use the 'ranger' package in R (Wright & Ziegler, 2017), a fast implementation of Breiman (2001).

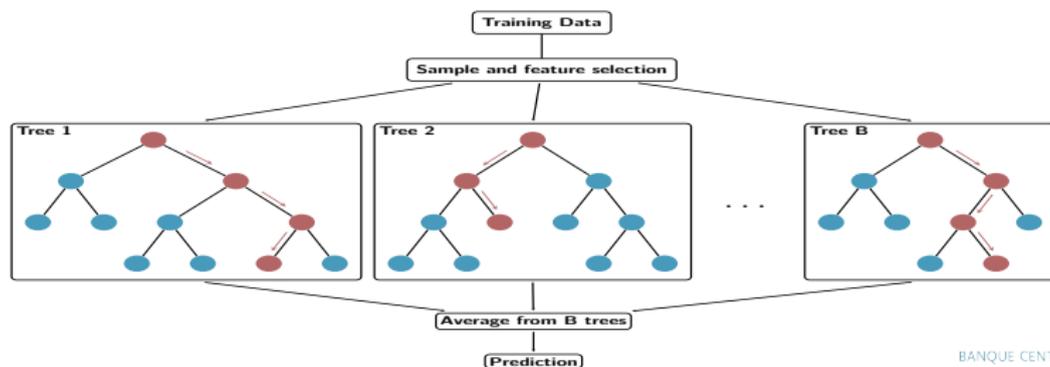
Random Forests for RPPI: Motivation & Principles (2)

- ▶ Decision trees: High variance and risk of overfitting
- ▶ Random forests (Breiman, 2001) mitigate these issues through:
 - ▶ **Bootstrap aggregation (bagging)**: build B trees on bootstrap samples to reduce variance
 - ▶ **Feature bagging**: at each node, randomly select a subset of predictors
- ▶ Prediction for a continuous target:

$$\hat{f}_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(x_b)$$

where x_b is the bootstrapped and feature-sampled variant of x .

Figure 4: Simplified random forest



Counterfactual methods Provide interpretability by comparing predicted values from machine learning models using hypothetical data points (counterfactuals) across a given input/feature while holding all other inputs/features constant.

Partial Dependence Plots (PDP).

- ▶ Introduced by Friedman (2001) to interpret ML models
- ▶ Show *marginal effect* of feature x_t on predicted outcomes by averaging predictions over all observations:

$$\text{PDP}(x_t = j) = \frac{1}{N} \sum_{i=1}^N \hat{f}_{\text{bag}}(x_t = j, x_{i,c})$$

- ▶ In RPPI context: effect of transaction period on predicted apartment prices.
- ▶ Assumption: Variable of interest (x_t) uncorrelated with other predictors → Avoid unrealistic counterfactual regions and misleading marginal effects.

- ▶ Reveal heterogeneity masked by PDP averages (Goldstein, Kapelner, Bleich, & Pitkin, 2015)
- ▶ For each observation, hold characteristics $x_{i,c}$ fixed and vary x_t across every observed value.
- ▶ One curve per observation:

$$\text{ICE}_i(x_t) = \hat{f}_{\text{bag}}(x_t, x_{i,c})$$

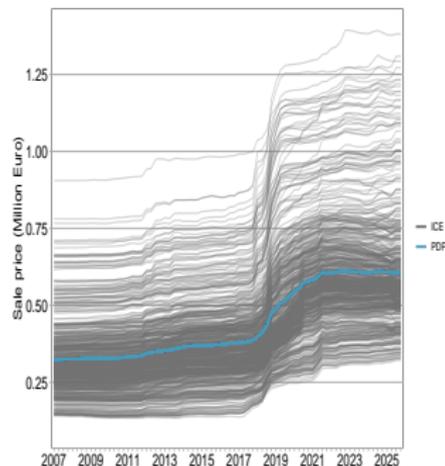
- ▶ ICE simulates “selling the same apartment in every time period”.

Link between ICE and PDP.

