CAHIER D'ÉTUDES
WORKING PAPER

N° 194

USING MACHINE LEARNING TO AGGREGATE APARTMENT PRICES: COMPARING THE PERFORMANCE OF DIFFERENT LUXEMBOURG INDICES

BOB KAEMPFF AND DAVID KREMER

JANUARY 2025

BANQUE CENTRALE DU LUXEMBOURG
EUROSYSTÈME

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

Bob Kaempff
David Kremer

Abstract

This paper presents three different methods to estimate an apartment price index for Luxembourg and evaluates their performance by comparing volatility, proneness to revisions, coherence and out-of-sample fit. In addition to the standard hedonic and repeat sales methods, we apply a machine learning algorithm (the interpretable random forest approach) to produce a new index for Luxembourg. The three methods indicate similar trends in residential property prices. The new random forest index closely tracks the two more traditional indices, providing evidence supporting the viability of this new approach. In comparing the three methods, the random forest index is more stable and therefore provides information that is easier to interpret. However, all three methods are subject to revisions when new observations are released and these tend to be larger for the random forest than for traditional indices.

JEL Codes: C40, C53, R31
Keywords: Residential property price index, hedonic model, repeat sales model, machine learning, random forest algorithm

February 10, 2025. Banque centrale du Luxembourg, Département Économie et Recherche, 2 boulevard Royal, L-2983 Luxembourg (contact: bob.kaempff@bcl.lu, david.kremer@bcl.lu). *For useful comments and suggestions, we thank Patrick Fève, Paolo Guarda, and BCL colleagues. This paper should not be reported as representing the views of the BCL or the Eurosystem. 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.*

RÉSUMÉ NON TECHNIQUE

Au Luxembourg, l'outil principal pour mesurer l'évolution des prix des logements est un indice trimestriel calculé par l'Institut national de la statistique et des études économiques (STATEC) sur base des transactions immobilières contenues dans les extraits des actes notariés. Cet indice se base sur une méthode dite "hédonique" qui consiste à établir une relation statistique entre le prix d'un bien et les caractéristiques de ce bien, permettant ainsi d'éliminer les effets liés à la composition changeante de l'échantillon. Or, comme tout autre outil statistique, la méthode hédonique se base sur différentes hypothèses sous-jacentes et présente un certain nombre de faiblesses.

Nous appliquons deux techniques alternatives pour mesurer l'évolution des prix de vente des appartements au Luxembourg à travers le temps. Une première technique souvent utilisée est la méthode des "ventes répétées" qui se base sur la comparaison du prix d'un même objet immobilier qui a été vendu à plusieurs moments. D'abord, nous estimons que depuis 2007, les ventes répétées concernent 44 % des transactions immobilières au Luxembourg portant sur des appartements. Ensuite, nous appliquons cette méthode pour créer un indice alternatif des prix des appartements.

Au délà des méthodes conventionnelles, l'intelligence artificielle (IA) et les techniques d'apprentissage automatique fournissent de nouvelles méthodes pour estimer les indices de prix. Bien que ces modèles sont souvent critiqués comme des "boîtes noires", des nouvelles approches permettent de les rendre plus interprétables. Ainsi, ce cahier applique un simple modèle d'apprentissage automatique aux transactions immobilières au Luxembourg pour créer un nouvel indice des prix des appartements.

Après avoir détaillé la methodologie de ces trois indices, nous évaluons la performance de chaque indice selon sa volatilité, la taille de ses révisions, sa cohérence avec les autres indices et ses indications hors échantillon. Nos principaux résultats sont les suivants : les trois approches produisent des indices avec des tendances similaires dans le temps, ce qui réduit l'incertitude liée à l'interprétation d'un seul indice de prix. En particulier, les trois méthodes confirment la rapide progression des prix entre 2018 et 2021. De même, nos résultats confirment le fort ralentissement en 2022 ainsi que la baisse des prix en 2023. L'indice basé sur des techniques d'apprentissage automatique suit de près les indices plus traditionnels, ce qui confirme empiriquement la validité d'une telle approche.

Le choix final de la méthode dépend souvent des besoins de l'utilisateur. Alors que l'indice basé sur des méthodes d'apprentissage automatique est considérablement moins volatil et donc plus facile à interpréter, il est par contre sujet à des révisions plus importantes que celles des deux autres méthodes conventionnelles.

NON-TECHNICAL SUMMARY

In Luxembourg, the main indicator for residential property price developments is a quarterly index calculated by the National Institute of statistics and economic studies (STATEC) based on property transactions extracted from notary deeds. The index is based on the so-called "hedonic" method, which consists of establishing a statistical relationship between the price of a dwelling and its characteristics, thus eliminating effects linked to the changing composition of the sample. However, like any other statistical tool, the hedonic method is based on different underlying assumptions and has a number of weaknesses.

We apply two other methods to measure the evolution of apartment prices over time. One commonly used alternative is the so-called "repeat sales" method, which focuses on changes in the sale price of properties that have been sold more than once. First, we estimate that since 2007, repeated sales have concerned 44 % of transactions in Luxembourg involving apartments. We then apply this method to create an alternative apartment price index.

In addition to these conventional methods, artificial intelligence (AI) and machine learning techniques provide new methods for estimating price indices. Although ML-based models are often criticized as black boxes, there are different tools to help interpret their output. In this paper, we apply a "random forest" algorithm to Luxembourg property transactions to generate an apartment price index from individual transactions in Luxembourg.

After detailing these three methods, this paper evaluates their performance according to volatility, proneness to revisions, coherence and out-of-sample fit.

Our main results are the following: all three methods identify similar trends over time, which reduces the uncertainty associated with the interpretation of any single price index. Rapid growth in apartment prices, especially from 2018 to 2021, is confirmed by all three methods. Similarly, all three methods confirm the sharp slowdown in 2022 and the decline in apartment prices during 2023. The random forest index closely follows the traditional indices, which empirically confirms the validity of this approach.

The final choice of the method often depends on the user's needs. While the random forest index is considerably less volatile and therefore easier to interpret, it is however subject to revisions as new observations are released and these revisions tend to be larger than those of the traditional indices.

1 Introduction

The need for reliable data on residential property prices is widely recognised by international institutions. In the midst of the Financial Crisis, the 2009 Report to the G-20 Finance Ministers and Central Bank Governors (Financial Stability Board, 2009) described data on dwellings and the associated price changes over time as "*critical ingredients for understanding household wealth, its evolution over time, and for the vulnerability of households' financial position*".

Understanding the drivers of a housing cycle and its implications for the broader economy is a key focus of attention for policy makers (Scatigna et al., 2014). That is because downturns in residential property prices have been a catalyst for major economic crises[1]. In practice, Residential Property Price Indices (RPPIs) are often used in conjunction with other macroeconomic statistics to monitor the broader economy. RPPIs have a number of other important uses, apart from serving as a measure of price changes. In its technical handbook, Eurostat (2013) identifies various potential use cases, including:

- A macro-economic indicator of economic growth.
- An input into estimating the value of housing as a component of wealth.
- A financial stability indicator.
- A means for within-country and international comparisons.

These many use cases emphasize the need for viable indices. However, methodological differences can undermine comparisons across countries, regions or time (Silver, 2012).

Property transactions are infrequent and apply to highly heterogeneous objects, so that changes in compilation methods or coverage can affect RPPIs. For example, the COVID-19 pandemic illustrated how geographical coverage can be an important feature. As the opportunity costs associated with commuting time decreased with the deployment of work-from-home arrangements, housing demand may have shifted away from city centres towards rural areas (Roma, 2021). Such changing preferences could lead to diverging price dynamics in different regions, which are not captured by a national index.

[1] Crowe, Dell'Ariccia, Igan, and Rabanal (2013) find that, of the 46 systemic banking crises for which data was available, more than two-thirds were preceded by house price boom-bust patterns. Mortgage booms have also been identified as a critical driver of financial instability, playing a central role in the recurrence of economic crises (Jordà et al., 2015, 2016).

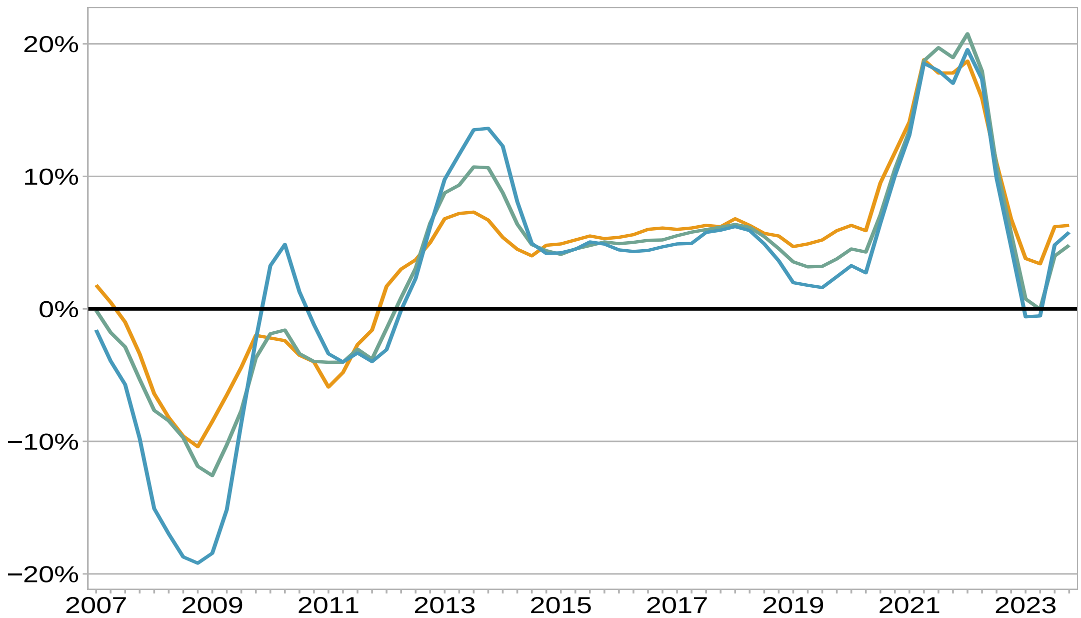
Some of the key choices in RPPI design are:

- **Adjustment method for changes in quality.**
- **Coverage by geographical region** (national, regional, urban centers,...).
- **Coverage by type of housing** (row or detached houses, apartments,...).
- **Data source** (real estate agents, online websites, land registry or notary office,...).

The first three items are particularly relevant for this paper. As a result of the different choices in the design of RPPIs, there are often several indices available for a given country. In general, these are published by national statistical institutes, with some complementary RPPIs published by private data providers and/or other public institutions. Because measurement difficulties can prevent a single index from covering all requirements, users often have to make compromises (e.g. RPPIs may only be available at the national level).

Using three different RPPIs available in the US, Figure 1 shows that, despite differences in methodology and data, there are clear similarities in the trend and timing of turning points. However, it also shows that there are differences among RPPIs and that these differences can vary significantly from quarter to quarter.

Figure 1: Examples of available Residential property price indices in the US



Note: S&P CaseShiller National; S&P CaseShiller 10 City Composite; Federal Housing Finance Agency. Data retrieved from Bloomberg Terminal. Sources: Standard & Poors, Federal Housing Finance Agency.

In Luxembourg, the National Statistical institute (STATEC) publishes quarterly statistics on acquisition prices for dwellings. These include the number of transactions, average prices per transaction, average prices per square meter and so-called hedonic price indices that correct for differences in housing characteristics. The price indices are broken down into different categories, distinguishing between existing houses, existing apartments and newly built apartments, allowing a comparison of price dynamics in these market segments.

In this paper we reproduce the hedonic index from STATEC and apply two additional methods to measure apartment prices in Luxembourg. First, we apply a so-called "repeat sales" method, which assesses how valuations change over time by focusing on the difference in sale price of the same apartment across sales at different moments in time. Second, we apply a machine learning (ML) technique, the so-called "random forest" approach, to generate apartment price indices for the different market segments in Luxembourg.

The remainder of this paper is organised as follows. Section 2 presents the methodology of the hedonic index, the repeat sales index and the random forest index. Section 3 describes the data. Section 4 reports the main results, broken down by type of index and region and evaluates the empirical performance of the apartment price indices by comparing interpretability, revisions and accuracy, meaning the ability to predict out-of-sample transaction prices. Section 5 concludes.

2 Methodologies

The simplest approach to build a price index would be to use the mean or median price across transactions, typically a month or a quarter. This would only require information on individual transactions (price and date) and the resulting index would capture a general trend in transactions over time.

However, as many authors have noted[2], average or median price indices can give a misleading impression of price trends over a given period. That is because the set of dwellings being sold each period will have different characteristics, which changes the composition of the sample from which average or median prices are calculated each period. For example, a period when transactions are mainly in high-end dwellings could be followed by a period when transactions are predominantly in low-end dwellings, creating a false impression that average or median prices declined, including for dwellings that were not sold. In the following, we present three methods to produce more reliable residential property price indices.

