Filling the Gaps with MICE-RF: Addressing Missing Data in Property Price Indices

Miriam Steurer*[a] and Sabrina Spiegel[b]

[a]Department of Economics, University of Graz, Austria, miriam.steurer@uni-graz.at
[b]ZTdatenforum, Graz, Austria

February 10, 2025
Preliminary draft – please do not circulate

Abstract

Missing data pose a significant challenge in the construction of real estate price indices, potentially distorting key economic indicators used in monetary and macroprudential policy decisions. Official index compilers frequently rely on complete-case analysis, which discards incomplete observations. However, when missingness is non-random, this approach can introduce systematic biases—particularly in commercial real estate (CRE) markets, where data gaps are more prevalent. Despite the critical importance of accurate real estate price measurement, international statistical organizations have yet to establish clear guidelines, and existing literature lacks robust imputation strategies tailored for price index construction.

This paper evaluates whether Multiple Imputation by Chained Equations (MICE), combined with the Random Forest algorithm (MICE-RF), can effectively mitigate these biases. Using a complete transaction dataset for Viennese apartments (2015–2023), we introduce artificial missingness and demonstrate that MICE-RF reliably corrects biases, even under strong systematic missingness patterns. We then apply this method to Austrian office transaction data, where property size — a key explanatory variable — is missing in over half of the cases. Our results show that multiple imputaiton via MICE-RF significantly alters the estimated price index trajectory for the period 2021 to 2024: while complete-case analysis suggests rising office prices, the imputed index indicates stable or declining prices, aligning more closely with market fundamentals.

Our findings underscore the importance of incorporating imputation techniques into official index compilation to enhance the reliability of real estate price indicators. As transaction-based indices play an increasingly central role in economic analysis and financial stability assessments, addressing missing data is not merely a statistical refinement but a necessity for informed policy decisions.

Keywords: Commercial Real Estate price index; Hedonic Price Index; Missing Data; Multiple Imputation by Chained Equations (MICE); Random Forest Imputation
JEL codes: R31, E31, C81, R33, C55, E44

Acknowledgments: We thank *ZTdatenforum* for providing us with the data for this analysis.

*Corresponding author.

1 Introduction

This paper examines the impact of missing data on price indices and investigates whether imputing missing values via the Random Forest (RF) algorithms within the *Multiple Imputation by Chained Equations* (MICE) framework can enhance the quality of transaction-based real estate price indices. Using residential and commercial real estate data from Austria (2015–2024), we first demonstrate how different types of missingness influence price index results. We then illustrate how multiple imputation via MICE-RF can significantly improve the validity of price indices even when data missingness is highly systematic, addressing limitations of current approaches such as complete-case analysis and simple imputation methods.

The real estate sector is a key economic driver and serves as a leading indicator of broader macroeconomic trends (Leamer, 2007). Real estate price indices, produced by national statistical institutes (NSIs) and Central Banks, are fundamental tools for guiding monetary policymakers in assessing market dynamics. However, constructing high-quality price indices for real estate markets presents significant challenges. Unlike standardized goods, real estate properties exhibit substantial heterogeneity in terms of size, quality, and location, complicating efforts to measure price inflation accurately. In contrast to markets where inflation can be measured from the price change for uniform products, real estate markets require more sophisticated methods. Furthermore, as real estate markets are highly heterogeneous, comprehensive market coverage is essential; incomplete property price data may subsequently lead to inaccuracies in property price indices, which, in turn, could affect the reliability of monetary and macroprudential policy decisions.

Internationally, hedonic regression based on transaction data is the primary method for constructing official property price indices (Eurostat, 2013, 2017; Hill et al., 2018; International Monetary Fund, 2020; Commission Implementing Regulation, 2023). While purchase contracts consistently document transaction prices and dates, they often lack detailed property characteristics such as size, condition, or amenities, leading to systematic missing data entries in real estate transaction data. This is particularly the case for commercial real estate (CRE) data, which typically contain many more missing values than residential real estate (RRE) data.

When faced with incomplete data, NSIs and central banks typically decide between two options: the first is to use only transactions with complete descriptive variables to compile the property price indices (a method known as "list-wise deletion" or "complete-case" analysis). Option two consists of estimating property price indices without using descriptive variables, thereby compromising the quality adjustments in the resulting hedonic price index.

A third approach, imputing missing values, is generally not employed. There are two reasons for this omission. First, international statistical agencies like Eurostat or the IMF have yet to provide

clear guidelines to real estate price index compilers.[1] Second, and maybe more importantly, the real estate price index literature has been slow to engage in the discussion on if and how the imputation of missing values should be performed. However, other research fields have actively explored imputation techniques. Especially in medical research, imputation techniques have been thoroughly debated for decades, beginning with seminal contributions from Rubin (1976, 1987) and Little (1988).

The potential consequences of biases in official real estate price statistics can be large and highlight the need for a robust discussion on how to manage missing values in transaction datasets. This issue is not new. In a wider price index context, guidelines on how to deal with missing observations are mainly concerned with finding appropriate substitutes for missing product prices than employing imputation methods (Armknecht and Maitland-Smith, 1999). Feenstra and Diewert (2003) examine the effects of single imputation methods (i.e., the carry-forward, the cell-mean, and the linear interpolation method) for missing data within the ICP program. The authors conclude that imputation improves price index construction, but also note that the choice of method significantly impacts the index's properties. While this study establishes that imputation can improve price indices, it does not explore modern multiple imputation techniques, leaving a critical gap in the real estate price index literature that our paper seeks to address. It also concentrates on general price data rather than property transaction data. In a different context, Kofman and Sharpe (2003) discuss the use of multiple imputation with finance data (an area which also suffers from high percentages of missing data) and conclude that multiple imputation provides a robust framework for handling missing data.

Multiple imputation offers significant advantages over single imputation and complete-case analysis (Rubin, 1987; Little and Rubin, 2019). Advances in computing power and statistical software have made multiple imputation techniques faster and possible for large real estate datasets. Among multiple imputation strategies, Multiple-Imputation-by-Chained-Equations (MICE), has, in particular, made imputation tasks more accessible, flexible, and robust. MICE is a widely used framework for multiple imputation. While it has been the gold standard in medical research (Erler et al., 2016), its application in real estate price index construction remains unexplored. To our knowledge, this paper is the first to illustrate the feasibility and benefits of incorporating multiple imputation via MICE into official property price indices. MICE is relatively easy to implement. It is also very flexible, allowing the user to specify parametric and non-parametric estimation techniques within the MICE process.

MICE imputes missing values through sequential estimation rounds (van Buuren and Groothuis-Oudshoorn, 2011). The algorithm considers one variable at a time, taking all observed data and the

[1] In the context of Consumer Price Index (CPI) inflation measurement, there exist some recommendations of how compilers should treat missing values (Eurostat, 2018, 2020; Fund, 2020). However, these recommendations take the form of simple imputation rules rather than state-of-the-art imputation methods. Specifically, both Eurostat and IMF recommend "mean imputation" (i.e., substituting the mean for a missing value), the "carry-forward method" (i.e., the last observed price is substituted for a missing price), and using similar products to impute the missing price. In the context of real estate price indices, where each property is different and missing values can occur in a multitude of variables, these guidelines are of little use.

imputed values from the previous round as input. This iterative process continues until the imputed values converge. Once one dataset is filled up in this way, the process is repeated, and further imputed datasets are produced. Separate analyses are then performed on each of these imputed datasets, and the results are aggregated. The aggregation of results produced from different datasets can be interpreted as a form of "model averaging", which is known to increase the reliability of the final outcome (Moral-Benito, 2015; Steel, 2020).[2] Additionally, the variations between the imputed datasets give rise to variations between the individual sub-results and can thus illustrate how sensitive the results are to the missing values (Rubin, 1976).

Imputation via MICE makes the imputation process easier, but by itself does not solve all our imputation worries: One flaw of standard multiple imputation procedures is that model selection has to be done manually (or default regression methods used). Thus, they fail to automatically account for nonlinear relationships and interaction effects (Doove et al., 2014; Deng and Lumley, 2024). This means that apropriate imputation modeling requires time and significant domain knowledge to succeed well. A second potential flaw of standard multiple imputation procedures is that they do not adequately account for model uncertainty (Goh and Kim, 2019). Typically, each variable with missing data is imputed using a single model, which the researcher specifies. This approach can lead to underestimation of variability and potential bias in subsequent analyses, as it fails to incorporate the uncertainty inherent in the model selection process. Related to this aspect is the critique that the flexibility inherent in MICE leaves room for potential misspecification and thus biased imputations if the standard parametric imputation algorithms (e.g., linear regression, PMM) are used (Hayes, 2018; van Buuren, 2018).

While multiple imputation via chained equations (MICE) simplifies the imputation process, it does not address all associated challenges. One notable limitation of standard multiple imputation procedures is that model selection must be performed manually — or rely on default regression methods. Thus, there is no automatic mechanism that ensures that nonlinear relationships and interaction effects between variables are captured appropriately (Doove et al., 2014; Deng and Lumley, 2024). This implies that good imputation modeling requires time and significant domain knowledge to succeed well. A second potential flaw of standard multiple imputation moprocedures is that they do not adequatley account for model uncertainty (Glaeser et al., 2005). Typically, each variable with missing data is imputed using a single model specified by the researcher. Using only one model for each variable can lead to an underestimation of variability and introduce bias in subsequent analyses, as it does not fully incorporate the uncertainty inherent in model selection.

However, these shortcomings of standard multiple imputation techniques can be solved by using tree-based methods instead of the standard parametric regression estimation within the multiple imputation framework. Tree-based methods, like the random forest algorithm, will automatically account for both

[2]The relationship with the model averaging literature is further enhanced when slight differences between the models used for analysis are introduced (Schomaker and Heumann, 2014).

model uncertainty and interaction effects without additional modelling input from the researcher. As real estate data often feature complex, nonlinear relationships between property characteristics (e.g., size, location, amenities), tree-based methods are particularly suitable for our imputation task. Tree-based methods that vary the available variables that the algorithm can choose at each stage (e.g., the random forest or the XGBoost algorithms) also automatically force some model variation. Thus, they provide a computationally easier way to include the idea of model averaging (Moral-Benito, 2015; Steel, 2020) than incorporating Bayesian Model Averaging into the imputation process (as done e.g. by Goh and Kim (2019)). Another positive aspect of tree-based methods is the automatic sub-sampling and training on sections of the available data, and the possibility to handle numeric as well as categorical variables. Finally, MICE in conjunction with random forest algorithm can handle imputation problems where the number of variables exceeds the number of observations (Hayes, 2018).

We assess the effectiveness of MICE in conjunction with Random Forests (MICE-RF) in a high-missingness, real-world setting, using two real estate transaction datasets from Austria. Our first dataset consists of apartment transaction data for Vienna from 2015 to 2023, which does not contain missing values. We artificially introduce different missingness patterns into the dataset to assess their impact on price indices. We then impute missing values within the MICE-RF framework and compile price indices on the imputed datasets. Our approach is similar to that of Deng and Lumley (2024), who illustrate the superior performance of tree-based multiple imputation methods for a medical dataset.

Our findings reveal three key insights: (i) Real estate price indices are quite resilient to randomly missing data, because hedonic regression models can still capture price trends using the available observations. It is thus not necessary to impute missing data values when data are missing completely at random. (ii) Systematic missingness, on the other hand, can lead to substantial biases in real estate price indices. (iii) Multiple imputation with MICE-RF successfully corrects for biases in the price index, provided that the algorithm is given a sufficient number of relevant descriptive characteristics for each property. This is the case even when missingness is severely unbalanced.