- ▶ PDP is the pointwise average of ICE curves:

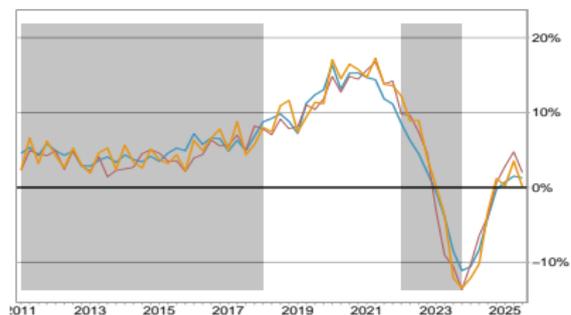
$$\text{PDP}(x_t) = \frac{1}{N} \sum_{i=1}^N \text{ICE}_i(x_t)$$

Figure 5: ICEs and PDP



PDPs can be slow to calculate and computationally expensive. For example, this figure consists of 528 transactions \times 76 periods = 40,128 predictions to construct the ICEs, which have each been derived from a random forest composed of 500 individual trees

Figure 7: Existing apartments



Note: ■ Random forest index; ■ Hedonic index; ■ Repeat sales index.

Main results

- ▶ Random forest HPI (■) closely tracks hedonic and repeated sales indices ⇒ **Viable alternative**.
- ▶ Random forest HPI is more stable and therefore **easier to interpret** (removes measurement errors that may arise from small estimation samples, cf. New apartments Hedonic Index).
- ▶ Similar trends in three quality-adjusted indices ⇒ **Less uncertainty than relying on a single index**.

Sources of Revisions

- ▶ **Repeat Sales:** Each resale provides information on both past and current transaction prices.
- ▶ **Random Forest:** Inherently unstable due to random node splits.
- ▶ **Hedonic:** Regression coefficients re-estimated with each release.

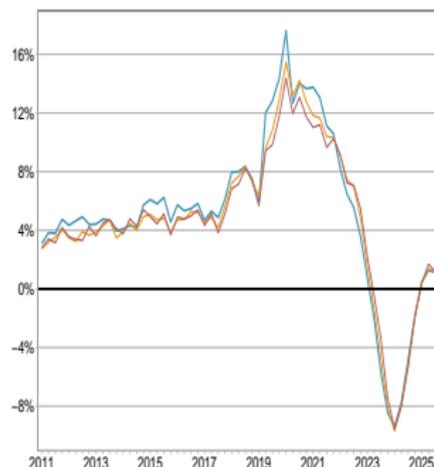
Framework

- ▶ Re-estimate indices for each quarter to obtain *real-time*, *quasi-real* (transmitted with up to 1-quarter delay), and *final* series.
- ▶ Revisions reflect both new transactions and model-induced adjustments (Orphanides & Norden, 2002).

Empirical Results

- ▶ **Hedonic index:** Minimal revisions once delayed transactions are included.
- ▶ **Random Forest:** Largest and most persistent revisions; especially in high-growth periods.

Figure 8: Illustration of index revision - Random forest



Note: Year-on-year growth for real-time (■), quasi-real (■) and the final estimate (■).

Objective: Evaluate how well RPPIs predict individual apartment price changes *out-of-sample*, addressing aggregation bias due to heterogeneity.

Method (K-fold CV, repeat-sales based):

- 1 Randomly split full transaction sample into $K = 5$ folds.
- 2 Estimate each index on 4 folds (training).
- 3 Select all repeat-sales in the held-out fold (validation).
- 4 Predict second-sale price:

$$\hat{p}_{i,t} = p_{i,\tau} \times \frac{ind_t}{ind_\tau}$$

- 5 Compute errors: $e_{i,t} = \hat{p}_{i,t} - p_{i,t}$ and MAPE

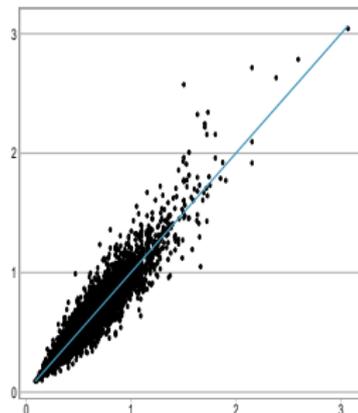
$$MAPE = \frac{1}{N} \sum_i \left| \frac{e_i}{p_{i,t}} \right|.$$

- 6 Repeat for all 5 folds.