2.1 Hedonic methods

The hedonic approach is based on an estimated relationship between the price of a dwelling and several of its characteristics, such as its living surface, its geographical

[2] See for example Smith et al. (1988); Englund et al. (1999); Jansen et al. (2008).

location or the number of parking spaces[3]. A hedonic RPPI therefore accounts for the fact that the dwellings sold in a given period will have different characteristics from those sold in another. In this paper, we follow the methodology used by STATEC (see Lamboray (2010)), estimating the following "hedonic function", that links the transaction price p of a property i to the following k characteristics of i (X):

$$ln(p_i) = \alpha + \sum_{k=1}^{K} \beta_k X_{k,i} + \varepsilon_i \tag{1}$$

- The surface of the dwelling.
- The square root of the surface of the dwelling.
- The type of dwelling (Newly built/Existing)
- Dummy indicating whether the dwelling is sold with a cellar
- Dummy indicating whether the dwelling is sold with a parking space
- The distance between the dwelling municipality and Luxembourg City

As in Lamboray (2010) we use a hedonic imputation approach to construct our hedonic index. More specifically, we run separate hedonic regressions for a base period and for the current period and use the estimated regression parameters to form imputed measures of constant quality price changes. We thus reproduce the methodology used for the official RPPI in Luxembourg to provide a benchmark for our results[4].

Although the hedonic method is often used to built RPPIs, it requires choosing the explanatory variables to include and the specification of the functional form (F. T. Wang & Zorn, 1997). These choices are not obvious in practice. For example, the rapid deployment of work-from-home arrangements during the COVID-19 pandemic may have affected the additional value represented by a balcony or a garden. Although this information is not available at the level of individual transactions in Luxembourg, it seems plausible that such features would contribute positively to the price of the dwelling.

2.2 Repeat sales methods

The repeat sales approach assumes that changes in the price of a given property when it is sold at different points in time (repeat sales) reflect the change in the general price level for all dwellings. This implies that the characteristics of the property have remained unchanged between sales. This approach avoids the hedonic regression requirement to choose a set of characteristics that are observable for all dwellings or even to specify the functional form linking prices to observed characteristics (Eurostat, 2013).

[3] In its handbook, Eurostat (2013) recognizes the hedonic approach as the "standard" one.

[4] There are some minor differences between the official RPPI and our internal estimates, particularly for new apartments, mainly due to differences in the approach to clean raw data prior to estimation.

By construction, the repeat sales method also controls for location at the finest level of detail, while other methods are often unable to identify the exact location. For instance, data available in Luxembourg only indicates the municipality where the dwelling is located.

In practice, the repeat-sales model uses ordinary least squares to estimate a relationship between the price progression (logarithm of the ratio between the first and second sale price of each property) and a set of dummy variables taking the value -1 in the first sale period (τ), 1 in the second sale period (t) and 0 for all other periods until the end of the sample (S). Thus, the regression equation is:

$$ln\left(\frac{p_{i,t}}{p_{i,\tau}}\right) = \sum_{s=0}^{S} \Theta_s D_{i,s} + \varepsilon_{i,t} \textit{ with } (0 < \tau < t < S) \tag{2}$$

where $D_{i,s}$ denotes the dummy variable identifying object i in period s and $\varepsilon_{i,t}$ is the error term with zero mean and constant variance σ^2. The repeat sales index from period 0 to period S can be obtained by exponentiating the corresponding regression coefficients Θ_s

One drawback of the repeat sales method is that property that is sold at two different points in time may no longer have the same characteristics due to renovations or changes in the surroundings. Therefore, the longer the time span between two sales, the less plausible the constant-quality assumption underlying this approach. We relax this assumption by applying the weighted repeat sales approach introduced by Case and Shiller (1987), putting less weight on dwellings sold after long time intervals. This approach is discussed in more detail in Kaempff and Kremer (2021).

2.3 Random forest algorithm

In addition to these two "traditional" methods to construct RPPIs, this paper applies an index based on a relatively new approach to modeling residential property prices using machine learning (ML) algorithms. While several authors proposed ML solutions to residential property price estimation (see for instance C. Wang and Wu (2018); Fan et al. (2006); Baldominos et al. (2018)), their application to RPPIs remains limited. This section describes the random forest algorithm[5], a common ML technique and follows a methodology proposed by Krause (2019).

Random forest is an ensemble technique that combines simple models ("building blocks"), in this case regression trees, to obtain more stable and powerful statistical learning models (James et al., 2013). Over time, random forest became a standard non-parametric regression tool, which can use different types of predictor variables without making any prior assumption on their association with the dependent variable.

Regression trees are easy to interpret and can be used to approximate nonlinear relationships. However, they suffer from high variance, which makes them unstable. High variance means that the structure of a single tree (and hence its predictions) may

[5]In this paper,we use the 'ranger' package in R (Wright & Ziegler, 2017), which is a fast implementation of the methodology proposed by Breiman (2001).

change considerably after small changes in the training data[6]. Regression trees, especially if grown very large, tend to fit very well in-sample but do not perform as well when predicting from a new set of observations, a phenomenon known as overfitting.

The random forest therefore works by *bagging* (**b**ootstrap **agg**regation) multiple trees. Bagging is a popular method to reduce the variance of a statistical learning method and relies on the fact that for any set of k independent observations $n_1, ..., n_k$ with variance σ, the variance of the sample mean $\overline{n}$ is given by $\frac{\sigma}{k}$ (James et al., 2013). Since we lack independent training datasets, bootstrapping randomly selects subsamples (bootstrap samples) from our dataset.

A second important feature of the random forest is that the algorithm does not take into account all possible predictors in each tree. Instead, at each node in the tree, it randomly selects a subset of predictors, or so-called features, and chooses the best node splits from this subset (feature bagging)[7]. For our random forest, we use the following set of predictors X:

- The period of the transaction.
- The surface of the dwelling.
- The distance between the dwelling municipality and Luxembourg City
- Dummy indicating whether the dwelling is sold with a cellar
- Dummy indicating whether the dwelling is sold with a parking space
- The total surface of cellar(s)
- The total surface of parking space(s)

An individual tree "grows" by randomly selecting the subset of predictors at each node, choosing the best split and then passing on to the next two nodes. A set of trees grown on the same dataset makes up a random forest. For a continuous outcome variable, the final prediction is made by averaging the predictions from individual trees, which are themselves obtained by averaging the observations grouped in the final node (leaf) that the tree assigns to the inputted observation. We draw B separate bootstrap samples (with replacement) , grow a tree for each bootstrap sample and then obtain an aggregate prediction by averaging the predictions of all B trees:

$$\hat{f}_{bag}(x_b) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x_b) \tag{3}$$

where $\hat{f}^b(x_b)$ denotes the predicted outcome from the b^{th} tree which has randomly selected a subset x_b of the predictors available in X. Thus, the random forest extends simple bagging by not only randomly selecting subsets of the training data but also randomly selecting subsets of the predictors.

[6] Using ML terminology, regression models are "trained" rather than "estimated" and the estimation sample is referred to as the training sample.

[7] See appendix A.2 for more details on node splitting.

The basic steps for growing our random forest composed of B trees are as follows:

Random forest algorithm

1. Draw a bootstrap sample b with n random observations (transaction price with the complete set of characteristics x_b) from the data set with N observations[8].

2. For subset x_b drawn in Step 1, one individual tree is constructed by:

   - randomly selecting, at each node[9], a subset from the complete set of predictors (features) X[10].

   - splitting the observations in the given node, on the basis of the previously selected subset, into two new nodes and continue the process to split each new node until splitting the newest node would create nodes below the minimum node size s (see appendix A.1).

3. Inputting a new observation into the root node of the tree from Step 2 will generate an output $\hat{f}^b(x_b)$ by moving down the branches of the tree until it reaches the corresponding leaf (final node), where the prediction is given by the average value of the output variable across the observations in that leaf.

4. Steps 1 to 3 are repeated until B regression trees are created, making up a forest. For the entire forest, the final prediction $\hat{f}_{bag}(x_b)$ is computed by averaging the outputs from input x_b obtained from the B trees (equation 3).

The parameters of the random forest algorithm are set by K-Fold cross validation (See Appendix A.3 for more details):
- number of regression trees B = 500
- number of predictors f = 4
- minimum node size s = 5

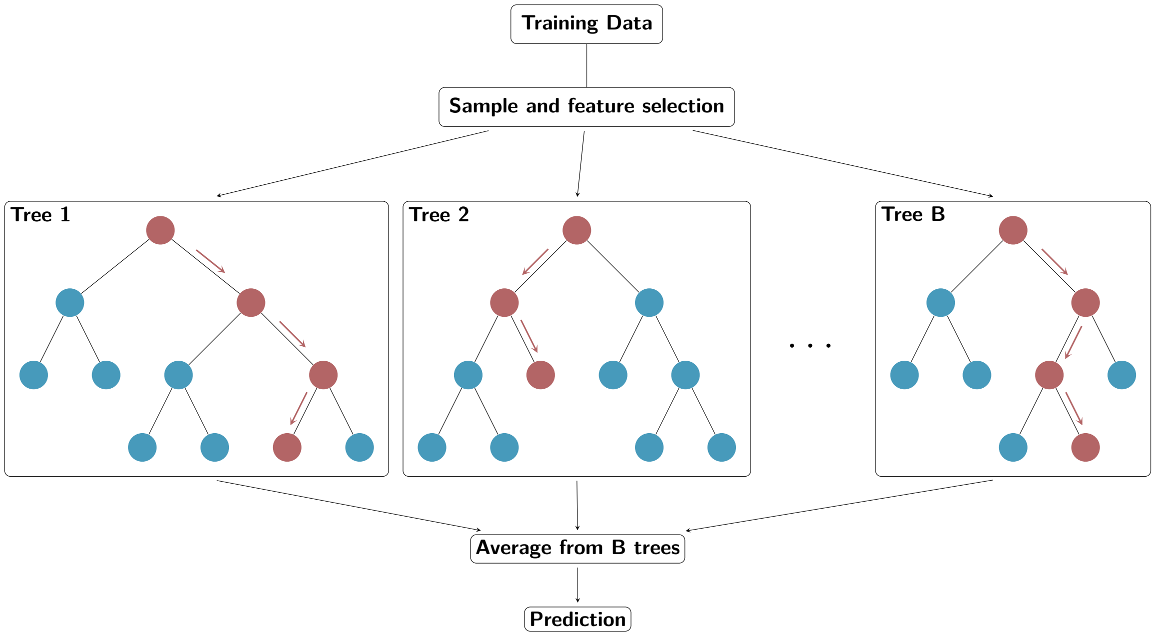
.

[8]By randomly drawing observations with replacement from the data set, the bootstrap sampling algorithm selects $1-(1-\frac{1}{N})^N$ of the observations from the data set. For large data sets, this approximately equals 62.3 % of the N observations in the data set.

[9]Using ML terminology, a node is a sub-sample of data and an associated decision rule.

[10]In practice, we constrain the algorithm to use the transaction period as predictor for each node, so that f predictors are randomly selected in addition to the transaction period.

Figure 2: Simplified random forest - Illustration



Although the random forest is a widely used algorithm, it is often criticized as a "black-box model" for producing predictions that are difficult to interpret (Ribeiro et al., 2016), This is because, random forests do not create coefficient estimates (like more traditional methods), which could help to understand which predictors are driving a particular outcome.

2.3.1 Interpretability methods

Model interpretability methods can be applied to black-box models to improve predictions from machine learning[11]. In this paper, interpretability is defined as information that helps the user determine how small changes to a given input affect the model output (Slack et al., 2019).

Counterfactual methods provide interpretability by comparing predicted values from ML models using hypothetical data points (counterfactuals) across a given input/feature while holding all other inputs/features constant. Partial dependence plots (PDP) and individual conditional expectations (ICE) (Goldstein et al., 2015) are common examples of counterfactual methods.

- **Partial dependence plots**

[11] Breiman (2001) explicitly points out that "*a forest of trees is impenetrable as far as simple interpretations of its mechanism go. In some applications, (...) it is critical to understand the interaction of variables that is providing the predictive accuracy.*"

Partial Dependence Plots (PDP) were introduced by Friedman (2001) to interpret the results of ML algorithms, including the random forest. They visualise the marginal effect of a selected input variable on the response variable. For our RPPIs, the objective is to predict the impact of time on apartment prices.

Conceptually, a partial dependence function sets the feature of interest (transaction period) to the same value for each observation and uses a trained algorithm to find predictions (transaction price) over all the observations in the data set. These predictions are then averaged across the entire data set to get the partial dependence value for that period in time.

Formally, the partial dependence function for a given value j of a feature x_t based on the trained model $\hat{f}_{bag}$ can be approximated as follows:

$$\widehat{PDP}(x_t) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{bag}(x_t, x_c^i \mid x_t = j) \tag{4}$$

In the context of our RPPIs, x_t identifies a transaction period from the dataset and x_c^i contains the values of all other characteristics (predictors) for a given observation i. To construct a PDP, x_c^i is set to the same value for all observations and (4) is calculated for each time period t in the sample, after which the predicted value for that period is obtained by averaging predicted values over all the N observations.

PDPs visualise the process described above, by plotting every value of x_t (time periods) against the average predicted value (from Equation 4) across all observations assuming they occurred in the given time period. The resulting plot shows the average value (transaction price) as a function of time (see Figure 3 for an illustration of a PDP)[12]. Since the PDP considers every observation in the dataset and produces a single average prediction as a function of the variable of interest, it is referred to as a "global interpretation method".

One key assumption underlying PDPs is that the dwelling characteristics in x_c are not correlated with the time of sale x_t. If this assumption is violated, the partial dependence plots will include regions of the data that are very unlikely to be observed in practice [13]. In our example, this assumption would be violated if regulations set a maximum dwelling surface until a certain point in time, or if dwellings could not be sold with cellars or parking spaces until a certain date, or if the maximum distance from Luxembourg City was suddenly extended by new construction in distant areas that were previously uninhabited. For our dataset, the correlation between the transaction period and dwelling characteristics (predictors) is negligible [14].