We next examine a dataset of commercial office transactions, which, unlike the first dataset, contains genuine missing data, making it difficult to ascertain the true values behind these gaps. Missing data are the norm rather than the exception in Commercial real estate (CRE) datasets. This situation is problematic for index compilers because CRE markets are already characterized by greater heterogeneity and significantly lower transaction volumes than residential markets. Furthermore, the missingness in CRE data may be more systematic with some market segments being more reluctant to share property details than other segments. As a result, the type of property or buyer could influence the likelihood of missing data, introducing potential biases.

We find that multiple imputation via MICE-RF changes the trajectory of the Austrian office price index. While ignoring the data gaps (i.e., performing a complete-case analysis) leads to a positive price trajectory over the last four years, multiple imputation via the MICE framework changes the

estimate to a flat price index. The imputed price index aligns more closely with expert opinions, Austrian residential trends, and international CRE indices, reinforcing its credibility.

The remainder of the paper is structured as follows: In section 2, we discuss multiple imputation with MICE. In section 3, we experiment with transaction data for apartments in Vienna, Austria, between 2015 and 2023. We present how different types of missingness as well as subsequent imputations via MICE-RF affect the hedonic price indices compiled with these data. In section 4, we investigate how imputing missing values for office transaction data changes the trajectory of the resulting price index. We consider all office transactions in Austria between 2015 and 2023 for this task. Section 5 concludes the paper. We provide supplement material in the Appendices A to C.

2 Multiple Imputation via MICE-RF

2.1 Multiple Imputation in the Context of Real Estate Price Data

Imputation is a statistical technique that replaces missing values with plausible estimates that reflect what those values would have been if they had been observed (Rubin, 1987).[3] Traditional imputation techniques, such as mean imputation or regression imputation, replace missing values with a single estimated value. However, unless the data are missing completely at random, single imputation strategies will introduce bias into the imputed dataset and produce standard errors that are too low, leading to artificial certainty in the estimates and biased estimates and hypothesis tests (Rubin, 1978; Little and Rubin, 2019; Schafer, 1999; Honaker and King, 2010; Little et al., 2014). In real estate markets, every property is different and imputing missing values (e.g. size) using the mean can lead to serious biases in the resulting index unless data are missing-completely-at-random (MCAR). These issues with single imputation are particularly prominent when the proportion of missing data is large, or when missingness is not confined to one variable but affects multiple variables (Schafer, 1999).

To address these single imputation issues, multiple imputation introduces a probabilistic framework that better reflects the inherent uncertainty in missing data. The concept of multiple imputation goes back to Rubin (1978). The difference between single and multiple imputation is not one of algorithm complexity (similar algorithms can be used in single and multiple imputation) but one of the degree of certainty allotted to the imputed values. As articulated by Rubin (1978), "Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing)." In contrast, modern imputation methods such as maximum likelihood estimation, Bayesian estimation, and multiple imputation—adequately account

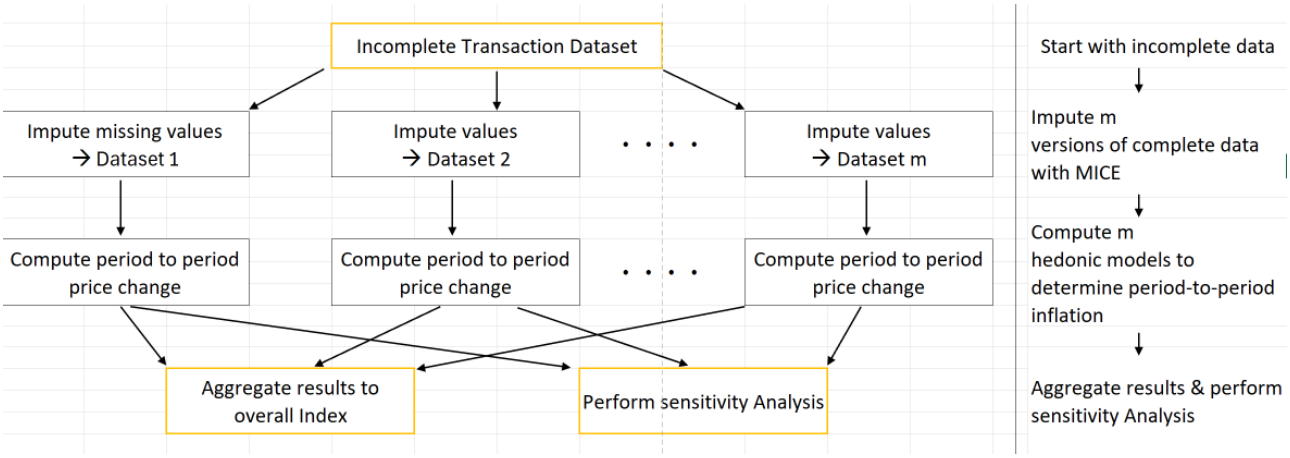
[3] We want to alert the reader that the term "imputation" can be confusing in the hedonic price index literature because there already exists a specific hedonic regression technique called "imputation method" (also known as the double imputation, double regression, or double prediction method) (OECD et al., 2013; Hill, 2013; Hill et al., 2018).

for uncertainty in imputed values.

As the name suggests, multiple imputation produces multiple predictions of each missing data point, which are presented in multiple "completed" datasets. The spread between the different imputation results allow us to capture the uncertainty around the imputed data point. Subsequent analysis is performed separately on each of the imputed datasets, before being aggregated into the final results and estimating test statistics. By pooling results from multiple imputed datasets (using Rubin's Rules) ensures that the uncertainty of imputation values is taken into account, preventing overly confident conclusions (Rubin, 1987).

Multiple-Imputation-by-Chained-Equations (MICE) is a widely used framework for handling missing data through multiple imputation (van Buuren and Groothuis-Oudshoorn, 2011; Azur et al., 2011). MICE is flexible and easy to implement, allowing users to specify both parametric and non-parametric estimation techniques for imputing missing values. The algorithm operates iteratively, imputing one variable at a time while conditioning on all other observed and previously imputed values (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2018). This sequential process continues until the imputed values stabilize, ensuring convergence. Once a single imputed dataset is generated, the process is repeated multiple times to create several imputed datasets. Each dataset is then analyzed separately, and the results are combined to account for imputation uncertainty. The variability across these datasets reflects the uncertainty introduced by missing data, providing a more robust statistical inference than single imputation or complete-case analysis. Figure 1 provides an overview of the MICE process, illustrating the sequential nature of imputations in our approach.

Figure 1: The MICE imputation framework



Currently the main alternatives to imputation via MICE is to engage in a joint Bayesian approach (Gelman et al., 2013; Bartlett et al., 2015; Erler et al., 2021). However, we do not consider these Bayesian approaches in this paper because (1) they may be too complicated to be implemented by some statistical agencies, (2) despite their differences, MICE imputations tend to produce similar

outcomes to those obtained using a joint Bayesian model approach, even in more complex hierarchical settings (Kunkel and Kaizar, 2017), (3) several aspects of MICE are grounded in Bayesian principles, even if the method is implemented in a computationally simpler, frequentist framework, (4) MICE is a widely accepted approach for data imputation in medical research and other related fields, and (5) it is easily implementable in R, which is the language many NSIs already use to compile their real estate price indices.

The MICE algorithm specifies the multivariate imputation model on a variable-by-variable basis using conditional densities, one for each variable with missing values (van Buuren, 2007; van Buuren and Groothuis-Oudshoorn, 2011). The algorithm imputes missing values through iterative rounds of prediction modelling, drawing plausible values from a distribution specifically formulated for each missing value van Buuren and Groothuis-Oudshoorn (2011). In each round, each variable with missing values is imputed, potentially using all other variables in the imputation dataset. This process should be run until the previously predicted and the newly predicted values converge.[4]

Following van Buuren and Groothuis-Oudshoorn (2011), the steps in the MICE algorithm are as follows:

- Let $Y = y_{i,j}$ be a data matrix , where $y_{i,j}$ represents the value of the j-th variable for the i-th observation (i.e., property transaction). The matrix Y is partitioned into observed (Y_{obs}) and missing (Y_{mis}) components.
- The goal of MICE is to impute Y_{mis} by estimating its conditional distribution given Y_{obs}.
- assume that our dataset Y has p variables: $Y = (Y_1, Y_2, Y_3, ...Y_p)$
- then for each variable Y_j with missing data, we define its conditional distribution given all other variables[5]: $P(Y_j|X_j, \theta_j)$. Here, $X_j = Y \setminus Y_j$ are predictors, and θ_j represent model parameters.
- The MICE imputation process is iterative:
  (a) Start with an initial guess for Y_{mis}.
  (b) For each iteration $t = 1, \ldots, T$, for each variable $j = 1, \ldots, p$, impute the missing values $\tilde{Y}_j^{mis,(t)}$ by sampling from the conditional posterior[6]:

$$\tilde{Y}_j^{mis,(t)} \sim P(Y_j|X_j^{(t-1)}, \hat{\theta}_j). \tag{1}$$

[4]Note that convergence usually happens quite fast, mainly because previously imputed values only enter newly imputed values via their relationship with other variables and not directly (van Buuren and Groothuis-Oudshoorn, 2011).

[5]Using all available variables to impute the missing values is the default setting. However it is also possible to specify a selection of the X_j that should be used to impute the missing values $\tilde{Y}_j^{mis,(t)}$

[6]Note that this step is similar to Bayesian posterior predictive sampling. However, while a fully Bayesian approach would sample θ_j from its posterior distribution before generating imputations, in MICE these θ_j values are typically treated as fixed (Erler, 2019).

(c) Update $X_j^{(t)} = (X_j^{(t-1)}, \tilde{Y}_j^{mis,(t)})$.
(d) Repeat until convergence (e.g., stabilization of imputed values).

MICE is a framework for imputation rather than a particular method. A separate estimation model can be specified for each variable with missing values. Thus, the user can adjust the conditional distributions in equation (1) to fit the type of each variable Y_j.[7] We used the random forest approach (method = "rf") to impute values for missing entries in both numeric and categorical variables.[8]

2.2 Random Forest

The Random Forest (RF) algorithm is a Machine Learning technique based on decision trees, first introduced by Breiman (2001). Ensemble learning methods, like Random Forest, leverage the combined power of multiple models to generate more accurate and robust predictions than individual models. The algorithm works by constructing many decision trees during training and aggregates their outputs to make predictions.[9]. Each decision tree in the forest is trained on a random subset of the data, selected through bootstrapping (sampling with replacement). Additionally, a random subset of features is considered for splitting at each node in the tree, which introduces some model variation between trees. These mechanisms – random data sampling and random feature selection – help reduce overfitting and increase the generalizabilty of the random forest model.