Key Findings:

- ▶ No method significantly outperforms others
- ▶ MAPE: Hedonic 9.96%, RF 10.02%, Repeat-Sales 9.96%.

Figure 10: Out-of-sample predictions - Price level (Mio. EUR) (Random forest)



Note: X-axis corresponds to observed values and Y-axis to the out-of-sample predictions. The blue line represents the linear fit.

- ▶ All three methods (hedonic, repeat sales, random forest) identify similar trends at the aggregate level.
- ▶ The random forest index closely tracks traditional indices, supporting its validity.
- ▶ Improvements could come from enhanced data quality, particularly additional dwelling characteristics (e.g. information on major renovations).
- ▶ Consulting multiple indices can reduce uncertainty in property price developments.

Thank you for your attention.

- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.
- Goulet Coulombe, P. (2024). The macroeconomy as a random forest. *Journal of Applied Econometrics*, 39(3), 401–421.
- Krause, A. (2019). A Machine Learning Approach to House Price Indexes. *Working Paper, Zillow Group*.
- Orphanides, A., & Norden, S. v. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, 84(4), 569–583.
- Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

Appendices

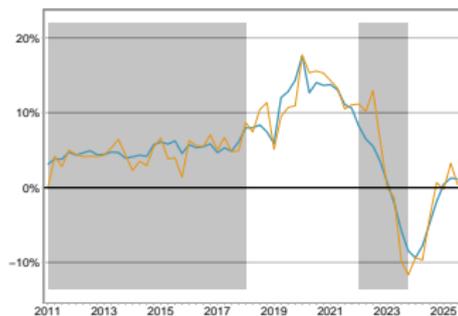
		2011-2023	2011-2017	2018-2021	2022-2023
Overall	Hedonic	6.3%	4.5%	11.7%	2.2%
	Random forest	6.4%	4.8%	11.5%	1.9%
Existing	Hedonic	6.4%	4.6%	12.5%	0.6%
	Random forest	6.4%	4.8%	12.1%	-0.1%
	Repeat sales	5.9%	4.2%	12.1%	-0.7%
New	Hedonic	6.2%	4.4%	10.6%	3.9%
	Random forest	6.7%	5.0%	10.7%	4.8%

Figure 12: Apartment prices, average year-on-year growth (%)

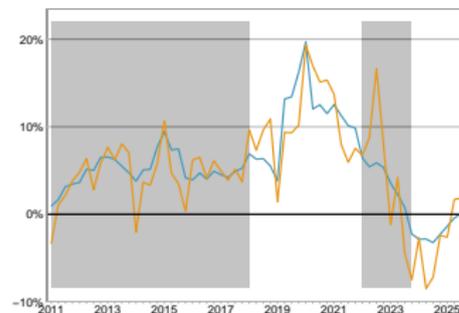
Note: By construction, the repeat sales index excludes new apartments, so we report results along with those of the hedonic index and the random forest index for existing apartments.

Estimated annual price changes per index (%)