[12]PDPs can be slow to calculate and computationally expensive. For an input variable x_t with T points in the range $[x_t^{min}, x_t^{max}]$ and N data points to evaluate the PDP, there are $T * N$ evaluations of the fitted model $\hat{f}_{bag}$, which in turn consists of B regression trees, to construct the PDP. For example, Figure 3 consists of 468 transactions * 68 periods = 31,824 predictions to construct the ICEs, which have each been derived from a random forest composed of 500 individual trees

[13]Suppose we were interested in the effect of height on a person's running speed. If we were to include age as an explanatory variable, we would run the risk of estimating speeds for individuals aged 10 years and measuring 1m95, which are unlikely to be observed.

[14]Annex A.4 provides the correlation matrices for the set of predictors.

- **Individual conditional expectations**

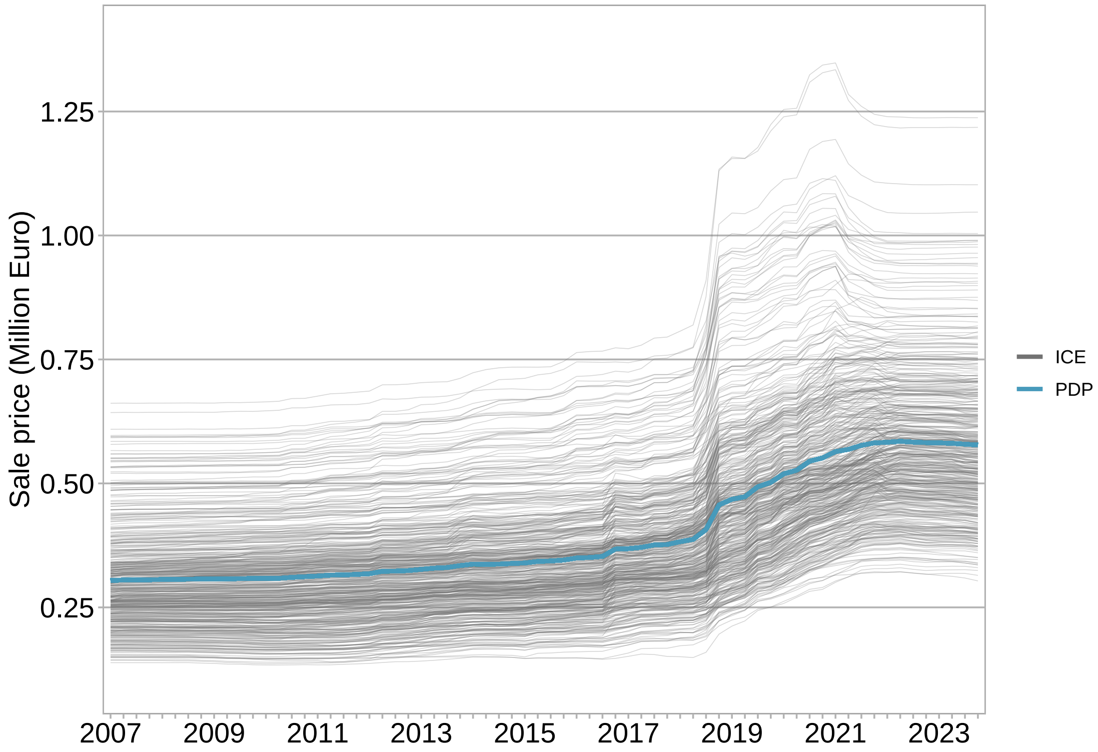
Since an average curve derived from the PDP can mask the complexity of the modeled relationship, Goldstein et al. (2015) proposed to refine the PDP by plotting the relationship between the predicted response and any predictor of interest for each individual observation, hence the name "local interpretation method".

Specifically, individual conditional expectations (ICE) plot the predicted outcome (transaction price) for each observation as a function of time conditional on dwelling characteristics of the given observation.

For each observation, an ICE curve can be plotted by varying its date across all possible dates in the sample while holding its other characteristics constant. In other words, ICE's simulate the value of a given dwelling as if it had been sold once in every time period in the sample. Repeating this step for each observed transaction in the dataset and averaging across all predicted prices in each period gives the full PDP.

The ICE algorithm can thus provide insights into individual conditional relationships estimated by a random forest. In Figure 3, we condition on time. However, PDPs can condition on any characteristic that is not correlated with other predictors. For example, assuming the distance to Luxembourg city is not correlated with other dwelling characteristics, we can plot the predicted transaction prices of apartments based on their distance to the capital.

Figure 3: Illustration of individual conditional expectations and partial dependence plot



Note: Illustration of ICE's and PDP using 1 % of existing apartment transactions (468 observations). Vertical axis shows average predicted prices and horizontal axis time periods. Grey lines show the predicted prices for each individual apartment (ICE) and the blue line shows the average per quarter (PDP). Own calculations using the 'pdp' R package (Greenwell, 2017).

2.3.2 Constructing price indices using random forest algorithms

The random forest indices from this study will thus be derived from PDPs (Equation 4). Following the approach described in Lamboray (2010), we treat new and existing apartments separately, and construct two random forest algorithms obtaining separate PDPs for the two subsamples ($\widehat{PDP}^{new}$ and $\widehat{PDP}^{existing}$). The final index is calculated as the weighted average of the two sub-indices, where the weights in year T are the corresponding shares of the value of all transactions in the previous year T-1.

As shown in Figure 3 above, PDPs and ICEs in levels are expressed in euros at current prices. To make the random forest output comparable to the repeat sales index and the hedonic index, we convert the PDPs into an index using the same base year (2015) as the other indices.

3 Data sources

The indices are constructed based on information on property transactions extracted from notary deeds collected by the *Administration de l'enregistrement, des domaines et de la TVA* (AED). These are complemented with information regarding the living surface of apartments and annexes from the *Administration du Cadastre* (Land Registry). The initial sample used in this study includes 100,537 apartment transactions[15] registered by the AED between the first quarter of 2007 and the fourth quarter of 2023.

Initial data processing harmonizes the treatment of the sales of new apartments and eliminates irrelevant transactions[16]. This processing follows the common methodology developed by a national working group on real estate statistics composed of members of STATEC, the National Housing Observatory, the AED and the BCL[17]. The resulting sample, used for the compilation of the hedonic price index and the random forest index, includes 72,050 transactions (*"the hedonic sample"*).

Next, repeat sales involving the same apartment at different points in time are identified. Although the database does not contain an explicit identifier for the dwelling sold, this can be constructed for each apartment from the following information:

- The location of the apartment (municipality)
- The surface of the building plot
- The living surface of the apartment
- Textual description of the apartment[18]
- The total surface of parking space(s)
- The total surface of cellar(s)

If all these dimensions match for any two transactions at different points in time, then the second transaction is assumed to be a repeat sale of the same apartment. With this procedure, 31,746 transactions, or 44 % of the hedonic sample are identified as repeat sales, i.e. transactions involving apartments that were sold more than once between 2007 and 2023. Three additional treatments are performed to eliminate certain sales that could bias the index. First, we follow Jansen et al. (2008) by excluding repeat sales where the interval between sales is less than 12 months, as these generally feature very

[15]In this paper we focus on apartment transactions as the AED database does not contain sufficient information for houses (e.g. surface of dwelling) to calculate the indices described in Section 2.

[16]For example, office space transactions are excluded from the sample.

[17]For detailed descriptions of this processing, see Paccoud et al. (2024) and `https://statistiques.public.lu/dam-assets/fr/methodologie/methodes/economie-finances/Prix/prix-logements/note-prix-de-vente.pdf`

[18]The AED uses a textual description to identify the apartment withing the building (e.g. apartment n. ... 1st floot, block C). This description remains fixed for each resale.

high price gains, suggesting these dwellings were bought to renovate and resell. Second, we exclude dwellings that were sold five times or more between 2007 and 2023 since this suggests that they may have some hidden flaws. In a last step we trim the data by excluding repeat sales with very high average annual price increase between sales[19]. These cleaning steps are explained in more detail in Kaempff and Kremer (2021).

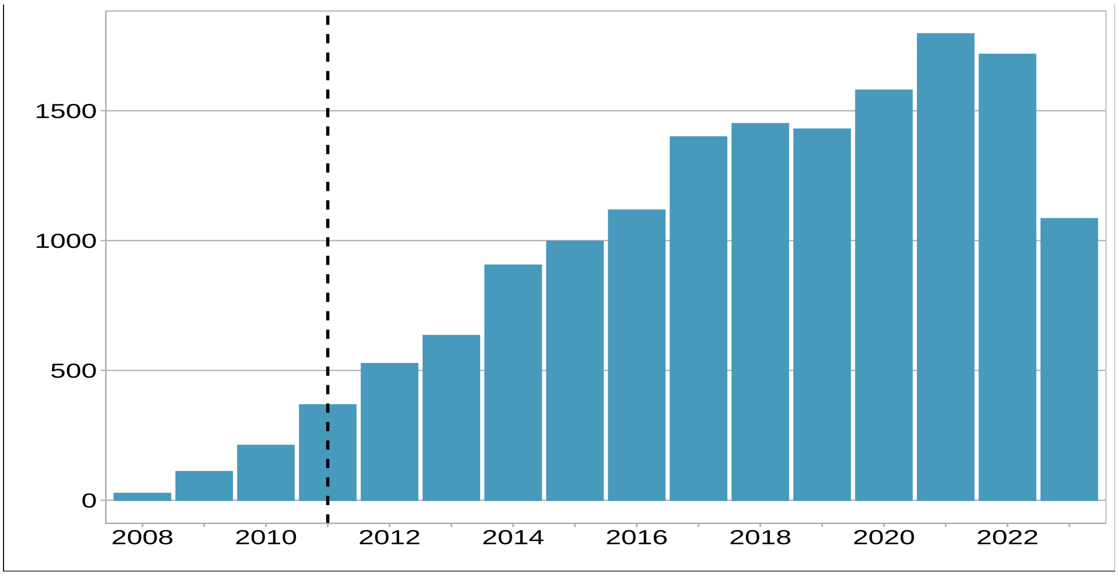
Table 1: Summary statistics for the hedonic and repeat sales samples

	Hedonic			**Repeat Sales**	**Repeat Sales**
	All	*New*	*Existing*	**(Initial)**	**(Clean)**
Number of transactions	72,050	25,461	46,589	31,746	27,594
Average price per m2 (thousand euros)	5,641	6,168	5,353	5,350	5,348
Average living surface (m2)	81.4	82.0	81.2	79.9	81.1

Sources: AED, Own calculations

The final sample used to calculate the repeat sales index includes 27,594 transactions, representing 38 % of the hedonic sample. Table 1 compares the number of transactions, the average price per square meter, and the average living surface for the different subsets of data. Naturally, there are fewer observations of repeat sales (on the right), but the average prices per square meter and average living surface are very similar to those of the hedonic sample for existing apartments.

Figure 4: Number of repeat sales per year



Sources: AED, Own calculations

[19] In a first step we stratify repeat sales by region (Canton Luxembourg and rest of the country) and by year of the latest sale. We then eliminate, per stratum, repeat sales where price gains exceed median price gains by 1.5 times the inter quartile range. Such price changes are likely to indicate apartments that underwent major renovations. A similar treatment is applied for the hedonic index (Lamboray (2010))

The analysis will focus on the period beginning in 2011 because there are obviously few to no repeat sales at the beginning of the sample. By construction, the number of repeat sales observed per year gradually increases over time as shown in Figure 4. Since the sample begins in 2007 and we require a minimum holding period of 12 months, the first year that we can observe a repeat sale is 2008. However, after data cleaning (see Section 3), only 30 repeat sales are observed in that year. We therefore focus on results starting in 2011, when there were roughly 370 repeat sales (92 per quarter). Despite the sharp drop in transactions in 2023 due to higher interest rates and price declines, we still observe enough repeat sales to consider the sample informative in that year.

4 Results

We not only differentiate between existing and newly built apartments but also divide the sample into three geographical regions. We distinguish an initial period of moderate price increases (2011-2017), a period of strong price increases (2018 - 2021) and a period characterized by weaker prices and fewer transactions (from 2022). After a visual comparison of the results, we aim to identify the advantages and disadvantages of each index by comparing volatility, proneness to revisions, coherence and out-of-sample fit.

4.1 National price developments

The upper part of Table 2 reports average annual growth rates for the hedonic index and for the random forest price index. Average growth rates for the whole sample are very closely aligned and vary between 6.4 % based on the random forest and 6.3 % based on the hedonic index. According to all indices, price increases were clearly higher between 2018 and 2021. Between 2011 and 2017, the highest average annual growth rate was 4.8 % and the lowest was 4.5 %. Between 2018 and 2021, these shift up to 11.7 % and 11.5 %. From 2022 to 2023, the two indices capture slower price growth, with the hedonic index averaging 2.2 % and the random forest index averaging 1.9 %. Standard T-tests for equality of means confirm that, for the complete sample as well as for the different subsamples, none of the differences in mean growth rates are statistically significant.