Random forests have been employed across various domains. Within the real estate field, random forest models have been shown to consistently outperform traditional linear regression models in house price estimation (Steurer et al., 2021; Ho et al., 2021). For these reasons, automatic valuation models typically rely on random forest and other decision tree methods to estimate property prices. Probably the main reason for the success of Random Forest and other tree-based algorithms within the automated valuation literature is their ability to handle non-linear dependencies and complex interactions among predictors. For example, while factors like location, size, and age of a property interact in ways that cannot be captured well by linear models, tree-based methods are able to capture these non-linear complex dependencies well *and* automatically. Thus, Random Forests do not require the definition of a functional form or the explicit selection of variables, making them particularly easy to apply even for users who are less familiar with advanced econometric techniques. Other benefits are that the ensemble nature of random forests ensures high robustness to noise and overfitting and that the randomness in the variable provision within the random forest algorithm ensures that there are slight model variations each time the algorithm is run, addressing the issue of model uncertainty

[7] Most commonly used regression and classification methods can be used within the MICE framework. Linear, logistic, and multinomial logistic regression are available as well as some semi- and non-parametric approaches like LASSO or Random Forest algorithms.

[8] We also tried regression methods but found that imputations were better with the random forest ("rf") method, especially as the number of included variables increases.

[9] Aggregation is done via majority voting (for classification tasks) or averaging (for regression tasks)

within the imputation process (Stekhoven and Bühlmann, 2011). Building on the advantages discussed above, we propose that Random Forests' ability to handle complex relationships can also be beneficial for estimating individual property characteristics based on observed prices and other attributes.

Several papers show that random forests and other tree-based algorithms in conjunction with multiple imputation outperform other imputation methods for missing values in a medical setting (Burgette and Reiter, 2010; Getz et al., 2023; Deng and Lumley, 2024). Outside the medical literature, there are only few empirical studies investigating how random forests compare with other algorithms within a multiple imputation setting. One of them is Schwerter et al. (2024), who find that Random Forest-based multiple imputation consistently perform better than non-tree based alternatives. We argue that the advantages of Random Forests in price predictions also apply when imputing missing values based on complex patterns of related variables. For this reason, we specify the random forest method as the imputation method for each variable with missing values.

2.3 Hedonic price index estimation

Most national statistical offices employ hedonic methods to compile property price indices.[10] The fundamental assumption of hedonic modeling is that the price of property n in period t, denoted as p_n^t, is a function of the K characteristics that describe it. In this paper, we employ the hedonic time dummy method, originally proposed by Court (1939) and further discussed in (de Haan and Diewert, 2013; Hill, 2013; Hill et al., 2018; Crawford and Neary, 2023).

Note, that we analyze each imputed dataset separately, i.e., we compute separate hedonic price indices for each of the m imputed datasets, which are then aggregated according to index aggregation rules as well as Rubin's Rules (Rubin, 1987).

Following the notation in de Haan and Diewert (2013) — and omitting the subscript indicating which of the m imputed datasets is considered — the hedonic time dummy regression equation is specified as follows:

$$\ln p_n^t = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k z_{nk}^t + \epsilon_n^t, \tag{2}$$

where β_0 denotes the intercept term, D_n^T is the time dummy matrix that indicates whether property n was sold in period t, and $z_n k^t$ describe the k characteristics present in property n that was sold in period t. β_k^t represent the coefficients on the characteristics and the δ^τ are the time dummy coefficients. ϵ_n^t is a random error term. The time dummy coefficients (δ^τ) measure how time affects the logarithm of

[10] In the United States, the repeat sales (RS) method is more commonly used. However, the RS method is not viable in most countries due to significantly lower turnover rates compared to the United States and other English-speaking countries. Additionally, hedonic methods offer several advantages over RS methods, including the absence of the need for index revisions as new data become available and greater robustness against sample selection bias (Gatzlaff and Haurin, 1997; Clapp and Giaccotto, 1999).

the price level. Thus, by exponentiating these time dummy coefficients, we can thus obtain a measure of quality-adjusted price inflation between the base period 0 and period t:

$$P^{0,t} = \exp(\hat{\delta}^t), \tag{3}$$

with $t \in 0, ..., T$.

2.4 Aggregation of Multiple Imputation Results

It may be intuitive, but incorrect, to derive an overall aggregated price index by averaging the price indices obtained from the m regression analyses. Instead, the appropriate approach is to aggregate the bilateral indices between adjacent periods (Eurostat, 2013). These bilateral price indices can be readily derived from the aggregate price index $P^{0,T}$. We illustrate the process below.

Reintroducing the subscripts to denote the m imputed datasets, the bilateral hedonic price index between periods $t-1$ and t is given by:

$$P_i^{t-1,t} = \frac{\exp(\hat{\delta}_i^{\,t})}{\exp(\hat{\delta}_i^{\,t-1})}, \tag{4}$$

where $i = 1, \ldots, m$ indicates the version of the imputed dataset used.

Following Rubin's Rules (Rubin, 1976, 1987; van Buuren, 2018), the combined bilateral price index after imputation is obtained by averaging across the m bilateral hedonic price indices:

$$P^{t-1,t} = \frac{1}{m} \sum_{i=1}^{m} P_i^{t-1,t}, \tag{5}$$

where $P_i^{t-1,t}$ represents the bilateral price index between periods $t-1$ and t, computed using imputed dataset i. This aggregation process is applied iteratively across all adjacent periods in the sample.

The overall aggregated price index from period 0 to t is then obtained through sequential multiplication:

$$P^t = P^{0,1} \times P^{1,2} \times \cdots \times P^{t-1,t}. \tag{6}$$

Beyond facilitating the combination of results from multiple imputations, *Rubin's Rules* provide a principled framework for valid statistical inference in the presence of missing data. These rules account for two key sources of variability: The first is *within-imputation variability*, which reflects the sampling variance within each imputed dataset and is computed by averaging the variances of the regression estimates across all imputations (Rubin, 1976, 1987). The second is *between-imputation variability*, which captures the variance of regression estimates across imputed datasets, quantifying

the uncertainty introduced by the missing data and the imputation process (Rubin, 1987). *Rubin's Rules* combine these two sources of uncertainty to derive the total variance of the pooled estimates, ensuring statistically robust inference (van Buuren, 2018).

2.5 A note on missingness patterns

Missing data mechanisms, which describe the underlying processes that generate missing values in a dataset, are typically classified into three categories introduced by Rubin (1976): *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). The implications of incomplete data, as well as the feasibility and effectiveness of imputation, depend on which of these mechanisms is responsible for the missingness in a given dataset.

MCAR occurs when the probability of a missing value is entirely independent of both observed and unobserved data. Under MCAR, complete case analysis remains unbiased, as the missing observations represent a random subset of the total dataset. However, the standard errors are typically larger because less information is used with a smaller sample sizes. In contrast, if the missing data is generated by a MAR or MNAR process, the resulting sample is no longer a simple random sample of the complete dataset, which means subsequent estimates may be biased. MAR holds when, conditional on the observed variables, the probability of a value being missing does not depend on the unobserved data (Rubin, 1976; Austin et al., 2021). In other words, MAR missingness occurs when the missing data mechanism can be fully explained by the observed data (Rubin, 1976). MNAR arises when missingness is neither random nor explainable through the observed data, meaning that the probability of missingness depends on the values that are themselves missing.

Like other multiple imputation techniques, *multiple imputation by chained equations* (MICE) assumes that missingness follows a MAR process (Rubin, 1976; Little and Rubin, 2019; Enders, 2022).[11] This assumption implies that after controlling for all available variables in the imputation model, any remaining missingness is effectively random. To maximize the imputation model's ability to uncover the latent relationships within the data, it is crucial to include as many potentially relevant variables as possible. This enhances the likelihood of identifying the missing data mechanism, as suggested by Rubin (1976). Moreover, incorporating additional external but relevant information can aid in transforming an MNAR dataset into one that follows a MAR structure (Enders, 2022).[12]

While Little (1988)'s MCAR test provides a means to assess whether missingness follows an MCAR

[11] van Buuren and Groothuis-Oudshoorn (2011) notes that while MICE can accommodate MNAR data, doing so requires additional modeling assumptions. Specifically, MICE cannot directly account for dependencies between missingness and unobserved values, which can result in biased estimates and an underestimation of imputation uncertainty.

[12] However, variable selection is generally necessary to ensure computational tractability. The *mice* R-package (van Buuren and Groothuis-Oudshoorn, 2011) now allows the use of penalized regression methods, such as LASSO, to address this issue.

process, it does not differentiate between MAR and MNAR mechanisms. In fact, despite extensive research efforts, no conclusive diagnostic methods currently exist for distinguishing MAR from MNAR (Enders, 2022).

Although the inability to definitively determine whether missingness follows a MAR or MNAR mechanism can pose challenges in some applications, this issue is less problematic in the context of real estate transaction data. Unless the set of available variables is severely limited, it is unlikely that missingness follows an MNAR process. This is because property characteristics tend to exhibit strong correlations with one another and with transaction prices. As first noted by Rosen (1974), this interdependence is the foundation of hedonic regression models. If the advertised price of a property is too high given its set of characteristics, it simply won't sell; if the price is too low, the potential seller will be inundated with hopeful buyers and will raise the price. Either way, the final transaction price will be pushed towards a level consistent with the current market situation.[13]

3 Application 1: An apartment price index for Vienna, Austria

To assess the performance of MICE-RF in handling missing data in real estate indices, we conducted a series of simulations under different missing data mechanisms. The primary goal is to examine whether MICE can effectively reconstruct missing observations and produce indices that closely approximate those derived from complete data. We introduce missingness in all variables except for the price and date variables.[14]

3.1 A price index with the complete data

To analyze the impact of missing data on price indices and to demonstrate the effects of multiple imputation using MICE, we utilize a dataset of apartment price transactions in Vienna spanning from January 2015 to September 2023. This dataset, provided by *ZTdatenforum*, comprises 67,292 transactions, each characterized by a diverse set of variables.

For this analysis, we focus on a carefully selected subset of descriptive variables that collectively

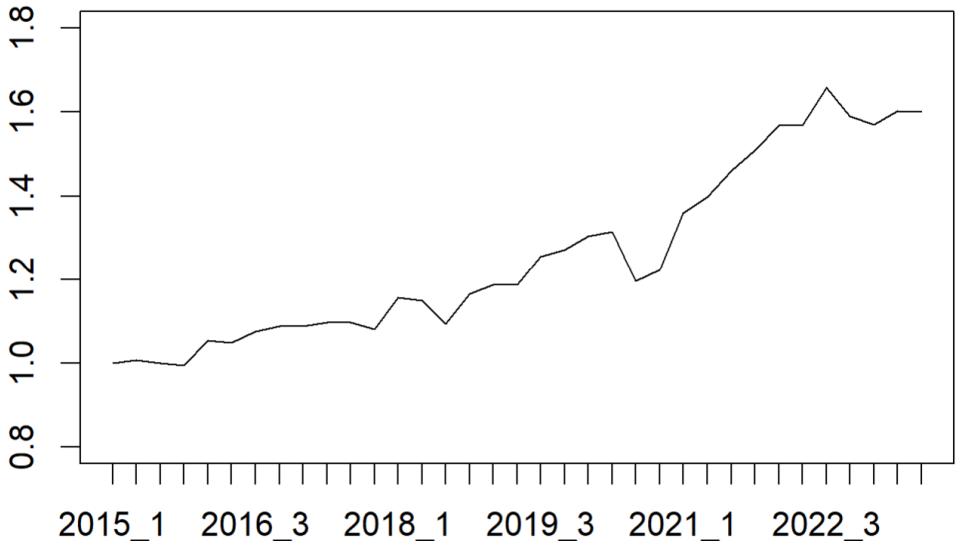
[13] This observation also underscores the general superiority of transaction data over advertised prices from online platforms. Online listing prices often include speculative pricing strategies, with some sellers deliberately setting low prices to attract interest, while others list properties at unrealistically high prices.