(a) Overall

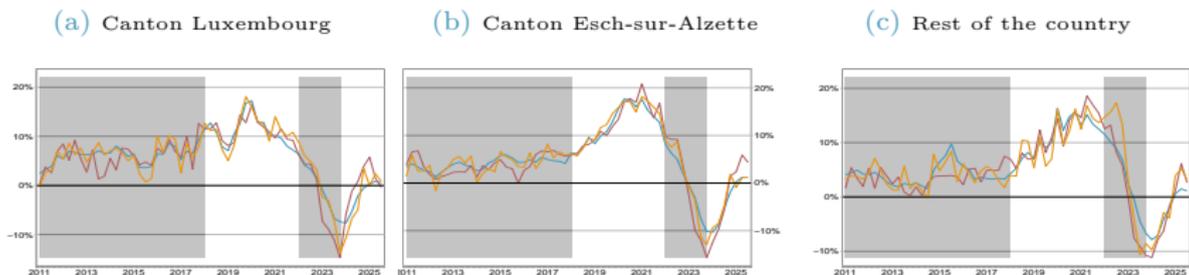


(b) New apartments



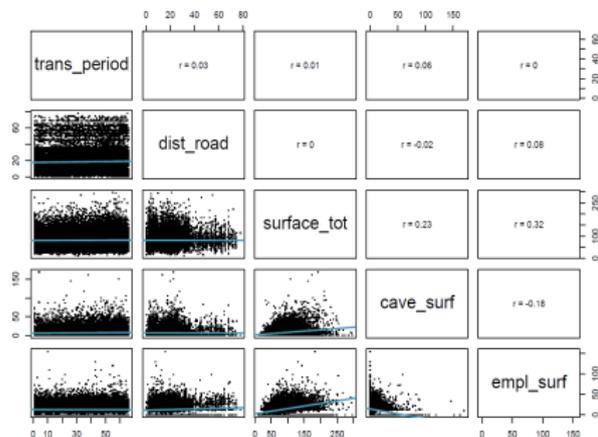
Note: ■ Random forest index; ■ Hedonic index

Figure 14: Annual changes by region (%)

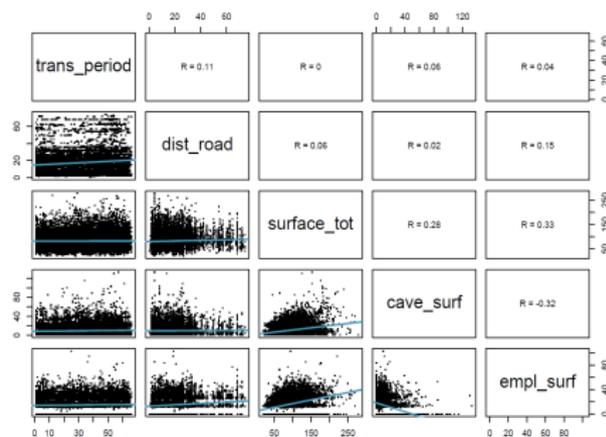


Note: ■ Random forest index; ■ Hedonic index; ■ Repeat sales index.

(a) Existing apartments



(b) New apartments



Valid PDPs require that there is no strong correlation between the variable of interest and the other explanatory variables, hence we are interested in the first row and first column of the plots. The upper triangle provides the Pearson correlation coefficients between the variables and the lower triangle shows the fitted line from a simple linear regression.

Key Hyperparameters:

- ▶ **min.node.size**: minimum observations required to split a node
- ▶ **mtry**: number of splitting variables randomly selected at each node
- ▶ **num.trees**: number of trees in the forest (controls stability vs. computation time)

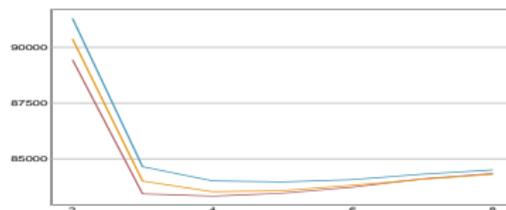
Tuning Procedure:

- ▶ Grid search over *mtry* and *min.node.size* $\in \{5, 15, 25\}$
- ▶ 3-fold cross-validation: train on 2 folds, validate on 1; repeat for all folds and hyperparameter combos

Results

- ▶ Optimal model: *mtry* = 4, *min.node.size* = 5
- ▶ No strong guidance for *num.trees*: larger forests usually improve stability but increase computation

Figure 17: Hyperparameter finetuning



Note: Each line plots the *mtry* value (*x*-axis) against the average RMSE (*y*-axis) obtained from 3-fold CV. The different colors indicate the three minimum node sizes selected to train the models: ■ 5; ■ 15 and ■ 25