Table 2: Apartment prices, average year-on-year growth (%)

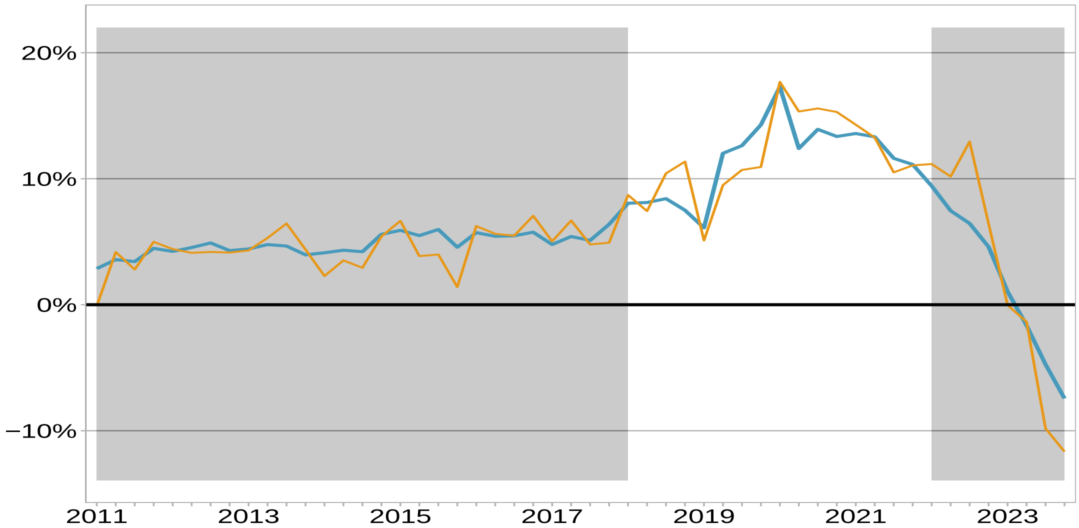
		2011-2023	2011-2017	2018-2021	2022-2023
Overall	Hedonic	6.3%	4.5%	11.7%	2.2%
	Random forest	6.4%	4.8%	11.5%	1.9%
Existing	Hedonic	6.4%	4.6%	12.5%	0.6%
	Random forest	6.4%	4.8%	12.1%	-0.1%
	Repeat sales	5.9%	4.2%	12.1%	-0.7%
New	Hedonic	6.2%	4.4%	10.6%	3.9%
	Random forest	6.7%	5.0%	10.7%	4.8%

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

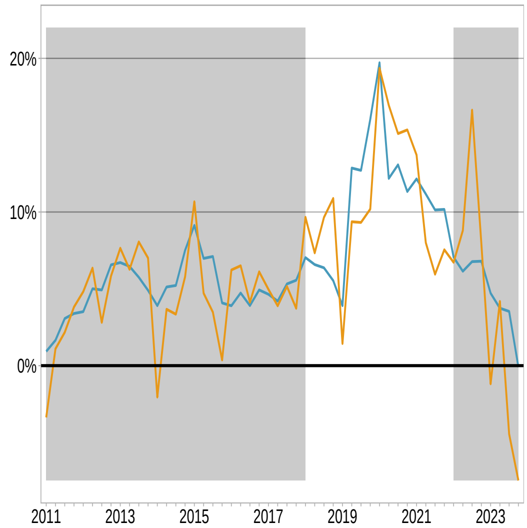
The lower part of Table 2 reports average annual growth in the sub-indices for new and existing apartments. All indices confirm that price growth between 2011 and 2017 was similar for existing and newly built apartments. From 2018 to 2021 increases where slightly higher for existing apartments, possibly because by 2017 newly built apartments were much more expensive (6,500€/m2) than existing apartments (5,100€/m2). However, the slowdown in prices from 2022 to 2023 was more pronounced for existing apartments than for newly built apartments, at least partly reflecting unprecedented increases in construction costs between 2022 and 2023. For both existing and new apartment prices the differences in average growth rates across different indices are not statistically significant.

Figure 5: Year on year growth rates by type of apartment(%)

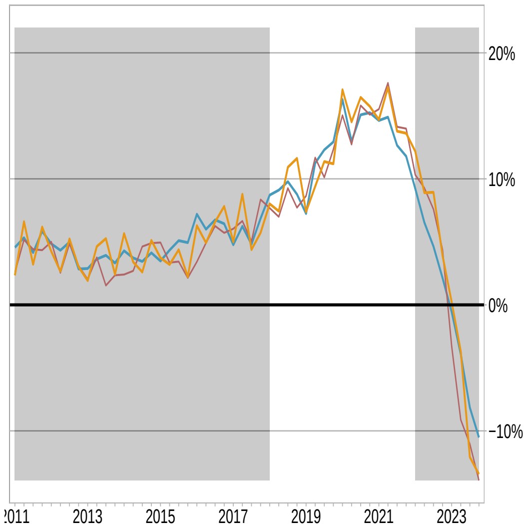
(a) Overall



(b) New apartments



(c) Existing apartments



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

Figure 5 plots annual changes in each index by type of apartment. Comparing our three methods, growth rates for the overall indices are fairly similar over the whole sample. Occasionally, hedonic and random forest subindices for new and existing dwellings diverge in terms of direction (e.g. 2022) or magnitude (e.g. 2021). For existing apartments, where we observe roughly twice as many transactions (Table 1), the series are much less volatile. We will discuss these properties in sections 4.3 and 4.5.

4.2 Regional price developments

Although national indices of residential property prices can be a poor guide for local prices, there are only few indices for comparisons within countries (Fraisse & Pionnier, 2020). In Luxembourg, the hedonic index is computed at the national level. However, the National Housing Observatory ("*Observatoire de l'Habitat*") does publish average prices per square meter at the regional level. These prices are averaged for each municipality where at least 5 transactions took place over the year[20]. Although these statistics provide useful insights into price developments at the municipal level, they should be interpreted with care because there is no quality adjustment.

The limited number of transactions at the municipal level is insufficient for an application of the different approaches presented in this paper. However, accurate regional measures of residential property prices can be important for policy. For instance, they can serve to assess the impact public policy can have on property prices, such as the construction of new roads or the improvement of existing public transport links.

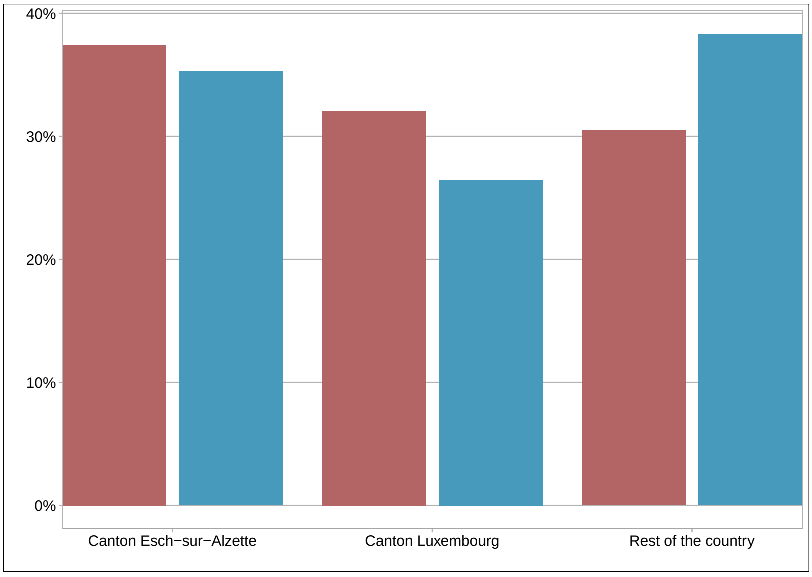
Buyer preferences may also vary over time, which can lead to diverging price dynamics across different regions. The rapid deployment of work from home in the context of the COVID-19 pandemic stimulated a debate on the changing preferences for proximity to the workplace. In Luxembourg, such a change in preferences would lead to an increase in demand in regions further away from the capital.

We split our sample of apartment transactions into three geographic areas, the canton of Luxembourg, the canton of Esch-sur-Alzette and the rest of the country[21]. This allows us to distinguish between price movements in the two largest urban areas of the country and to split the sample into three groups with roughly equal size (see Figure 6)

[20] See: `https://logement.public.lu/fr/publications/observatoire/logement-en-chiffres-15.html`

[21] In 2023, the canton of Luxembourg represented 33 % of the population and the canton of Esch-sur-Alzette another 30 %. The rest of the country combines the 10 remaining cantons that together represent about 36 % of the population.

Figure 6: Share of transactions by geographic region (%)



Note: ■ *Hedonic sample;* ■ *Repeat sales sample. Shares add up to 100%.*

Figure 6 shows the share of transactions across the different regions. The hedonic sample contains more transactions from the canton of Luxembourg and the repeat sales sample contains more from the Rest of the Country. This may reflect the concentration of newly built apartments (which have not yet been resold) in the canton of Luxembourg. Nonetheless, the regional distribution of transactions in the hedonic sample and the repeat sales sample remain relatively close.

Table 3: Average annual price changes per region (%)

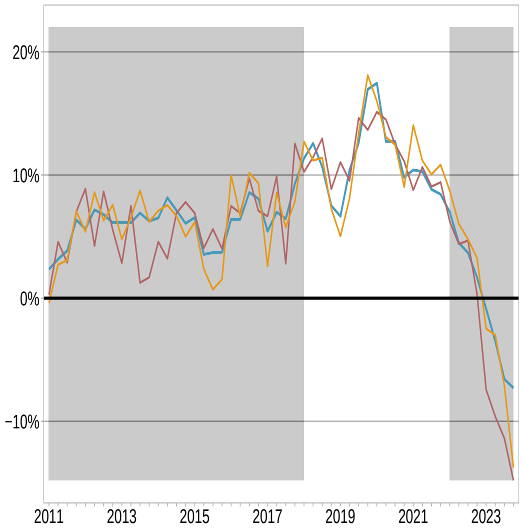
		2011-2023	**2011-2017**	**2018-2021**	**2022-2023**
Canton Luxembourg	**Hedonic**	6.6%	5.9%	11.5%	-0.4%
	Repeat sales	6.1%	5.8%	11.5%	-3.6%
	Random forest	6.7%	6.1%	11.2%	-0.2%
Canton Esch-sur-Alzette	**Hedonic**	6.1%	4.0%	12.9%	-0.1%
	Repeat sales	5.8%	3.7%	12.9%	-1.2%
	Random forest	6.1%	3.9%	12.6%	0.5%
Rest of the country	**Hedonic**	6.1%	3.9%	10.5%	5.0%
	Repeat sales	5.7%	3.5%	11.8%	1.4%
	Random forest	6.2%	4.1%	10.7%	4.6%

Table 3 reports annual price changes since 2011 onwards for the three geographic regions. Between 2011 and 2017, our price indices are closely aligned, showing somewhat stronger price growth in the canton of Luxembourg. Between 2018 and 2021 the three indices show the strongest price growth for the canton of Esch-sur-Alzette. Between 2022 and 2023 all three indices point to a more pronounced slowdown in the cantons of

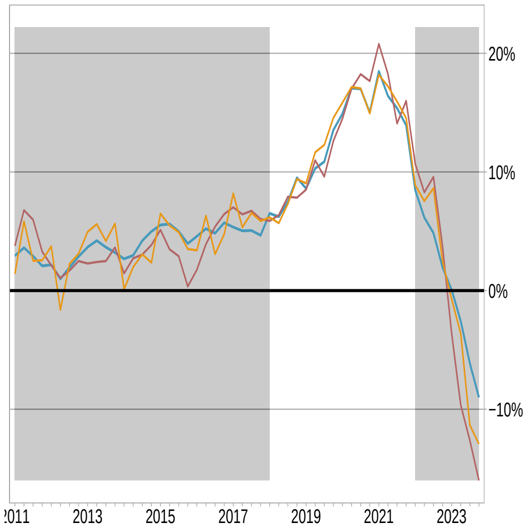
Luxembourg and Esch-sur-Alzette compared to the rest of the country. Furthermore, the repeat sales index seems to show somewhat slower price growth than the two other indices. Again, this reflects the fact that sales of newly built apartments cannot be included in this index.

Figure 7: Apartment prices, year-on-year growth (%)

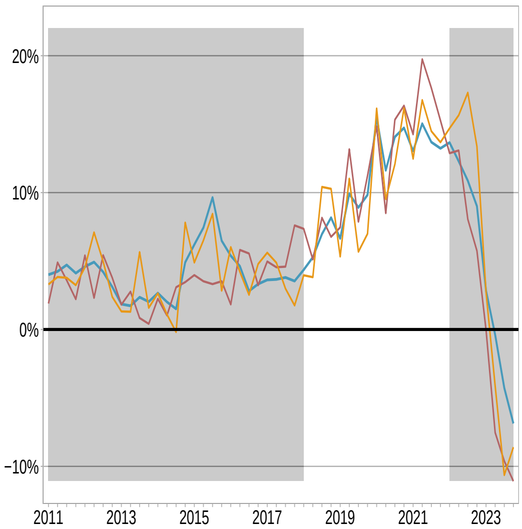
(a) Canton Luxembourg



(b) Canton Esch-sur-Alzette



(c) Rest of the country



Note: *Random forest index;* *Hedonic index;* *Repeat sales index ; Existing and newly built apartments*

Figure 7 plots annual growth rates by region. Two results stand out. First, apartment prices started to grow strongly in the canton of Luxembourg before other regions followed. Around the capital, annual growth rates started to exceed 10 % in 2017, while this did not happen until 2019 for Esch-sur-Alzette and the rest of the country. Second, in the canton of Luxembourg price growth already slowed in 2020, while for Esch-sur-Alzette and the rest of the country this did not happen until 2022. As previously discussed, this second result could in part reflect buyer preference shifting to put less emphasis on commuting times or it could reflect the very high level of prices in the canton of Luxembourg (7,300 €/m2 in 2017) compared to the rest of the country (4,700 €/m2 in 2017).