[14] There are two reasons for excluding these values from the imputation exercise. First, price and date are the two variables that are always available in any transaction dataset as they are crucial to prove ownership of a property. Second, while MICE is flexible and can handle missing data in independent variables (van Buuren and Groothuis-Oudshoorn, 2011), there is a general view in the imputation literature that the imputation of missing values in the dependent variables should be avoided as it can distort the relationship between predictors and dependent variables. In the context of MICE imputation, there is also the issue that imputation of predictor variables can lead to a breakdown of Rubin's Rules (van Buuren, 2018).

capture key aspects of a property's location, construction quality, and size—factors widely regarded as the primary determinants of property prices (OECD et al., 2013). An additional criterion for selecting these seven variables is their minimal proportion of missing values. The selected variables include the property's purchasing price (*price*), the transaction period (*Quarteryear*), the internal size in square meters (*size*), postcode dummies (*PLZ*), an indicator for whether the property was sold by a builder or construction firm (*builder*), a variable denoting whether a parking spot price was mentioned in the contract (*car_price*), and the distances (in meters) to the nearest pharmacy (*pharm_dist*) and doctor's surgery (*doc_dist*).[15][16]

We estimate a log-linear hedonic regression model using the selected variables, following the specification outlined in equation (2). The resulting time-dummy hedonic price index is denoted as P_{orig}. It is displayed as the solid black line in Figure 2 and all subsequent figures in this section. P_{orig} serves as the benchmark index for comparison with alternative indices derived from the same hedonic regression model but applied to subsets of the dataset that simulate various missingness patterns. The index reveals that inflation remained positive for most of the study period, with the exception of 2020, when Austria experienced multiple COVID-19-related lockdowns. Additionally, the price trend exhibits a noticeable flattening beginning in mid-2022.

Figure 2: Full sample price index for apartments 2015-2023



[15] Distance variables serve as useful proxies for a property's degree of urbanity and accessibility.

[16] To maintain analytical clarity, we apply only minimal data cleaning, excluding transactions between close relatives, sales following bankruptcies, and apartments outside the size range of 20 to 400 square meters.

3.2 Introducing different types of missingness

We now introduce different missing data patterns into the complete transaction dataset and investigate their effects on the hedonic price indices compiled on these sub-samples. We consider three broad cases: (1) random missingness across all variables, (2) non-random missingness across all variables, and (3) non-random missingness concentrated on one variable.

Case 1: Random missingness

To produce samples with random missingness we randomly delete the same proportion of values for each variable, except for the price and date variables. We introduce between 10 and 50 percent missing values in this way and label them *MCAR10* to *MCAR50* (with the last two digits denoting the degree of missingness introduced).[17] Under this scenario, a single observation can miss none, one, or many entries. As the missingness is introduced independently for each variable, there is no linkage of missingness across variables. Performing Little's MCAR test on these data samples we indeed find that we cannot reject the Null hypothesis that the data are generated via an MCAR process (see Table 1).

Case 2: Non-random missingness

Next we produce samples of data afflicted by non-random missingness in all variables but price and date. To make the situation as difficult as possible, we follow an algorithm developed by de Sousa and Mayer (2021) to produce MNAR missingness. This algorithm produces missing values on a variable-by-variable basis, but this time the missingness patterns are correlated across variables. One effect of this missingness process is that it leaves us with more complete rows than the random process we used in case 1 (shown in Table 1). As before, we introduce missingness of between 10 to 50 percent. We label these datasets as nonMCAR10 to nonMCAR50, with the last two digits indicating the percent of missing values. However, even though the algorithm is designed to produce MNAR-type missingness, it is more likely that the missingness produced in this way is of the MAR type.[18] In any way, Little (1988)'s MCAR test shows that we can clearly reject the Null hypothesis that the data are generated via an MCAR-type missingness process Table 6.

Case 3: Concentrated missingness in one variable

For case 3, we assume that only one variable, size, is affected by missingness. Size is obviously an important variable to describe any property and we want to see how our price indices behave this variable is unavailable for large parts of the transactions. Not only do we remove large percentages of the size information from our datasets, we also do so in a highly concentrated way. To produce

[17] These percentages of missing data are far higher than typically observed in actual datasets.

[18] This is the case because of the correlation that tends to occur between real estate characteristics. Just as expensive properties tend to perform well on a number of characteristics, cheap properties tend to perform badly on a number of characteristics. This correlation between property characteristics make transaction data a good candidate for imputation, which generally assumes MAR-type missingness.

our case 3 sub-samples, we first rank all properties based on their purchase prices and then remove the size information from either the top or the bottom of this distribution. In this context, the sample *top75* refers to a dataset where the top 75 percent of properties (in terms of purchase price) contain information on size while the remaining 25 percent do not. Similarly, *top50* and *top25* indicate that size data is available only for the top 50% and top 25% of properties, respectively. Conversely, *bottom75*, *bottom50*, and *bottom25* indicate that size information is available only for the bottom 75, 50, or 25 percent of the price distribution. Our other reason for investigating the effects of missing size information is that this situation closely resembles the missingness structure in our second application (that of office transactions in Austria) where we find that internal size of the property is the variable missing most often.

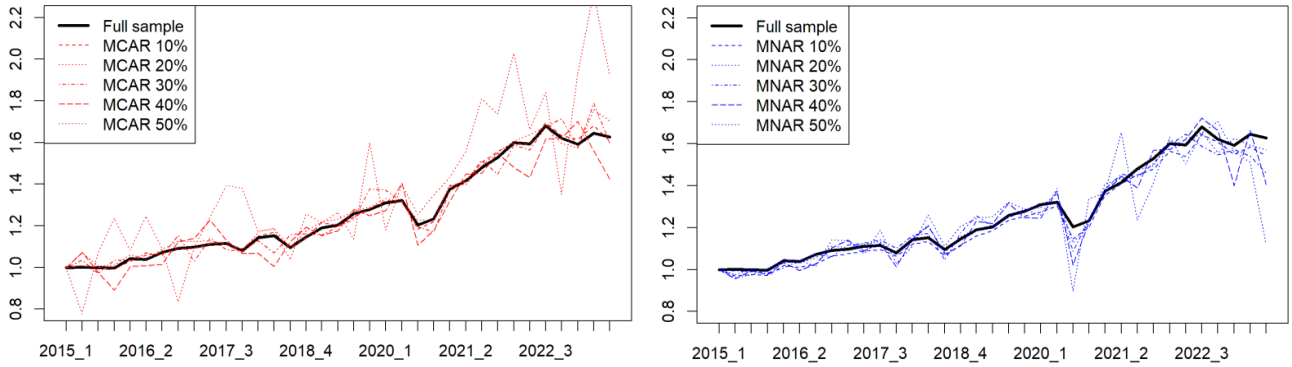
3.3 Price indices when missingness is ignored

The usual treatment of missingness by NSIs is to engage in a *complete case analysis*, i.e., to remove all transactions for which data are missing in the core variables. Figure 3(a) illustrates the impact of randomly deleting values from the original transaction dataset according to our case 1. Figure 3(b) illustrates the impact of the non-random deletion of values according to case 2. The first thing to notice in these figures is that hedonic price indices perform well given the severe degrees of missingness we introduced. However, we see some differences between the MCAR and non-MCAR samples.

The samples with MCAR missingness do not show any signs of systematic bias. This finding is consistent with prior simulation studies in other domains, which indicate that the proportion of missing data is not a significant predictor of bias, as long as the data are genuinely missing completely at random (MCAR) (Madley-Dowd et al., 2019).

In contrast, the samples with non-MCAR-type missingness exhibit some negative bias in the price indices. However, these samples are less affected by volatility than the MCAR samples. The reduced volatility in the non-MCAR samples can be attributed to larger sample sizes, resulting from our non-MCAR data-generating process, which concentrates missingness on some observations.

Figure 3: Price indices for Vienna (2015-Q1 to 2023-Q3) when values are (a) randomly deleted, or (b) non-randomly deleted



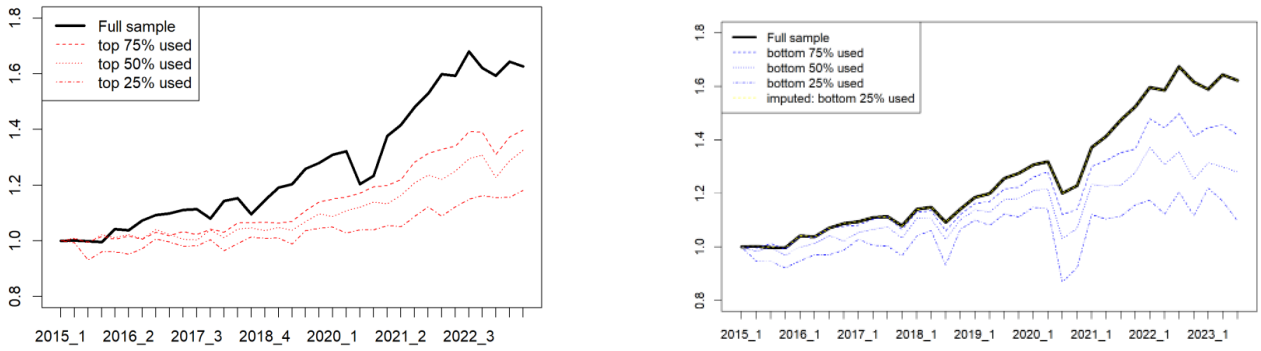
The original price index derived from the full sample is indicated in black, while indices compiled with the missingness subsamples are indicated in red or blue. **Figure (a)** illustrates the effect of deleting individual values according to a MCAR data generating process. Between 10 to 50 percent of all entries are deleted in this way. Note that with this data generating process, together with complete-case analysis- means that indices are compiled with rapidly falling sample sizes. For example, our sample with 50 percent missing in each variable has only 241 complete observations over 34 periods (on average 7 per quarter). Given these low observation numbers, it is remarkable that the subsample indices follow the original index quite well. **Figure (b)** illustrates the price indices when between 10 to 50 percent of the values for each variable are missing according to the MNAR data generating process discussed above. Compared with the MCAR cases, these price indices are less prone to volatility but show a slight downward bias. The lower volatility compared with the MCAR samples is an effect of the data generating process which introduces more clustered missingness than in the MCAR case and therefore a much larger number of complete observations remaining in the sample.

Systematic missingness has a significant impact on price indices. To examine this effect, we estimate the same log-linear hedonic regression model for each of the Case-3 subsets as previously applied to the full dataset (i.e., without artificially introduced missing data). The resulting price indices, compiled from these Case-3 subsets, are presented in Figure 4(a) and (b). As complete-case analysis (listwise deletion) removes all observations with missing values, in the case of the Case-3 subsets, we are left only with either the highest or the lowest segments of the market for index construction. Unsurprisingly, this systematic deletion introduces substantial bias into the resulting indices.

A disaggregated analysis of the upper and lower market segments reveals divergent price trends during the COVID-19 lockdown period in 2020. Lower-priced properties experienced significant price declines during this period, whereas higher-priced properties remained largely unaffected. We hypothesize that this pattern results from the interplay of two distinct market dynamics. First, investor demand for lower-priced properties appears to have weakened. Second, owner-occupiers increasingly sought high-quality properties, potentially driven by the realization—amid the COVID-19 crisis—that they preferred larger or higher-quality living spaces.