4.3 Volatility

After comparing the three methodologies visually, we want to evaluate them.

Price volatility is an important feature of any housing market, not just because the subset of dwellings sold may have different characteristics from one period to the next, but also because housing supply is generally slow to adjust to changes in demand. However, volatility is undesirable because it makes it difficult to identify peaks and troughs in real time, as required for policy decisions.

For most households, housing constitutes the largest asset on their balance sheet, so excessive price volatility can affect household welfare[22] (Banks et al., 2017). In this context, it is important to limit volatility due to measurement errors that arise from small estimation samples.

Longer-term trends are often assessed from moving averages or average changes over a longer period. While such methods reduce volatility, they may be unable to provide timely signals of changes in trend (Eurostat, 2013). As a result, constructing a price index always involves a tradeoff between timeliness and volatility.

One of the most common measures of volatility is the standard deviation, reflecting the average deviation from the mean over a period of time. Table 4 reports the standard deviation of year-on-year growth for the overall apartment price indices and separately for existing and new apartments.

[22] According to the 2021 LU-HFCS wave, the main residence and other real estate property account for 91 % of Luxembourg households' total real assets (Mathä et al., 2023).

Table 4: Apartment price index volatility

		2011-2023	**2011-2017**	**2018-2021**	**2022-2023**
Overall	Hedonic	5.5	1.6	3.3	9.5
	Random forest	4.6	0.8	3.0	6.1
Existing	Hedonic	6.1	1.8	3.4	9.7
	Repeat sales	6.2	1.6	3.4	9.8
	Random forest	5.3	1.2	2.8	7.0
New	Hedonic	5.3	2.9	4.5	7.9
	Random forest	3.8	1.7	4.1	2.4

Note: Volatility is measured by the standard deviation of year-on-year price changes.

Standard F-tests cannot reject the null hypothesis of equal volatility for the Random Forest and hedonic indices of overall apartment prices over the complete sample. However, on the subsample of moderate price growth (2011-2017), the random forest index seems to be significantly less volatile than the hedonic index. For new apartment prices, the random forest index is significantly less volatile than the hedonic index over the complete sample and for all subsamples except for the period of high price growth in 2018-2021. For existing apartment prices, all three indices have similar variance for the complete sample, even though, the repeat sales index is estimated on a smaller set of observations. Again, the random forest index is only significantly less volatile for the period of moderate price growth. These results confirm the visual evidence already provided in Figures 7 and 5.

On this criterion, the random forest index seems to be slightly more suitable to identify price trends, especially for new apartments, as the price signal is less blurred by index volatility.

4.4 Revisions

Index reliability is often overlooked as a desirable feature of price indices, especially for residential property prices, given their wide use (Clapham et al., 2006). However, all methodologies presented in this paper produce indices that are subject to revisions over time.

The repeat sales index is revised over time because new apartments are only gradually added to the estimation sample once they are resold. When the property is first resold, information is also added on the first sale in the past.

The random forest index is subject to revision due to the random nature of the algorithm itself. With new observations becoming available, the algorithm may choose different node splits, with an impact on the predicted values. PDPs also simulate the hypothetical sale price of newly added transactions over the entire time span and the predicted prices are then averaged over each period (see Section 2.3). The addition of new transactions to the estimation sample will thus inevitably lead to revisions.

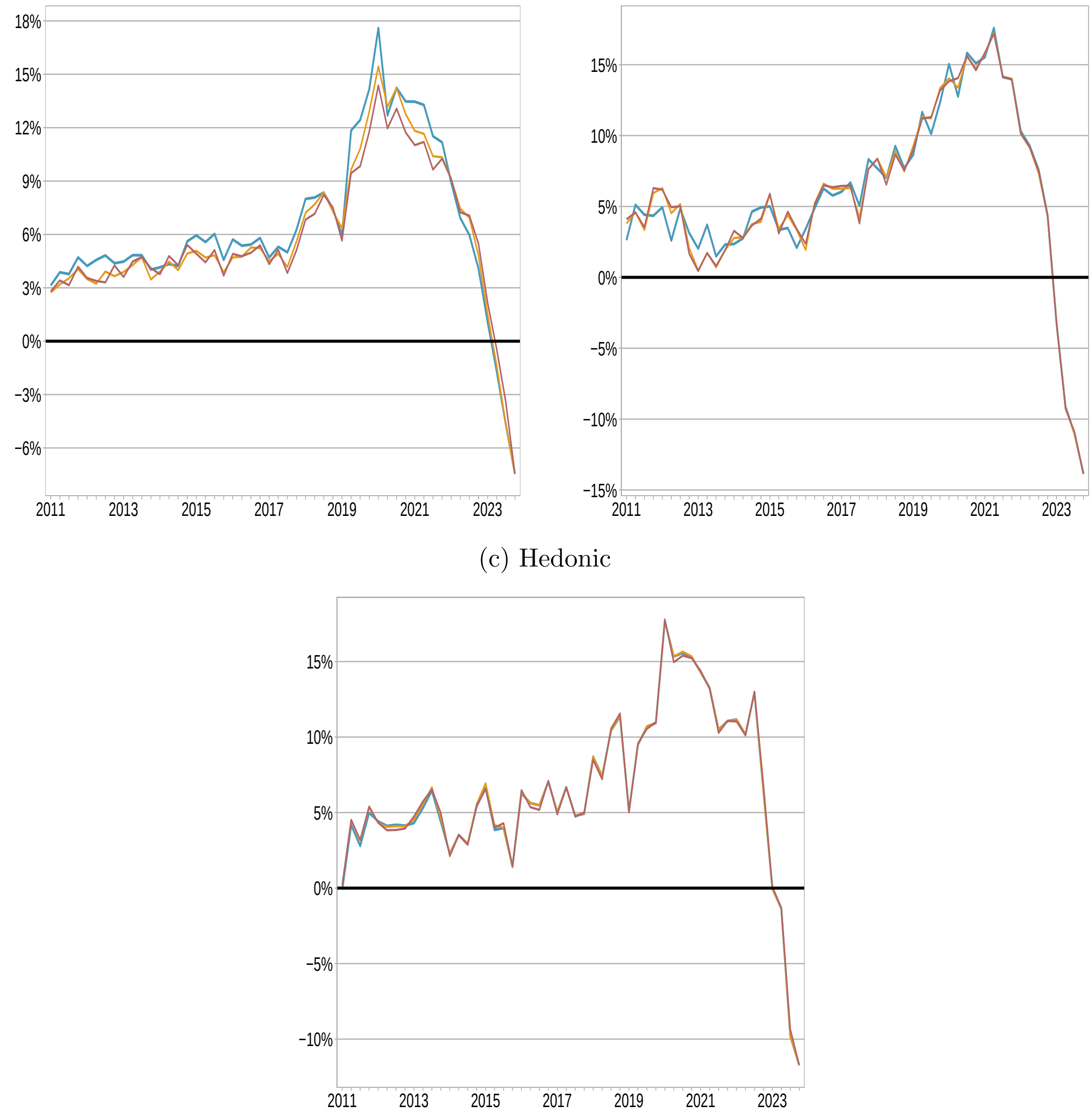
The hedonic method may also be subject to revisions if the regression coefficients in

Equation 1 are re-estimated with every release of new observations and the fitted value of past observations are revised with the new parameter estimates. In Luxembourg, the national indices computed by STATEC are frozen after 2 quarters and only published in Q+1 to account for the fact that certain transactions are transmitted with a delay. In practice, a similar approach could be applied to the repeat sales and random forest method, by fixing the index value after a certain delay if one judges that later revisions are negligible. However, since we are interested in the sensitivity to revisions of the three indices, we chose not to fix any of them after a certain cut-off period.

To assess index reliability, we follow Orphanides and Norden (2002) and measure the extent of revisions by successively re-estimating the indices with every new data release. By doing so, we capture index revisions from new data releases and from the estimation methods themselves. For every period, we estimate three indices, one "real-time", one "quasi-real" and one final index. The real-time index is based on data transmitted until the end of a given quarter, while the quasi-real index is estimated on data transmitted with a delay of up to one quarter. The final index is based on all observations up to the latest available quarter, which is 2023Q4 here[23]. Figure 8 compares year-on-year growth of (quasi) real-time and final estimates for each method separately.

[23]For example, the real-time estimate for 2015Q1 corresponds to the first available estimate in 2015Q1, while the quasi-real estimate takes into account 2015Q1 transactions that were transmitted until 2015Q2. The final estimate, calculated in 2023Q4 hence also takes into account 2015Q1 transactions transmitted later than 2015Q2.

Figure 8: Revision of indices (annual price changes, overall, %)



(a) Random Forest

(b) Repeat sales

(c) Hedonic

Note: The charts compare year-on-year growth for real-time (▬), quasi-real (▬) and the final estimate (▬). Until the final estimate, a new index is obtained by adding new observations up to a cut-off point and then re-estimating the model.

Table 5 shows the results. It is divided into two sections. The left section compares the real-time estimate with the final estimate, while the right section compares the quasi-real-time estimate with the final estimate. We calculate the correlation and root mean squared error (RMSE)[24] for year-on-year growth (Columns (a) and (c)). In addition, we provide summary statistics for the differences between the (quasi) real-time series and the final series (Columns (b) and (d)).

The correlation between the real-time and final series is already high, but it increases slightly when considering transactions with up to a one-quarter delay (quasi real-time). Since most late transactions are transmitted within a one-quarter delay, the right side of the table can be seen as representing the revisions due to the estimation method itself, rather than the addition of past transactions to the estimation sample.

As expected, for the hedonic index the final and quasi real-time series are very closely aligned once most transactions have been collected (Column c), and the differences between the quasi-real and final estimates decrease (Column d). From the point of view of the national statistical agency, it can therefore make sense to delay publication until all transactions are received for the reference period, as this can eliminate most revisions in the series.

It's also noticeable that revisions to the random forest index continue over long periods. Even for the quasi-real estimate, where the estimation sample is almost complete, the maximum difference with respect to the final estimate can be over 2 percentage points. This revision was particularly significant during periods of strong price growth, where the random forest index adjusted more slowly to incoming data.

Table 5: Revision of average annual growth rates (%)

	Final vs Real-Time						**Final vs Quasi Real-Time**					
	YoY Series (a)		**Difference with final estimate (p.p.) (b)**				**YoY Series (c)**		**Difference with final estimate (p.p.) (d)**			
	Correlation	*RMSE*	*Mean (bps)*	*SD*	*Min*	*Max*	*Correlation*	*RMSE*	*Mean*	*SD*	*Min*	*Max*
Hedonic	99.92%	0.00	-0.03	0.23	-0.59	0.38	99.98%	0.00	-0.05	0.12	-0.35	0.12
Random forest	98.43%	0.01	0.64	0.99	-1.39	3.23	99.02%	0.01	0.52	0.67	-0.97	2.20
Repeat sales	98.97%	0.01	-0.01	0.89	-2.35	1.99	99.16%	0.01	0.01	0.80	-1.93	1.99

Note: Statistics based on year-on-year growth rates over 2011q1-2023q4.

4.5 Coherence

The availability of several alternative price indices reduces the uncertainty linked to the interpretation of any single index, provided that they yield coherent results. To analyse coherence across our indices, we follow Mink et al. (2012), who considered both synchronicity and similarity measures to compare different estimates of the euro area output gap. The synchronicity measure, originally proposed by Harding and Pagan (2002), found numerous applications, including to house price cycles (Miles, 2023). We define synchronicity between an index i ind_i and an index j in quarter t as:

[24] The root mean squared error is defined as $RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i^f - y_i^r)^2}$ where n denotes the number of periods, y_i^r the year-on-year growth rate predicted by the real-time or quasi-real time index and y_i^f the growth rate from the final index.

$$\varphi_{ij}(t) = \frac{\Delta ind_i(t) * \Delta ind_j(t)}{|\Delta ind_i(j) * \Delta ind_j(t)|} = \begin{cases} 1 & \text{if } \Delta ind_i(t) * \Delta ind_j(t) > 0 \\ -1 & \text{if } \Delta ind_i(t) * \Delta ind_j(t) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where Δ indicates quarter-on-quarter growth. Less formally, the synchronicity measure φ_{ij} equals 1 if quarterly changes in index i and j have the same sign and -1 if they carry opposite signs. We set synchronicity to 0 if at least one of the two growth rates is equal to 0. We calculate synchronicity for both year-on-year changes and for the cyclical components[25] for all three indices . In the latter case, we calculate Equation 5 by replacing $\Delta ind_i(t)$ with the cyclical component of index i (respectively j) in quarter t. Figure 9 shows year-on-year growth and the cyclical components for existing and new apartments.

Figure 9: Year-on-year growth and cyclical component using Hodrick-Prescott filter (λ=1600)

(a) New apartments (b) Existing apartments

Note: *Random forest* ; *Hedonic* ; *Repeat sales. Dashed lines correspond to annual change rates and solid lines to cyclical components.*

[25]For all three indices, we obtain their cyclical components by applying the Hodrick-Prescott (HP) filter to their logarithms.

This methodology yields a time-varying measure of synchronicity, indicating each quarter whether our indices point the same direction. This constitutes an advantage compared to traditional metrics like the Pearson correlation coefficients, that are time-invariant and may suggest imperfect correlation even when all the indices are perfectly synchronized.

Since our indices have different volatilities, the amplitude of their average fluctuation will also differ. Therefore, we follow Mink et al. (2012) in measuring *similarity* as well as *synchronicity*, where the former is defined:

$$\gamma_{ij}(t) = 1 - \frac{|\Delta ind_i(t) - \Delta ind_j(t)|}{\frac{|\Delta ind_i(t)|+|\Delta ind_j(t)|}{2}} \tag{6}$$

The numerator is scaled by the average absolute year-on-year change of the index pair, where $\gamma_{ij}(t) = 1$ indicates that the annual change in index i (or its cyclical component) is equal to that of index j. By construction, the lowest value for $\gamma_{ij}(t) = 0.5$.

Table 6: Index synchronicity and similarity (2011Q2-2023Q4)

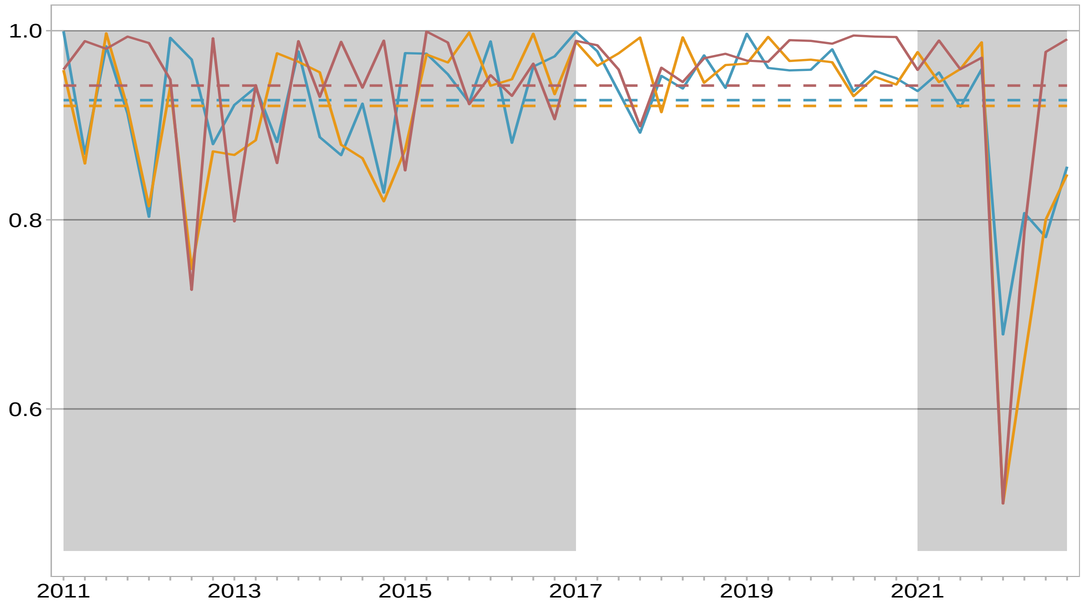
	Existing			New
	Random forest vs. Hedonic	Random forest vs. Repeat sales	Repeat sales vs. Hedonic	Random forest vs. Hedonic
Synchronicity				
Year-on-year changes	100%	98%	88%	75%
Cyclical component (HP filter)	87%	90%	88%	79%
Similarity				
Year-on-year changes	0.93	0.92	0.94	0.88
Cyclical component (HP filter)	0.79	0.82	0.83	0.75

Note: The table reports synchronicity and similarity measures of the respective index pairs. The synchronicity scores indicate the number of quarters in which the index-pairs point into the same direction over the entire sample. The similarity scores indicate the average score for each index pair as defined in Equation 6.

Table 6 presents the synchronicity and similarity measures between different index pairs, with the top panel showing synchronicity and the bottom panel similarity (see Figure 10 for an illustration). For synchronicity, existing apartments demonstrate higher alignment across all dimensions compared to new apartments. For instance, synchronicity between random forest and hedonic indices for annual changes is 100 % for existing apartments but only 75 % for new apartments. A similar pattern is evident for cyclical components, where the same index pair achieves 87 % synchronicity for existing apartments but only 79 % for new apartments. This disparity likely reflects the smaller sample size available for new apartments, which may reduce model robustness. The synchronicity scores also suggest that indices are better aligned in capturing annual changes than fluctuations in the cyclical components obtained by using Hodrick-Prescott filtering.

The similarity measures reveal strong alignment across index pairs, especially for year-on-year changes. This indicates that indices may occasionally point in different directions, but their overall ability to capture price movements remains largely consistent. For existing apartments, similarity scores range from 0.79 to 0.94, with the hedonic and repeat sales indices exhibiting the highest similarity for both annual changes and cyclical components. For new apartments, similarity measures are slightly lower overall, ranging from 0.75 to 0.88.

Figure 10: Similarity scores per index pair - Existing apartments



Notes: ▬ *Random forest - Hedonic;* ▬ *Random forest - Repeat sales;* ▬ *Hedonic - Repeat sales. Similarity scores are based on annual changes of the index series. Dotted line show average similarity scores over the entire period.*

In summary, the indices show generally high synchronicity and similarity across both apartment types and dimensions, highlighting their alignment in capturing market dynamics. New apartments, however, exhibit lower synchronicity and similarity scores compared to existing apartments, likely due to their smaller sample size and potentially greater variability in price movements (Section 4.3). In addition, similarity between index pairs deteriorates especially during turning points in the price cycle. For instance, annual price growth based on the repeat sales index turned negative in 2023Q1, one quarter ahead of the two alternative indices, which led to a deterioration in similarity scores as evidenced in Figure 10.

4.6 Out-of-sample fit

Having examined the coherence of the indices and their sensitivity to revisions, we now assess how well our residential property price indices can predict individual price changes outside the sample on which they were constructed. Although the indices show similar trends and cycles (Section 4.5), their levels differ, raising the question of whether one index can predict individual price movements more accurately than the others. To test this, we evaluate predictions on apartments that were not part of the estimation sample. This approach addresses the challenge of "aggregation bias" that underlies any price index. To construct RPPIs, heterogeneous and infrequently traded objects are pooled together, which may reduce the model "fit" for certain individual dwellings[26].

To assess the relative "accuracy" of different indices, we follow a procedure similar to the one proposed by Krause (2019) and Bogin et al. (2019) based on repeat sales. For each index, we estimate "out-of-sample predictions" of individual apartment prices, using the estimated price change between the first and second date of a repeat sales transaction. Using the notation from Equation 2, we define the predicted price of a repeat sales apartment that was initially sold in period τ and then resold in period t as $\widehat{p_{i,t}} = \frac{ind_t}{ind_\tau} p_{i,\tau}$, where ind_t corresponds to the index value in period t and $p_{i,\tau}$ to the observed price in the first transaction.

In a second step, we measure how close this estimate is to the observed price and we define the error as $e_{i,t} = \widehat{p_{i,t}} - p_{i,t}$. We use standard forecast analysis statistics to evaluate relative accuracy, computing the "*Mean Absolute Percentage Error*" (MAPE), where the MAPE for a given index is equal to $\frac{1}{N}\sum_i^N |\frac{e_i}{p_{i,t}}|$, where all N observations are repeat sales outside the estimation sample.

We use K-Fold cross-validation to perform our analysis, splitting the estimation sample into five "folds", four randomly chosen training samples and one evaluation sample from which we select all the repeat sales[27] (James et al., 2013). This "out-of-sample" evaluation method gives us an idea on how well the indices can predict price movements for individual apartments that were not included in the estimation sample. Implicitly, we thus assume that the transactions held out from the estimation sample follow the "true" price trend and that the observed sale price is the best indicator of an apartment's sales value. This might not always be true, for instance in case of a forced sale, but we address this point with our data cleaning procedure described in Section 3.

The steps for this evaluation method can be summarised as follows (see Annexe A.5 for an illustration):

---

[26] Bollerslev et al. (2016) construct a daily RPPI for 10 major US metropolitan areas to help alleviate potential "*(...) aggregation biases that may plague the traditional coarser monthly and quarterly indices if the true prices change at a higher frequency.*"

[27] Since we consider the repeat sales index informative from 2011 onwards (Section 3), we focus on repeat sales transactions where the first sale date occured after 2010. This provides approximately 9800 repeat sales transactions that we use to evaluate out-of-sample accuracy (Figure 11).

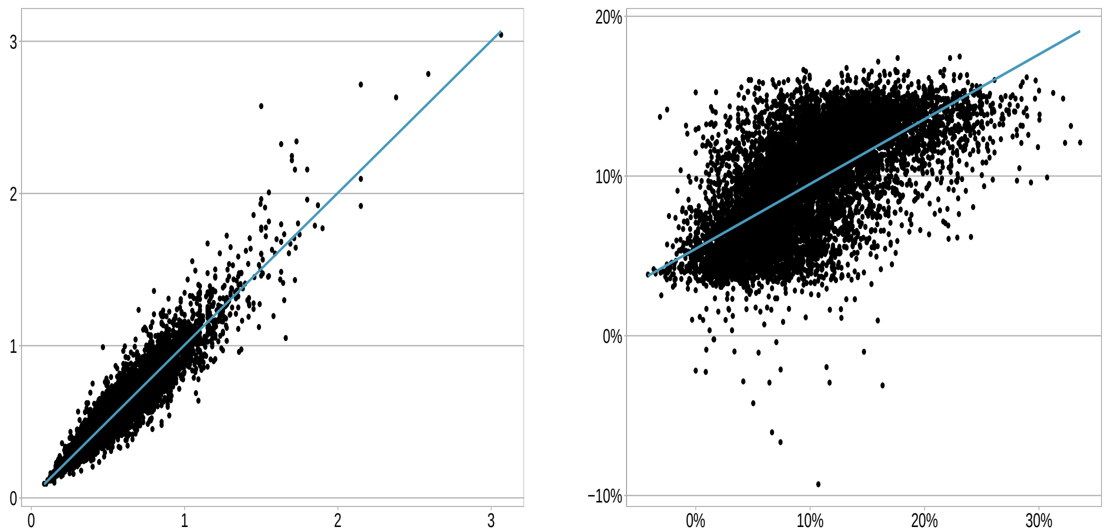
Estimating index accuracy

1. Split the entire sample into K = 5 random folds (subsamples).
2. Estimate each index based on four out of the K randomly chosen (training) folds[28].
3. Identify all the repeat sales transactions from the remaining fold (validation set).
4. For each index, compute the error $e_{i,t}$ between the observed and the implied price of the repeat sales identified in Step 3.
5. Repeat Steps 2-4 K times and report the MAPE for each index.

Table 7 reports the accuracy measures (MAPE) for the national sample and for the regional sub-samples. Figure 11 compares the observed and predicted price levels (left panel) and annualized price changes (right panel) for the random forest index. One limitation of this analysis is that the MAPE assigns the same weight to all transactions, regardless of the interval between sales and therefore does not consider that errors may increase with time (as documented in Bogin et al. (2019)). We relegate this adjustment to future research on this topic.

Figure 11: Out-of-sample predictions (Random forest)

(a) Price level (Mio. EUR) (b) Annualized price change (%)



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

[28] Bogin et al. (2019) refer to these indices estimated on a subset of the data as "trial" price indices.

Thus, we randomly split the entire sample of transactions into 5 folds and then treat each of these folds in turn as the evaluation fold, estimating the model on the other four folds to predict repeat sales in the evaluation fold. From Table 7 we can read that over all observations in Esch-sur-Alzette the random forest yields an average (out-of-sample) prediction error of approximately 8.8 %.

Table 7: Accuracy of out-of-sample predictions of repeat sales prices, by index

	Price level (MAPE)		
	Hedonic	**Random forest**	**Repeat sales**
Overall	9.96%	10.02%	9.96%
Esch-sur-Alzette	8.80%	8.80%	8.90%
Luxembourg	10.45%	10.40%	10.62%
Rest of the country	9.50%	9.30%	9.34%

Note: The national sample is composed of all repeat sales. Regional breakdowns consider only repeat sales in that given region.

The average forecast error is similar across the three methods, which is reassuring in the sense that the indices are fairly accurate, on average, in predicting individual price changes out-of-sample. To address the potential aggregation bias arising from apartments appreciating at different rates across regional submarkets grouped within a national index, we calculate accuracy measures for regional indices (Table 7). Looking at these regional breakdowns, we see that out-of-sample forecast accuracy is again similar across the three price indices and tends to be lowest for the canton of Luxembourg, possibly suggesting a higher heterogeneity in apartment price developments around the capital. For the two other regions out-of-sample forecasting accuracy appears to be higher despite the lower number of observations.

Contrary to some expectations, our findings at regional level reveal that no single approach considerably improves out-of-sample accuracy. One possible explanation is that larger markets, typically characterized by greater heterogeneity in apartment types (variations not fully captured by explanatory variables), may exhibit distinct price trends that justify more granular indices. In contrast, since the Luxembourg market is relatively small and therefore more homogeneous, it may not require such granularity. A national index that incorporates key characteristics, including the distance to the city center, may thus suffice. Moreover, the results highlight a trade-off between aggregation bias and estimation error, particularly when fewer transactions are available to construct regional indices (Bogin et al., 2019).

5 Conclusion

In this paper, we compare the performance of three different methods to calculate a residential property price index from individual transactions on apartments sold in Luxembourg. In addition to the hedonic method, used by the national statistical office to calculate the official index, we present the repeat sales method and the random forest algorithm, a common machine learning method that has been applied to aggregate prices of heterogeneous goods.