Another notable finding is that excluding both the upper and lower market segments results in indices that are systematically downward biased relative to the full-sample index, P_{orig}. While this outcome may seem counterintuitive, it arises from the method used to define our samples. Instead of excluding the most or least expensive properties on a period-by-period basis, we classified transactions based on the 25th, 50th, and 75th quantiles of the entire price distribution from 2015 to 2023. This approach introduces compositional distortions. In a market experiencing a general upward price trend, fewer early-period transactions and a greater number of later-period transactions fall within the top 25th, 50th, or 75th percentiles. Similarly, fewer properties remain below the thresholds for classification into the bottom 25th, 50th, or 75th percentiles over time. As a result, samples from later periods tend to represent properties of lower average quality than those from earlier periods. Consequently, price indices computed from these truncated datasets conflate changes in property quality with the effects of time, leading to a systematic downward bias.[19]

Figure 4: Price indices for Vienna (2015-Q1 to 2023-Q3) with either the bottom or the top end of the price distribution



Note: Significant bias is apparent when only the bottom or top end of the price distribution is available for index compilation.

3.3.1 Index Volatility and Bias measures

Each missingness scenario affects index bias and volatility differently. To evaluate the performance of indices compiled under different missingness conditions relative to the original index, P_{orig}, we introduce the following measures of bias and volatility.

To quantify the bias of subindex k, we compute its average deviation from the original index, P_{orig},

[19] Note, we are in the process of changing from this overall classification of price percentiles towards a period-by-period classification, but ran out of time to include these new estimations here.

over the last four periods:

$$Bias_k = \frac{1}{4} \sum_{t=T-4}^{T} (P_{orig}^t - P_k^t), \tag{7}$$

where T denotes the final period in the dataset, P_{orig}^t represents the value of the full-data price index in period t, and P_k^t is the corresponding value of the price index estimated under missingness scenario k.

Volatility measures the extent of fluctuations in an index. Lower volatility is generally preferred, as excessive fluctuations can obscure underlying price trends. However, higher volatility does not necessarily indicate poor index quality; in turbulent markets, price indices should reflect inherent market fluctuations. Nonetheless, excessive volatility may arise if the hedonic regression model fails to adequately control for variations in property quality across periods. We quantify the volatility of index k as the standard deviation of the log-price change between consecutive periods:

$$Vol_k = \text{sd}\left[\ln\left(\frac{P_k^{t+1}}{P_k^t}\right)\right]. \tag{8}$$

The bias and volatility measures for each missingness scenario are reported in Table 1. Case 1, representing MCAR-type missingness, exhibits minimal bias, even when a large proportion of data is missing. In contrast, indices compiled from Case 2 (i.e., non-MCAR) samples display greater bias than those from Case 1, with bias increasing as the proportion of missing values rises.

The most pronounced biases are observed in Case 3 indices. Volatility generally increases as the proportion of missing values rises across all three missingness scenarios. However, indices based on samples concentrated at the upper end of the market exhibit significantly lower volatility. We attribute this finding to the relatively homogeneous quality and pricing of newly built apartments, which likely reduces fluctuations in the estimated price indices.

3.4 The Impact of Imputing missing values with MICE-RF

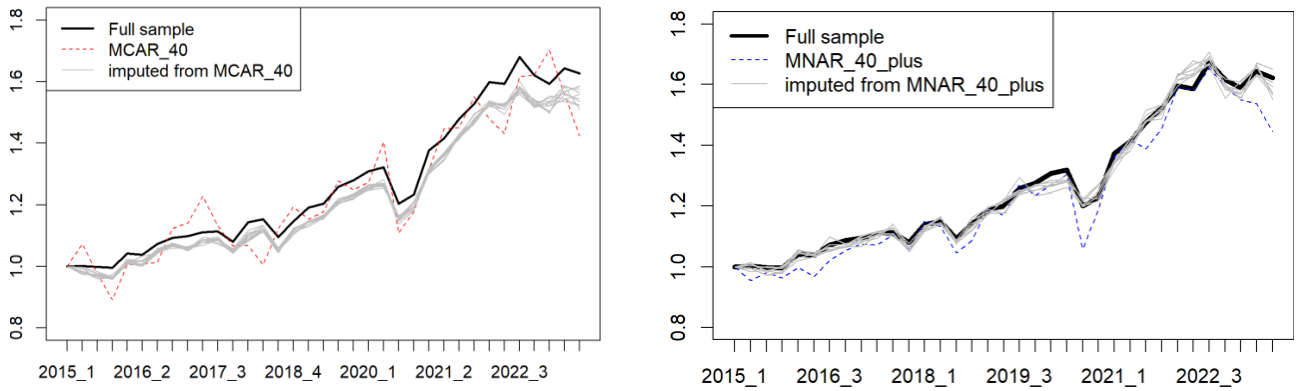
Our next step is to address missing data by applying multiple imputation using the R package *mice* (van Buuren and Groothuis-Oudshoorn, 2011) in combination with the Random Forest algorithm ($method = rf$) to impute all variables containing missing values. We set the number of imputed datasets to $m = 10$, as increasing this value (e.g., to $m = 20$) did not enhance outcomes but significantly increased computational time. The optimal number of imputations remains a topic of debate. Rubin (1987) suggests that 2–5 imputations are generally sufficient, a view supported by Schafer and Olsen (1998), who recommend 3–5. In contrast, Graham et al. (2007) proposes that the number of imputations should match the percentage of missing data.

Case 1 and 2:
Our findings align with expectations from the literature: multiple imputation is most effective under missing at random (MAR) missingness, as it allows patterns in observed data to inform imputation (Rubin, 1976). Even though we tried to introduce MNAR missingness, market forces establish a connection between the characteristics of a property and its price (Rosen, 1974), so that the MNAR cases exhibit MAR-like structures, which allow imputations to work (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2018). In contrast, MCAR data lack such structure, making imputation less reliable.

Our findings align with existing literature: multiple imputation is most effective when data are missing at random (MAR), as it leverages patterns in observed data to inform imputations (Rubin, 1976). Although we attempted to introduce MNAR missingness, market dynamics create a structural link between property characteristics and prices (Rosen, 1974), causing even MNAR cases to exhibit MAR-like patterns. This structural dependency facilitates the success of multiple imputation in real estate data. The patterns observed in Figure 5 extend to our other sub-samples.

As shown in Table 1, MICE-RF imputation reduces bias more effectively in sub-samples with non-random missingness than in MCAR-type samples. However, imputation consistently reduces index volatility, with a more pronounced effect as missingness increases. This suggests that the decision to apply imputation when data exhibit MCAR missingness should depend on (1) the degree of missingness and its effect on volatility and (2) the intended use of the index. If excessive volatility obscures underlying trends, MICE-RF imputation may be justified despite introducing potential bias. However, imputing MCAR-type missing data may create a false sense of precision, as any bias introduced by the random missingness process persists in the final results.

Figure 5: Dxxx



In Figure 5**(a)**, the red dotted line represents the price index derived from a random sub-sample where 40% of each variable's values are missing completely at random (MCAR). Without imputation, only 1,764 observations remain from the original 67,292 transactions, resulting in a more volatile and downward-biased index. The gray lines depict indices estimated from each of the 10 MICE-imputed datasets using Random Forest imputation. For MCAR missingness, imputation effectively smooths the index trajectory. However, since missing values are randomly distributed, the imputed index retains the inherent (randomly-introduced) bias of the complete-case index. Figure 5**(b)** illustrates the impact of MNAR missingness. The blue dotted line corresponds to the index constructed from a sub-sample where 40% of each variable's values are missing according to an MNAR data-generating process. A key observation is that the complete-case index (i.e., the blue dotted line) exhibits lower volatility than the index derived from the MCAR sample. This difference arises because missingness in the MNAR scenario is correlated across variables, resulting in a larger retained sample size (6,462 complete cases remain available for index computation). Another important finding is that imputation using MICE-RF successfully restores the trajectory of the original price index, demonstrating its effectiveness in handling structured missingness patterns.

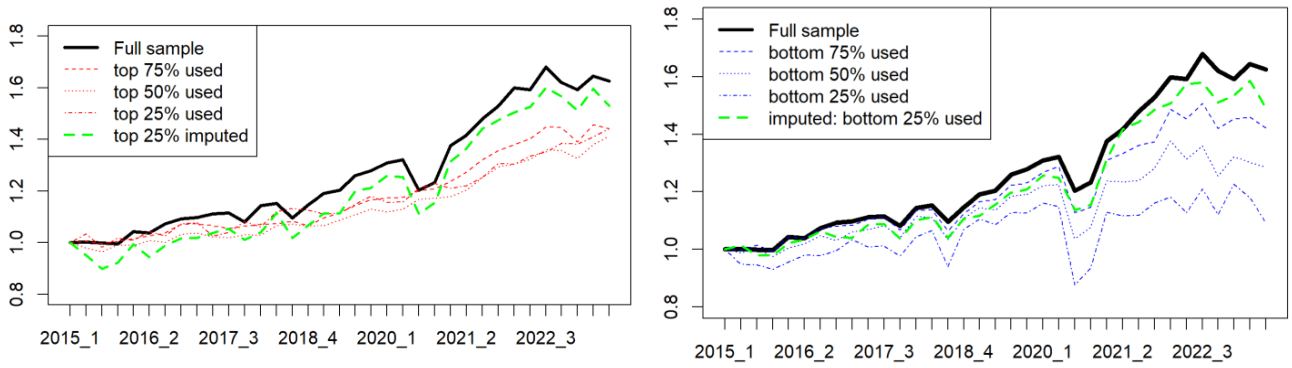
Case 3: Top and Bottom Quartiles

The green lines in Figure 6(a) and (b) depict the price indices obtained after applying MICE-RF imputation to the $top_2 5$ (left) and $bottom_2 5$ (right) subsamples. In the absence of imputation, complete-case analysis produces the lowest observed indices in Figure 2 (red for (a), blue for (b)). Although some bias persists in the imputed indices, they align significantly more closely with the full-sample index, P_{orig}, compared to indices computed without imputation.

These findings are highly promising. Even after removing 75% of size-related information from both high- and low-priced transactions and supplying MICE-RF with only seven weakly correlated variables, the imputed indices closely follow the trajectory of the full-sample index.[20] As shown in Table 1, cases with 50% or 25% missing size information yield results that are even more closely aligned with P_{orig}.

[20] Incorporating additional variables, such as plot size, further improves imputation performance.

Figure 6: Price indices for Vienna (2015-Q1 to 2023-Q3) with either the bottom or the top end of the price distribution



There are significant deviation from the full sample index when only the bottom or top end of the price distribution is available for index compilation. Imputing the missing values with MICE can dramatically improve the results. The green lines indicate the price indices compiled with the imputed datasets when only 25 percent of the bottom or top of the price distribution were available.

Summary table of imputation results

The results are summarized in Table 1. The first two columns report the test statistics and p-values from the MCAR test proposed by Little (1988). A high test statistic combined with a low p-value indicates non-random missingness. For subsets with MCAR missingness, the test fails to reject the null hypothesis, confirming that missing values were introduced in a completely random manner. Conversely, for all other data samples, the test strongly rejects the null hypothesis, providing evidence of non-random missingness.

In line with Rubin (1976), our findings indicate that imputing missing values is neither necessary nor inherently beneficial when data exhibit MCAR-type missingness. Since MCAR missingness does not introduce systematic bias into price indices, applying imputation in such cases may reduce accuracy rather than improve it. However, MICE-RF imputation does reduce index volatility, which may be desirable in certain applications. Thus, we recommend performing Little (1988)'s MCAR test before determining whether imputation is appropriate. If missingness is classified as MCAR, users face a trade-off between bias and volatility in index construction.