The main conclusions are the following. All three methods give a similar picture of trends at the aggregate level. The new random forest index closely tracks the two other (more traditional) indices, which empirically supports the validity of this approach. In addition, the random forest index is less volatile and therefore easier to interpret. Its out-of-sample forecast accuracy for individual transactions is comparable to those of the other indices. However, as new observations are released, the random forest method is subject to greater revisions than the repeat sales method, meaning it provides a less reliable signal of recent developments.

This paper presents a practical application of machine learning techniques to modelling residential property prices and the results from this exercise are encouraging. Despite some potential for improvement, the findings suggest that the random forest is at least comparable and occasionally preferable to the traditional methods.

Since our indices are fairly similar, improvements need not focus on the estimation method itself. Performance could be improved by focusing on the underlying data describing the characteristics of a dwelling. In Luxembourg, important characteristics such as the location or the size of an apartment are available via notary deeds. Other relevant information, such as the number of rooms or major renovations are not available from this source, although they could potentially improve overall performance.

Summing up, this paper compared the performance of three apartment price indices based on different estimation methods. Users may decide which index best suits their needs. However, consulting different, yet coherent, indices does reduce the uncertainty associated with relying on any single price index.

References

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied sciences*, *8*(11), 2321.

Banks, J., Blundell, R., Oldfield, Z., & Smith, J. P. (2017). *3. House Price Volatility and the Housing Ladder*. University of Chicago Press.

Bogin, A. N., Doerner, W. M., & Larson, W. D. (2019). Missing the mark: Mortgage valuation accuracy and credit modeling. *Financial Analysts Journal*, *75*(1), 32–47.

Bollerslev, T., Patton, A. J., & Wang, W. (2016). Daily house price indices: Construction, modeling, and longer-run predictions. *Journal of Applied Econometrics*, *31*(6), 1005–1025.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees - 1st edition.* Wadsworth.

Case, K. E., & Shiller, R. J. (1987). *Prices of single family homes since 1970: New indexes for four cities.* National Bureau of Economic Research Cambridge, Mass., USA.

Clapham, E., Englund, P., Quigley, J. M., & Redfearn, C. L. (2006). Revisiting the past and settling the score: index revision for house price derivatives. *Real Estate Economics*, *34*(2), 275–302.

Crowe, C., Dell'Ariccia, G., Igan, D., & Rabanal, P. (2013). How to deal with real estate booms: Lessons from country experiences. *Journal of financial stability*, *9*(3), 300–319.

Englund, P., Quigley, J. M., & Redfearn, C. L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *The journal of real estate finance and economics*, *19*(2), 91–112.

Eurostat. (2013). *Handbook on residential property prices indices (RPPIs)* (J. de Haan & E. Diewert, Eds.). Publications Office of the European Union.

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies*, *43*(12), 2301–2315.

Financial Stability Board, I. M. F. (2009). *Report to the G-20 Finance Ministers and Central Bank Governors.*

Fraisse, A.-S., & Pionnier, P.-A. (2020). Location, location, location - House price developments across and within OECD countries. *OECD Statistics Blog*. Retrieved from #https://oecdstatistics.blog/2020/07/10/location-location-location-house-price-developments-across-and-within-oecd-countries/#

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.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *R Project Journal*, *9*(1), 421.

Harding, D., & Pagan, A. (2002). Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics*, *49*(2), 365–381.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jansen, S. J. T., de Vries, P. A. U. L., Coolen, H. C. C. H., Lamain, C. J. M., & Boelhouwer, P. J. (2008). Developing a house price index for the Netherlands: A practical application of weighted repeat sales. *The Journal of Real Estate Finance and Economics*, *37*(2), 163–186.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2015). Betting the house. *Journal of international economics*, *96*, S2–S18.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2016). The great mortgaging: housing finance, crises and business cycles. *Economic policy*, *31*(85), 107–152.

Kaempff, B., & Kremer, D. (2021). Un indice des prix des appartements au Luxembourg basé sur la méthode des ventes répétées. *BCL Bulletins*, *2021*(3), 85–97.

Krause, A. (2019). A Machine Learning Approach to House Price Indexes. *Working Paper, Zillow Group*.

Lamboray, C. (2010). Un Indice Des Prix Hédonique Des Appartements. *STATEC Economie Et Statistiques*, *44*.

Mathä, T., Montes-Viñas, A., Pulina, G., & Ziegelmeyer, M. (2023). The Luxembourg Household Finance and Consumption Survey: Results from the fourth wave in 2021. *BCL Working Papers*(176).

Miles, W. (2023). Regional house price co-movement in the USA: the medium cycle is not the business cycle. *The Annals of Regional Science*, *71*(2), 437–462.

Mink, M., Jacobs, J. P. A. M., & de Haan, J. (2012). Measuring coherence of output gaps with an application to the euro area. *Oxford Economic Papers*, *64*(2), 217–236.

Orphanides, A., & Norden, S. v. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, *84*(4), 569–583.

Paccoud, A., Licheron, J., & Bérard, G. (2024). Panorama des transactions immobilières entre 2007 et 2023. *Observatoire de l'Habitat, Note 37*.

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, *18*(1), 6673–6690.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Roma, M. (2021). Euro area house price developments during the coronavirus pandemic. *Economic Bulletin Boxes*, *4*.

Scatigna, M., Szemere, R., & Tsatsaronis, K. (2014). Residential property price statistics across the globe. *BIS Quarterly Review September*.

Silver, M. (2012). Measurement Matters for House Price Indices. *IMF Research Bulletin*, *2012*(004).

Slack, D., Friedler, S. A., Scheidegger, C., & Roy, C. D. (2019). Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*.

Smith, L. B., Rosen, K. T., & Fallis, G. (1988). Recent developments in economic models of housing markets. *Journal of Economic Literature*, *26*(1), 29–64.

Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences*, *6*(4), 165–171.

Wang, F. T., & Zorn, P. M. (1997). Estimating house price growth with repeat sales data: what's the aim of the game? *Journal of Housing Economics*, *6*(2), 93–118.

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.

A Appendices

A.1 Regression trees

A regression tree is built through an iterative process that splits data into partitions or "branches" and continues splitting each partition into smaller groups ("recursive partitioning") (Breiman et al., 1984).

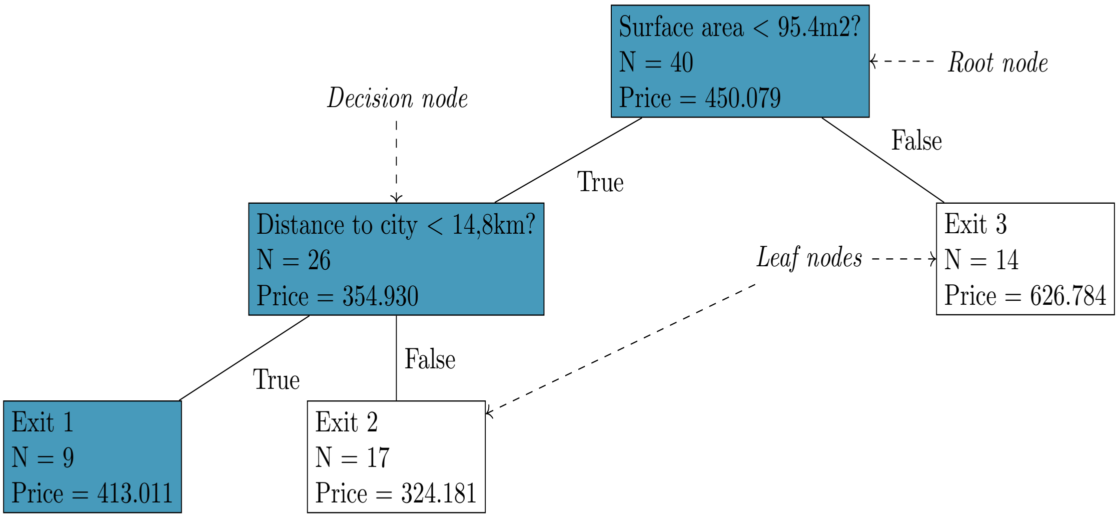
Initially, all observations from a sample $(x_{i1}, ..., x_{iK}, y_i)$, where $x_{i1}, ..., x_{iK}$ is drawn from the predictor space $(X_1, ..., X_K)$ and y_i is drawn from the outcome vector Y for $i = 1, ..., N$, are grouped into a single partition at the top of the tree (root node).

Then the algorithm uses a splitting rule to assign observations to two partitions based on the j^{th} predictor from the predictor space in a process called "node splitting" (Appendix A.2 explains how the predictor j and the splitting value s are determined at each node). The algorithm continues splitting the set of observations at each node (decision nodes) until it reaches a certain stopping criterion, such as the minimum number of observations per node.

Predictions from a regression tree are made by starting with an observation i for at the root node. To traverse the tree, we start at the root node and recursively move to the left or right child nodes depending on the decision rule at the root node and the value of the given feature in observation i. Depending on whether the value x_{ij} is less than or greater than the splitting value s, the decision rule dictates a move down to the left or to the right in the decision tree. This process is repeated in every subsequent decision node until the observation is assigned to a terminal (leaf) node.

For a continuous variable, the predicted outcome is equal to the mean value across all observations in that given leaf node. In general, regression trees tend to perform well if the relationship between the outcome variable and its predictors is complex and poorly approximated by linear forms such as the one described in Equation 1.

Figure A.1: Structure of a regression tree - Illustration



Notes. The blue path can be interpreted as follows: Out of 40 apartments, 26 were smaller than 95.4m2. This value was chosen by the algorithm as the best split because it minimizes the Residual sum of squares in the two subsequent decision nodes (see Annexe A.2 for more detail). Among those 26 apartments, 17 were located more than 14.8km away from the capital. In the subsample represented by the leaf at the far left (Exit 1), the average transaction price equals €413.011. For this very simplified example, the minimum node size was set to 20 observations.

Regression trees can be displayed graphically, which explains in part why they gained in popularity. Figure A.1 shows a stylized example of a regression tree, which has the transaction price as an outcome variable. Two predictors are used, namely the surface area of the apartment and its distance to the capital.

A.2 Node splitting rules

To build a regression tree, the set of possible values of the predictors $(x_1, ..., x_K)$ has to be divided into distinct regions. Using the notation from James et al. (2013), the process can be summarised as follows:

1. For each split, the objective is to find regions $R_1, ..., R_D$ that minimize the in-sample residual sum of squares (RSS), which is given by:

$$RSS = \sum_{d=1}^{D} \sum_{i \in R_d} (y_i - \hat{y}_{R_d})^2 \tag{7}$$

   where $\hat{y}_{R_d}$ is the mean of the outcome variable y for all observations that fall in region R_d.

   We start at the top of the tree and consider all K possible predictors at each possible value $x_{i1}, ..., x_{iK}$. Next, we construct two regions, R_1 and R_2 such that

$R_1(j,s) = \{X|X_j < s\}$ and $R_2(j,s) = \{X|x_j \geq s\}$, where $\{X|x_j < s\}$ denotes the region in which X_j takes a value less than a cutpoint s. The aim is to find two regions, which lead to the greatest reduction in RSS.

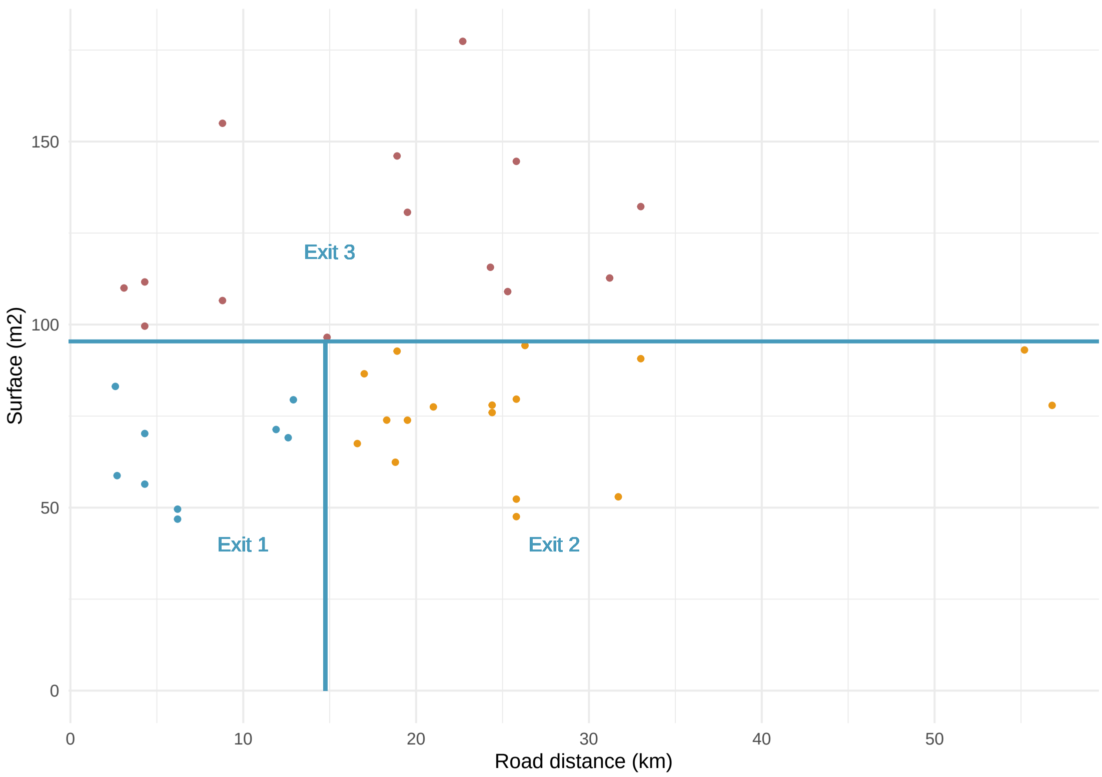
2. From 7, we thus seek j and s to minimise the equation at the root node:

$$\sum_{i:X_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:X_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{8}$$

3. We iterate this process in R_1 and R_2 and look for the best j and s to minimise the RSS within each of the regions and stop the process when we reach a pre-defined stopping criterion.

Figure A.2 shows a simple illustration of the three partitions from the stylized tree in figure A.1.

Figure A.2: Partitioning process - Illustration



Notes. The three region partition from Figure A.1. The tree has stratified the apartments into three distinct regions of the predictor space: Exit 1 = $\{X|Surface < 95,4; Distance < 14,8km\}$, Exit 2 = $\{X|Surface < 95,4; Distance \geq 14,8km\}$ and Exit 3 = $\{X|Surface \geq 95,4\}$.

A.3 Random forest tuning parameters

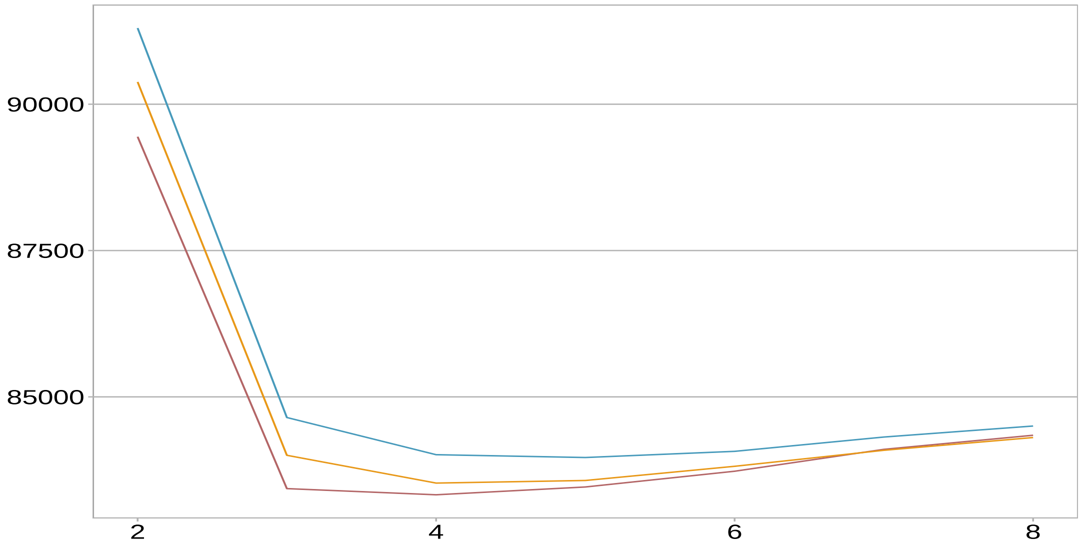
The random forest algorithm depends on the values taken by several hyperparameters. These control the structure of each tree (the minimum number of observations each node should have to be split, *min.node.size*), the size of the forest (the number of trees, *num.trees*) and its level of randomness (the number of variables considered as splitting variables at each split, *mtry*). Controlling the set of hyperparameters is referred to as "hyperparameter tuning" and relies on experimental results.

The random forest algorithm is known to perform well "out-of-the-box" (Athey & Imbens, 2019). It requires little parameter tuning to perform well and most random forest packages for statistical software define default settings. In our case, the default *mtry* is equal to the rounded down square root of the number variables and the minimum node size to 5. Still, there are no theoretical findings to support the default values and it is generally recommended to validate them by using different parameter combinations and to evaluate the performance of each setting. Parameter settings that correspond to complex trees or forests may indeed overfit the training data, i.e. yield predictions that are too specific to the training data and perform less well on other data. For example, Breiman (2001) emphazized that the mtry should be chosen such that "*the randomness injected minimizes the correlation while maintaining strength*".

The charts below show the results from a cross-validation (CV) procedure. We use K-Fold cross-validation, which splits the training data into K subsets, called folds. The model is iteratively fitted K times, each time training the data on K-1 of the folds and evaluating on the K-th fold. Here, K = 3 and in the first iteration we train on the first two folds and evaluate on the third. In the second iteration we train on the first and the third fold and evaluate on the second fold and so on. This K-Fold cross-validation process is repeated several times, each time using different hyperparameter settings and the final evaluations are based on the average outcomes from the 3 folds.

Figure A.3 shows the outcome of this process with the vertical axis indicating the RMSE obtained from 3-fold CV. This was used to select the optimal model. The final hyperparameters selected for the model are 4 randomly selected variables for each tree with a minimum node size of 5. The results confirm that random forest generally performs well with little finetuning as the selected parameters are very close to the default settings.

Figure A.3: 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

There is no clear guidance on the size of the forest, i.e. the number of individual trees grown, but it should be set sufficiently high to get stable predictions. In this sense, the number of trees is not a proper finetuning parameter because higher values are generally preferable (Probst, Wright, & Boulesteix, 2019). On the other hand, the computational intensity increases with the number of trees and the gains in terms of performance may be negligible compared to the costs in terms of computation times.

Figure A.4: Number of trees vs. Error

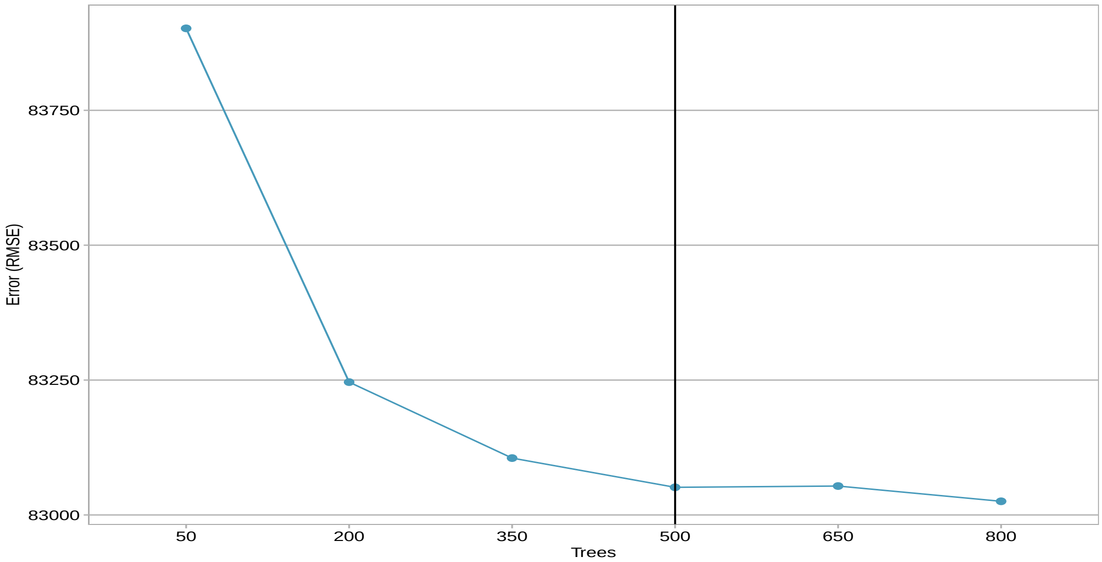


Figure A.4 shows the RMSE obtained from re-estimating the random forest algorithm with different settings for the number of trees. As we can see, there is only limited improvement in terms of error reduction after 500 trees, which was chosen for our model and which is a standard value in the literature. It also confirms empirical findings that the biggest performance gain is achieved when training the first 100 trees (Probst & Boulesteix, 2017).

A.4 Correlation matrices

Figure A.5: Existing apartments

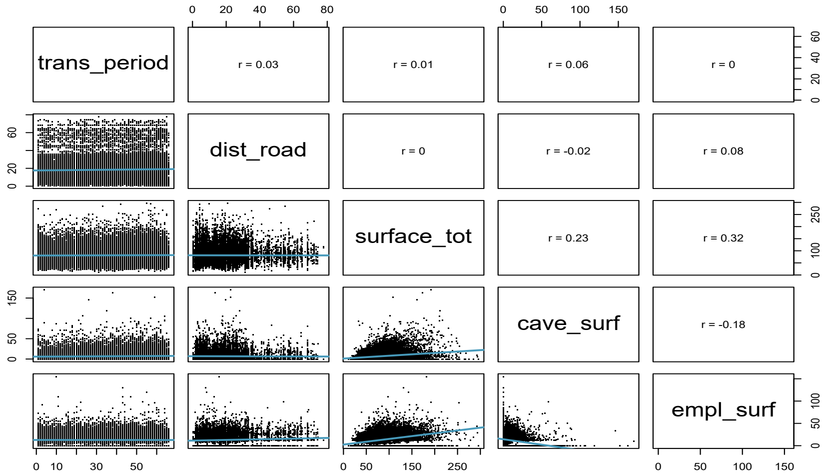
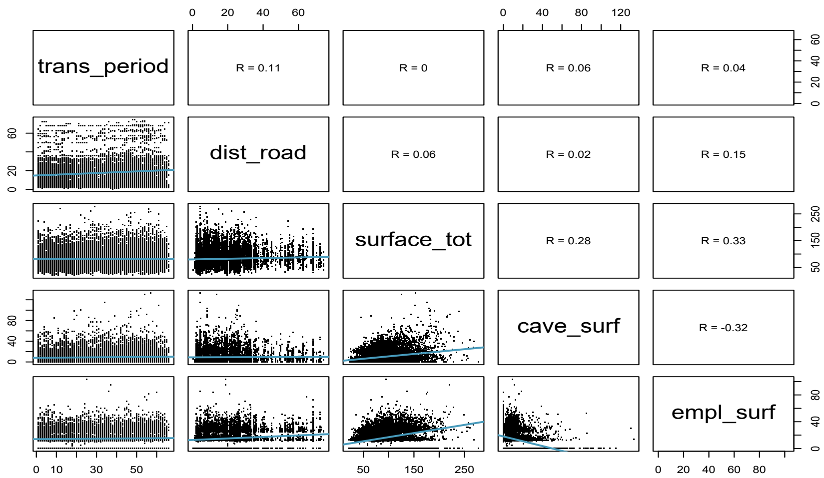


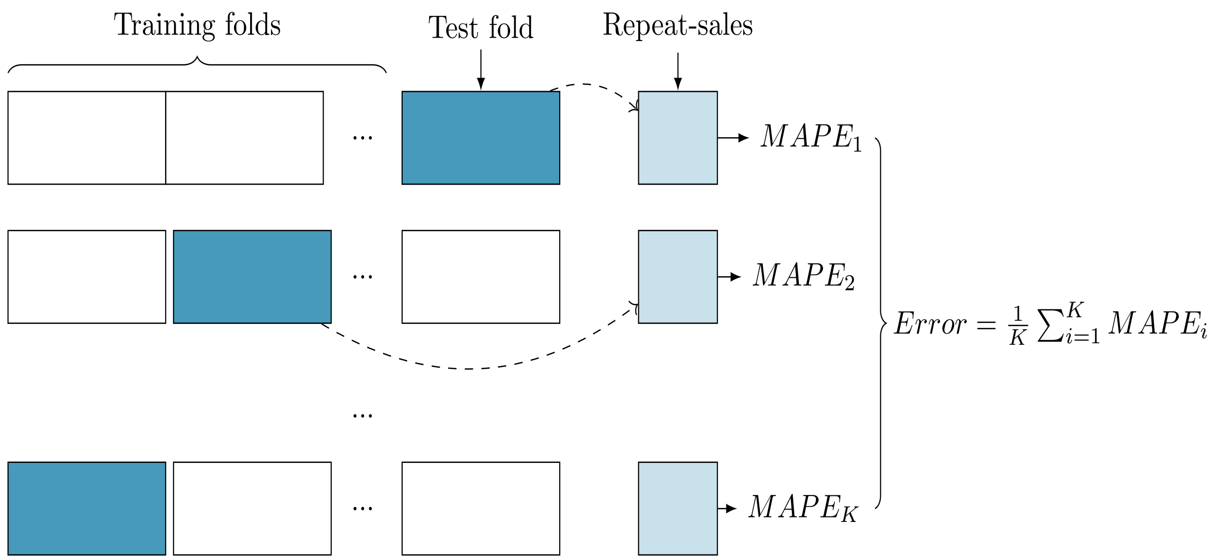
Figure A.6: New apartments



Notes. Figure A.5 and A.6 plot the correlation matrices for the numerical variables included in our model. We included the transaction period (*trans_period*), the distance to the capital (*dist_road*), the surface of the apartment (*surface_tot*), the size of the cellar (*cave_surf*) and the size of the parking space(*empl_surf*). 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.

A.5 K-fold cross-validation

Figure A.7: K-Fold Cross validation - Illustration



List of Figures

List of Tables

BANQUE CENTRALE DU LUXEMBOURG
EUROSYSTÈME

2, boulevard Royal
L-2983 Luxembourg

Tél.: +352 4774-1
Fax: +352 4774 4910

www.bcl.lu • info@bcl.lu