In summary, MICE-RF imputation is a powerful tool for handling missing data in real estate price indices. However, its application should be guided by a careful assessment of the underlying missing data mechanism to ensure that imputation enhances, rather than distorts, index accuracy.

Table 1: Effects of Missingness Patterns and MICE-RF Imputation

Missingness Type	Little's MCAR Test		N		Bias		Volatility	
	Statistic	p-value	Sample	Imputed	Sample	Imputed	Sample	Imputed
Case 1: Random Missingness								
MCAR10	557	0.956	32,246	67,292	1.002	0.986	0.037	0.036
MCAR20	692	0.761	14,059	67,292	1.013	0.976	0.044	0.036
MCAR30	699	0.447	5,413	67,292	1.065	0.954	0.042	0.038
MCAR40	690	0.545	1,764	67,292	0.974	0.943	0.061	0.037
MCAR50	681	0.644	499	67,292	1.168	0.915	0.070	0.038
Case 2: Non-Random Missingness								
nonMCAR10	41,417	0	36,427	67,292	0.952	0.997	0.043	0.036
nonMCAR20	38,371	0	22,830	67,292	0.946	0.997	0.052	0.038
nonMCAR30	35,916	0	12,697	67,292	0.907	0.963	0.059	0.038
nonMCAR40	32,025	0	6,462	67,292	0.881	0.956	0.078	0.038
nonMCAR50	27,609	0	2,826	67,292	0.926	0.931	0.098	0.037
Case 3: Concentrated Missingness in Size Variable								
top25	23,232	0	50,829	67,292	0.893	0.912	0.025	0.047
top50	17,072	0	33,657	67,292	0.861	0.982	0.017	0.050
top75	11,991	0	16,823	67,292	0.893	1.010	0.017	0.049
bottom25	11,611	0	50,469	67,292	0.731	0.915	0.073	0.047
bottom50	17,055	0	33,700	67,292	0.815	0.973	0.050	0.046
bottom75	23,232	0	16,891	67,292	0.901	1.019	0.044	0.042

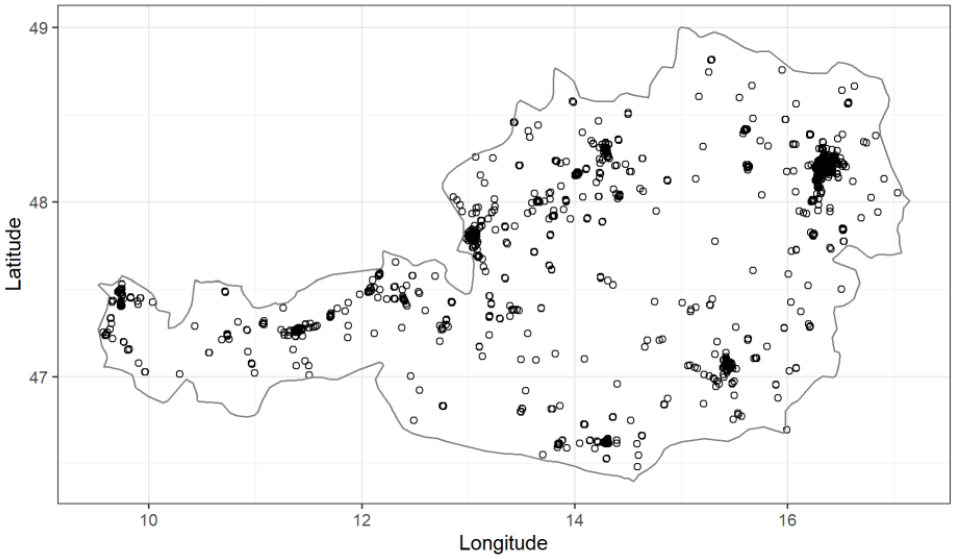
Note:
(i) In Little's MCAR test, a p-value close to zero (below 0.05), combined with a high test statistic, indicates that the null hypothesis—stating that the data are missing completely at random (MCAR)—can be rejected. (ii) Bias is measured relative to 1, meaning that an unbiased index receives a bias value of exactly 1. A bias value greater than 1 indicates a positive bias, whereas a value below 1 signifies a negative bias. The results show that imputing missing data before index compilation generally reduces bias in the price index—except in cases where missingness follows an MCAR pattern. (iii) Across all considered cases, imputing missing values with MICE-RF prior to index compilation consistently reduces the volatility of the resulting price index.

4 Application 2: A Commercial Office Price Index for Austria

4.1 Data

The transaction data for our second application also come from ZT*datenforum* and cover all office transactions for Austria between January 2015 and September 2024. These transactions are geographically distributed across the entire country. Figure 7 illustrates the longitude and latitude values of these transactions.

Figure 7: Geo-locations of office transactions in Austria between 20xx and 2024



Location of all office transactions between 2015-2024

A key challenge in this dataset is the high percentage of missing values for key property characteristics, particularly *size* and *legal_age*, which are absent in 1549 and 896 cases out of 3448 observations, respectively. Such missingness is common in commercial real estate (CRE) datasets, where property owners and tax offices are not required to report optional characteristics.[21]

To construct a hedonic price index using the Time Dummy method, we estimate a regression model where log-price is the dependent variable and property characteristics, including time dummies, serve as explanatory variables. Ideally, each property should be described by a comprehensive set of attributes to control for quality differences. When these quality adjustments function effectively, time dummy coefficients capture price changes attributable to temporal effects, which we use to derive the price index.

[21] Here, *optional* refers to the perspective of property owners and tax authorities rather than index compilers.

Table 2: Overview of Key Variables in the Dataset

Variable	Description	Type
	Physical Property Characteristics	
size	Interior size in square meters.	Size
legal_age	Years since official registration; may be lower than actual age for older properties.	Quality
builder	Indicates whether the property was sold by a construction firm (new or renovated).	Quality
hadPkwApPrice	Indicates whether a parking spot or garage was included in the sale.	Attributes
hadInventoryPrice	Indicates whether the purchase contract included inventory.	Attributes
	Locational Variables	
province	One of Austria's nine administrative regions, including Vienna.	Geographic
PB_number	The political district where the property is located (80 total).	Location
city dummies	Regional indicators for Austria's four largest cities (Vienna, Graz, Salzburg, Linz).	Location
postcodes	Austria's 2200 postcode areas (mandatory field).	Location
longitude, latitude	Precise geolocation of the property.	Location
	Distance-Related Variables	
shops1000	Number of shops within a 1000-meter radius (proxy for urbanity and amenities).	Accessibility
doctor_dist	Distance to the nearest doctor's office (proxy for urban accessibility).	Accessibility
citydistance	Distance to the nearest regional capital.	Accessibility

Notes:
Additional variables, such as building age, were available but omitted due to high missing rates. Most of the variables presented here exhibit negligible missingness; however, size and legal age are missing in 1547 and 895 cases out of 3438 observations, respectively.

4.2 Missingness patterns

We analyze the missingness patterns in our office dataset by repeatedly applying Little's MCAR test to subsamples of the data. Table 3 shows that the p-values from Little's MCAR test decrease as more variables are included in the analysis. This implies that as more explanatory variables are incorporated, the likelihood of MCAR missingness diminishes. This is an encouraging result, as multiple imputation

methods perform optimally when data are missing in a manner that is not completely at random (non-MCAR). Furthermore, as the number of variables included in the dataset increases, so does the number of identified missing data patterns, which –in turn– enhances the accuracy of multiple imputation.

Although the MCAR test cannot differentiate between MAR and MNAR missingness, the strong correlations between property attributes and price in real estate markets suggest that missingness is unlikely to depend directly on unobserved values (i.e., be of the MNAR type). Thus, naturally occurring missing values in real estate markets are more likely to exhibit MAR-type missingness than MNAR-type missingness (see Rosen (1974) and also discussion in the previous section).

Table 3: MCAR Test for Office Data Subsets

MCAR Test	Statistic	df	P-value	Missing Patterns
price+quarter+nearest-city	1.58	2	0.455	2
price+quarter+size	4.36	2	0.110	2
price+quarter+province+size	15.9	3	0.00117	2
price+quarter+province+size+ nearest-city	34.2	8	0.0000369	3
+more variables	8815	1642	0.000	87

Notes:
The table illustrates the results of Little (1988)'s MCAR test on our office transaction dataset for Austria. We find that as the number of variables included in the test increases, the probability of MCAR missingness diminishes.

4.2.1 Closer examination of the missingness patterns

As size is the variable most frequently missing, we first examine whether its missingness is systematically linked to price. We find that log-price distributions for properties with and without size entries show basically identical distributions in the overall dataset (see Figure 17(a) in the Appendix C). Similarly, Figure 17(b) confirms that properties across the entire price range exhibit missingness, except for four high-value properties. This suggests that missingness is not systematically linked to market value. The price distributions remain similar for properties with and without size data, when looking at the regional province level. Facet plots illustrating this point are presented in Figure 18 in Appendix C.

We next investigate geographic patterns in missingness. Across Austria's nine regions, price distributions remain largely similar, with the exception of Lower Austria. However, postcode-level analysis (Figure 19) reveals significant variation. Larger cities exhibit mixed missingness patterns due to higher notary density, whereas rural areas tend to either fully record or entirely omit size data. These

differences likely stem from variations in notarial recording practices across Austria's 405 notaries.

Table 4 quantifies missingness by province. The proportion of contracts without size information ranges from 30.4% in Burgenland to 57.7% in Lower Austria.

Table 4: Missing Data by Province

Province	Variable	Pct_Miss
B	Size	30.4
ST	Size	32.3
T	Size	39.0
K	Size	39.2
S	Size	39.7
O	Size	44.0
W	Size	50.2
V	Size	56.9
N	Size	57.7

Note: The proportion of contracts missing size information varies by province, with the highest rates in Lower Austria (57.7%) and the lowest in Burgenland (30.4%).

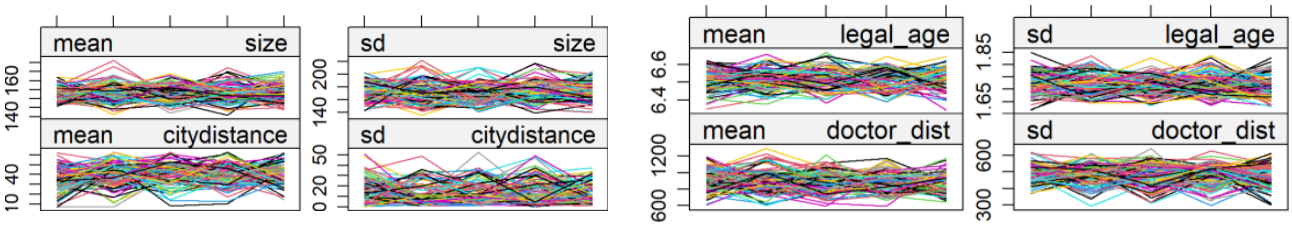
Overall, missing size data is not systematically related to property price, but it varies geographically. Larger cities exhibit mixed missingness. For example, in Vienna, size is missing in about 50% of cases, and similarly mixed size information is available in the other Austrian cities. On the other hand, in rural areas, size data tends to be either consistently recorded or consistently absent, likely reflecting local notarial practices. Figure 19 in Appendix C illustrates this pattern.

4.3 Imputing missing values with MICE

We again use MICE-RF to impute missing values (van Buuren and Groothuis-Oudshoorn, 2011). We chose a high number of imputations ($m = 100$) to illustrate the consistency of our results. We set the number of iterations to the default value within in the MICE framework (i.e., $maxit = 5$). The number of trees grown in each random forest is set to 20 (the default value is 10). All variables included in Figure 16 in Appendix C are included in this imputation analysis.

According to van Buuren and Groothuis-Oudshoorn (2011), convergence in multiple imputation is achieved when the streams of imputed values mix freely, showing no discernible patterns or trends over iterations. This behavior indicates that the imputation model has stabilized and is generating values that accurately reflect the underlying distribution of the data. The diagnostic plots in Figure 8 confirm convergence according to these criteria.

Figure 8: Imputation with mice: stream conversions of office dataset imputations



Comparing how parameter values change along the iterations is a good way to see if convergence tendencies are present. The figures above show the mean and standard deviations of the imputations for each of the 50 parallel imputation streams. We can conclude that convergence is present as the variance between the different streams is not greater than the variance within each stream van Buuren and Groothuis-Oudshoorn (2011)

4.4 Office price index for Austria 2015-2024

Following imputation, we estimate hedonic models separately for each imputed dataset according to Equation 2. We include the following variables in the hedonic regresson model: *log(price), size, citydistance, nearest_city, legal_age, builder, doctor_dist, Year, Quarteryear, province, Vie_dummy, Graz_dummy, Salzburg_dummy, Linz_dummy, noCity_dummy, shops1000, hadPkwApPreis*, and *PB_number*. Note that these are fewer variables than we used in the MICE-RF imputation process.

Table 5 presents a summary of the pooled regression coefficients across the 100 models, along with their associated standard errors, p-values, and confidence intervals. While the district dummies (*PB_number*) are statistically significant, their large number precludes individual reporting in the table (see Appendix for the full set of coefficients). The coefficient estimates in Table 5 reflect the joint regression outcome of the 100 imputed datasets, while the confidence intervals account for both within-imputation and between-imputation variability, following Rubin's Rules (Rubin, 1976, 1987).

Rather than deriving the aggregate price index from the joint regression results in Table 5, we construct it step by step using bilateral price indices, as described in Equations 4 to 6. To do so, we first compile bilateral price indices for each imputed dataset. Figure 9 illustrates these bilateral price indices, while Figure 10 illustrates the spread of these period-by-period index values for the 100 imputations. The results demonstrate a high degree of consistency across imputations, although the index values for 2023-2024 exhibit slightly greater variation. The final aggregate index is obtained by averaging the bilateral indices across imputations following Equation 5. According to Rubin's Rules (Rubin, 1976, 1987), each bilateral index is given equal weight in the aggregation. The overall aggregate index is then derived from the annual price changes, as described in Equation 6.

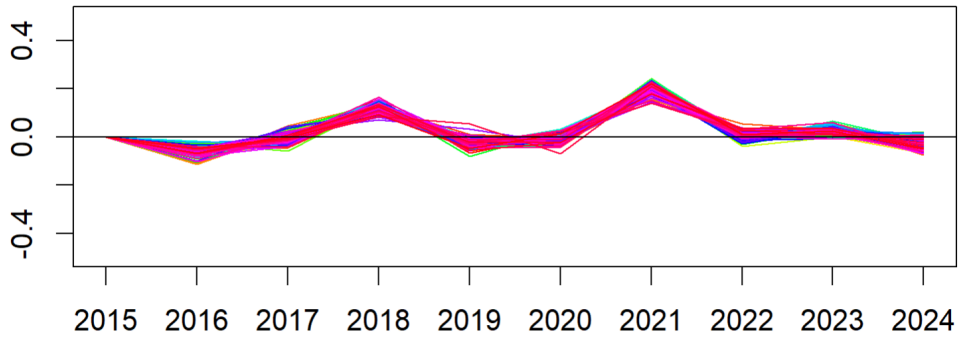
This similarity among the individual indices provides us with a high degree of confidence in the overall aggregate index. Figure 10 provides an alternative illustration for the closeness of individual results.

Table 5: Summary Joint Regression Output of 100 Hedonic Regressions

Term	Estimate	Std. Error	P-value	95% CI
(Intercept)	10.930	0.331	0.000	[10.281, 11.579]
size	0.002	0.000	0.000	[0.002, 0.002]
log(citydistance)	-0.109	0.033	0.001	[-0.173, -0.045]
builder	0.168	0.033	0.000	[0.104, 0.231]
doctor_dist	0.000	0.000	0.014	[0.000, 0.000]
hadPkwApPreisTRUE	0.316	0.039	0.000	[0.240, 0.393]
Year2016	-0.062	0.062	0.315	[-0.183, 0.059]
Year2017	-0.067	0.058	0.245	[-0.181, 0.046]
Year2018	0.053	0.059	0.371	[-0.063, 0.168]
Year2019	0.026	0.060	0.662	[-0.091, 0.143]
Year2020	0.016	0.058	0.785	[-0.098, 0.130]
Year2021	0.213	0.059	0.000	[0.099, 0.328]
Year2022	0.217	0.058	0.000	[0.103, 0.331]
Year2023	0.242	0.063	0.000	[0.118, 0.365]
Year2024	0.216	0.083	0.009	[0.053, 0.379]
Included Variables:				
legal_age	Included			
PB_number	Included			
Dummies for larger cities	Included			

Note: The table presents estimated coefficients, standard errors, p-values, and confidence intervals. Some variables (e.g., *legal_age*, *PB_number*, and city dummies) are included in the regression but not explicitly reported here. See Appendix C for the regression table including all variables.

Figure 9: The period-by-period price indices for each of the 100 imputed datasets



The figure illustrates the individual bilateral price indices for the office market in Austria between 2015 and 2024. Each line represents an index compiled with one of the 100 imputed datasets, forming the basis for the joint aggregate price index.

Figure 10: Bilateral index values for offices – result from 100 imputed datasets

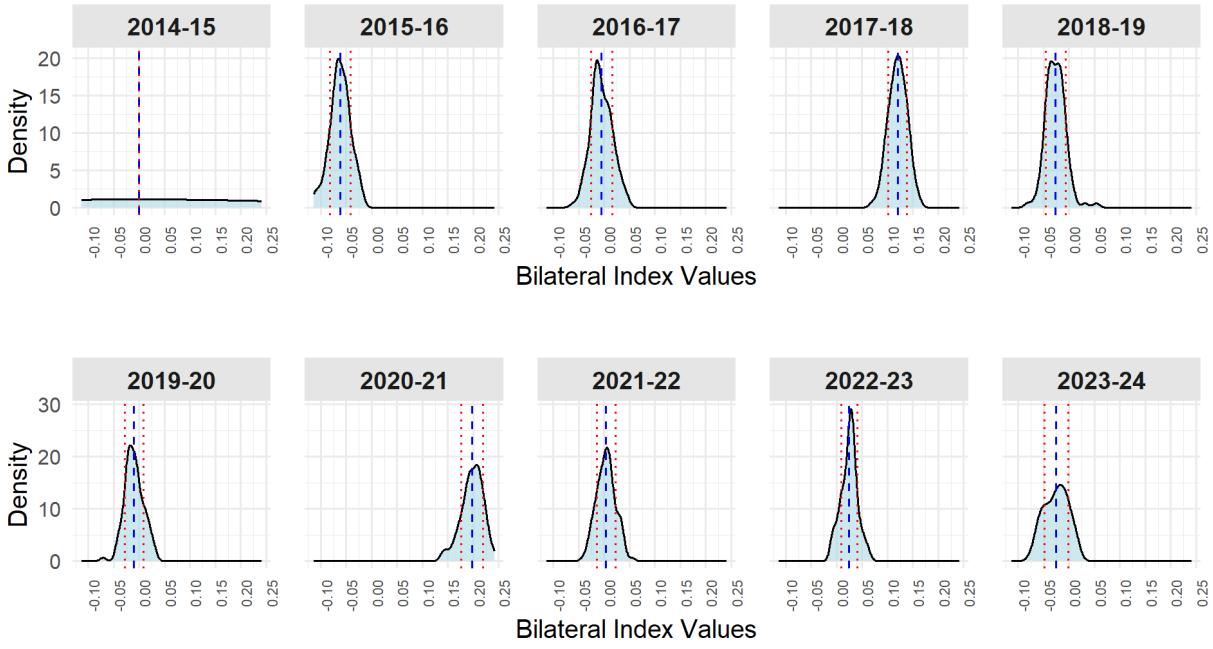
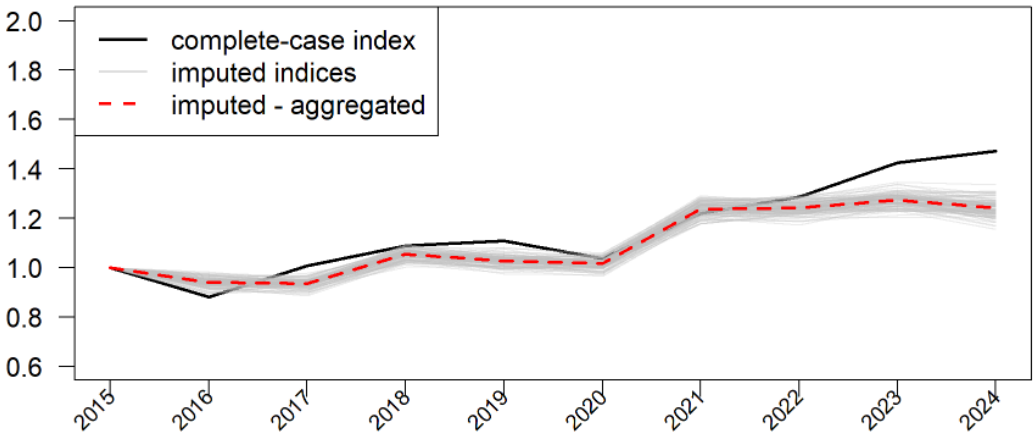


Figure 10 illustrates the variation in bilateral price index values for the 100 imputed datasets. The mean index value is indicated by the blue dashed vertical line. This value corresponds to the value of the bilateral index according to Rubin's Rules Rubin (1976, 1987). The red dotted lines indicate the standard deviation values. The first panel is empty as data only start in 2015.

Figure 11 presents the final aggregated office pirce index for Austria. The red dashed line represents the imputed index, while the solid black line represents the complete-case index. The gray lines indicate the individual price indices obtained from each of the 100 imputed datasets.

Figure 11: Office Price indices for Vienna – complete case and MICE-RF results



Note: In Figure 11, each gray line represents an index based on one of the 100 imputed datasets. The solid black line corresponds to the complete-case result, while the red dashed line represents the final aggregated office price index for Austria.

The complete-case and imputed indices exhibit similar trends up to 2021. However, from 2021 onwards, the two indices diverge. While the complete-case index indicates an upward market trend from 2021 to 2024, the index that includes imputed values suggests a flat price movement.

We argue that the imputed index provides a more accurate representation of market trends for three reasons: (i) it aligns more closely with professional surveyor assessments; (ii) the Vienna apartment market exhibits a similar trend flattening in 2021; and (iii) commercial office markets globally have experienced price stagnation or decline post-2021, largely attributed to the increased prevalence of working-from-home (Gupta et al., 2022; Deghi et al., 2022; Duca and Ling, 2024).

5 Conclusion

This paper highlights the critical impact of missing data on transaction-based property price indices and emphasizes the importance of robust imputation techniques. Using property transaction data from Austria, we demonstrated how different missing data mechanisms can distort price indices, affecting their reliability for policymakers, financial institutions, and market participants. While complete-case analysis remains a common approach in official statistics, our findings underscore the advantages of multiple imputation, specifically MICE with Random Forest, as a more reliable and methodologically

sound alternative.

Our results reveal that hedonic price indices for residential properties are generally stable when data are missing completely at random (MCAR), provided that a sufficient set of descriptive variables—such as location, size, and property condition—is available. Interestingly, imposing MNAR-type missingness proved difficult, as market forces tend to generate MAR-type missingness. This finding aligns with hedonic theory, which posits that property characteristics are implicitly linked to price, creating natural dependencies that facilitate imputation.

Our analysis further confirms that MCAR-type missingness does not introduce systematic bias into price indices. In line with existing literature (Rubin, 1976; Little and Rubin, 2019; Enders, 2022; Austin et al., 2021), we conclude that imputing missing values is unnecessary in cases where data are missing completely at random. However, non-random missingness—particularly systematic exclusions—can significantly bias price indices, making it imperative to address such missing data patterns.

Multiple imputation via MICE proved highly effective in correcting biases introduced by MAR-type missingness, preserving essential inter-variable relationships and improving the accuracy of price indices. This is particularly relevant in real estate markets, where missing data structures are often complex. Our findings highlight that the flexibility of the MICE framework, which supports both parametric and non-parametric imputation models, enhances its applicability. Specifically, our implementation of MICE with the Random Forest (rf) method for both continuous and categorical variables proved highly effective, reinforcing its suitability for real estate price estimation.

The benefits of multiple imputation extend beyond bias correction. Our applications to both residential and commercial real estate transaction data show that imputed price indices align more closely with the actual underlying price index (where known), expert assessments and observed market trends. This is especially critical for commercial real estate, where missing data are both more prevalent and more systematically patterned. Our study finds that imputing missing property characteristics -— particularly size and age —- yields a more plausible post-COVID-19 office price index for Austria than compiling a complete-case index. Notably, while the complete-case index suggests a continued price increase post-2021, the imputed index reflects a market stagnation that is more consistent with broader international trends in commercial real estate, particularly the effects of increased remote work (Gupta et al., 2022; Deghi et al., 2022; Duca and Ling, 2024).

Given these findings, we advocate for the broader adoption of multiple imputation techniques in constructing real estate price indices. We encourage further research to refine and standardize imputation methodologies, particularly in collaboration with international statistical agencies. Developing best practices for handling missing data will be essential in ensuring that price indices accurately reflect underlying market dynamics.

Ultimately, addressing the challenges of missing data is not merely a technical consideration but a fundamental requirement for maintaining the credibility and usefulness of real estate price indices. As transaction-based indices play an increasingly central role in economic analysis, investment decisions, and policy formulation, ensuring their accuracy through robust imputation techniques is crucial for informed decision-making.

References

Armknecht, P. A. and Maitland-Smith, F. (1999). Price imputation and other techniques for dealing with missing observations, seasonality and quality change in price indices. IMF Working Papers, 1999(078):A001.

Austin, P. C., White, I. R., Lee, D. S., and van Buuren, S. (2021). Missing data in clinical research: A tutorial on multiple imputation. Canadian Journal Cardiology, 37(9):1322–1331.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1):40–49.

Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical methods in medical research, 24:462–487.

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

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. American journal of epidemiology, 172(9):1070–1076.

Clapp, J. M. and Giaccotto, C. (1999). Revisions in Repeat-Sales Price Indexes: Here Today, Gone Tomorrow? Real Estate Economics, 27(1):79–104.

Commission Implementing Regulation (2023). Commission implementing regulation (eu) 2023/1470 of 17 july 2023 laying down methodological and technical specifications for the house price index (hpi) and the owner-occupied housing price index (oohpi). Accessed: 2025-01-01.

Court, A. T. (1939). Hedonic price indexes with automotive examples. In The Dynamics of Automobile Demand, pages 99–117, New York, NY. General Motors Corporation.

Crawford, I. and Neary, J. P. (2023). New characteristics and hedonic price index numbers. Review of Economics and Statistics, 105(3):665–682.

de Haan, J. and Diewert, E. (2013). Hedonic regression methods. Published by OECD, Eurostat, International Labour Organization, International Monetary Fund, The World Bank, and

United Nations Economic Commission for Europe. Available at: https://doi.org/10.1787/9789264197183-7-en.

Deghi, A., Natalucci, F., and Qureshi, M. S. (2022). Commercial real estate prices during covid-19: What is driving the divergence? Global Financial Stability Notes 2022/02, International Monetary Fund.

Deng, Y. and Lumley, T. (2024). Multiple imputation through xgboost. Journal of Computational and Graphical Statistics, 33(2):352–363.

Doove, L. L., van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis, 72:92–104.

Duca, J. V. and Ling, D. C. (2024). Work-from-home, covid-19, and online retail effects on commercial real estate prices. In 2025 ASSA Annual Meeting.

Enders, C. (2022). Applied Missing Data Analysis. Methodology in the Social Sciences Series. Guilford Publications.

Erler, N. S. (2019). Bayesian Imputation of Missing Covariates. PhD thesis, Erasmus University Rotterdam, Rotterdam, Netherlands. ISBN: 978-94-6323-623-2.

Erler, N. S., Rizopoulos, D., and Lesaffre, E. M. E. H. (2021). Jointai: Joint analysis and imputation of incomplete data in r. Journal of Statistical Software, 100(20):1–56.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W. V., Franco, O. H., and Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full bayesian approach. Statistics in Medicine, 35(17):2955–2974.

Eurostat (2013). Handbook on residential property price indices. https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF.

Eurostat (2017). Eurostat, commercial property price indicators - sources, methods and issues.

Eurostat (2018). Harmonised Index of Consumer Prices (HICP) Methodological Manual. Accessed: 2025-01-01.

Eurostat (2020). Estimation and Imputation of Missing Data for Short-Term Statistics. Accessed: 2025-01-01.

Feenstra, R. and Diewert, E. (2003). Imputation and Price Indexes: Theory and Evidence from the International Price Program. Working Papers 238, University of California, Davis, Department of Economics.

Fund, I. M. (2020). Consumer Price Index Manual: Concepts and Methods. International Labour Organization (ILO). Accessed: 2025-01-01.

Gatzlaff, D. and Haurin, D. (1997). Sample selection bias and repeat-sales index estimates. The Journal of Real Estate Finance and Economics, 14:33 – 50.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC.

Getz, K., Hubbard, R. A., and Linn, K. A. (2023). Performance of multiple imputation using modern machine learning methods in electronic health records data. Epidemiology, 34(2):206–215.

Glaeser, E. L., Gyourko, J., and Saks, R. E. (2005). Why is manhatten so expensive. The Journal of Law and Economics, 48(2).

Goh, G. and Kim, J. K. (2019). Accounting for model uncertainty in multiple imputation under complex sampling. arXiv preprint.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevention Science, 8(3):206–213.

Gupta, A., Mittal, V., and Van Nieuwerburgh, S. (2022). Work from home and the office real estate apocalypse. Working Paper 30526, National Bureau of Economic Research.

Hayes, T. (2018). Investigating the performance of cart- and random forest-based procedures for dealing with longitudinal dropout in small sample designs under mnar missing data. In Ferrer, E., Boker, S. M., and Grimm, K. J., editors, Longitudinal Multivariate Psychology, page 28. Routledge.

Hill, R. J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. Journal of Economic Surveys, 27(5):879–914.

Hill, R. J., Scholz, M., Shimizu, C., and Steurer, M. (2018). An evaluation of the methods used by european countries to compute their official house price indexes. Economie et Statistique, 500-502:221–238.

Ho, W. K. O., Tang, B. S., and Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1):48–70.

Honaker, J. and King, G. (2010). What to do about missing values in time series cross-section data. American Journal of Political Science, 54(3):561–581.

International Monetary Fund (2020). Residential Property Price Index: Practical Compilation Guide. International Monetary Fund, Washington DC.

Kofman, P. and Sharpe, I. G. (2003). Using multiple imputation in the analysis of incomplete observations in finance. Journal of Financial Econometrics, 1(2):216–249.

Kunkel, D. and Kaizar, E. E. (2017). A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Statistics in Medicine, 36(22):3507–3532.

Leamer, E. E. (2007). Housing is the business cycle.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association, 83(404):1198–1202.

Little, R. J. and Rubin, D. B. (2019). Statistical analysis with missing data, volume 793. John Wiley & Sons.

Little, T. D., Jorgensen, T. D., Lang, K. M., and Moore, E. W. G. (2014). On the joys of missing data. Journal of Pediatric Psychology, 39(2):151–162.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol, 110:63–73.

Moral-Benito, E. (2015). Model averaging in economics: An overview. Journal of Economic Surveys, 29(1):46–75.

Oberman, H. (2024). ggmice: Visualizations for 'mice' with 'ggplot2'. R package version 0.1.0.9000, https://amices.org/, https://amices.org/ggmice/.

OECD, Eurostat, Organization, I. L., Fund, I. M., Bank, T. W., and for Europe, U. N. E. C. (2013). Handbook on Residential Property Price Indices.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1):34–55.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3):581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association, pages 20–28.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley.

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research, 8(1):3–15. PMID: 10347857.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33(4):545–571. PMID: 26753828.

Schomaker, M. and Heumann, C. (2014). Model selection and model averaging after multiple imputation. Computational Statistics & Data Analysis, 71:758–770.

Schwerter, J., Joseph, A., Breunig, C., and Spindler, M. (2024). Evaluating tree-based imputation methods as an alternative to mice pmm for drawing inference in empirical studies. arXiv preprint arXiv:2401.09602.

Steel, M. F. (2020). Model averaging and its use in economics. Journal of Economic Literature, 58(3):644–719.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118.

Steurer, M., Hill, R. J., and Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38(2):99–129.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16(3):219–242. PMID: 17621469.

van Buuren, S. (2018). Flexible Imputation of Missing Data. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. Journal of statistical software, 45:1–67.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. Journal of Statistical Software, 45(3):1–67.

Appendix

A Pseudo-Algorithm for Random Forest

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ are the feature vectors and y_i are the corresponding targets. Number of trees T, and number of features to sample at each split m.
Output: An ensemble model $\mathcal{F}(\mathbf{x})$ for classification or regression.

1. **For** $t = 1$ to T:
   (a) Draw a bootstrap sample $\mathcal{D}_t$ of size n from $\mathcal{D}$.
   (b) Grow a decision tree $\mathcal{T}_t$ using $\mathcal{D}_t$:
      - At each node:
        - Randomly sample m features from the total p features.
        - Select the best feature and threshold for splitting based on a chosen criterion (e.g. RMSE for continuous variables).
      - Continue splitting until a stopping criterion is met (e.g., minimum node size or maximum tree depth).
   (c) Store the trained tree $\mathcal{T}_t$.
2. **Aggregate predictions:**
   - For continuous variables: $\mathcal{F}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{T}_t(\mathbf{x})$.
   - For factor variables: $\mathcal{F}(\mathbf{x}) = \text{mode}(\{\mathcal{T}_t(\mathbf{x})\}_{t=1}^{T})$.

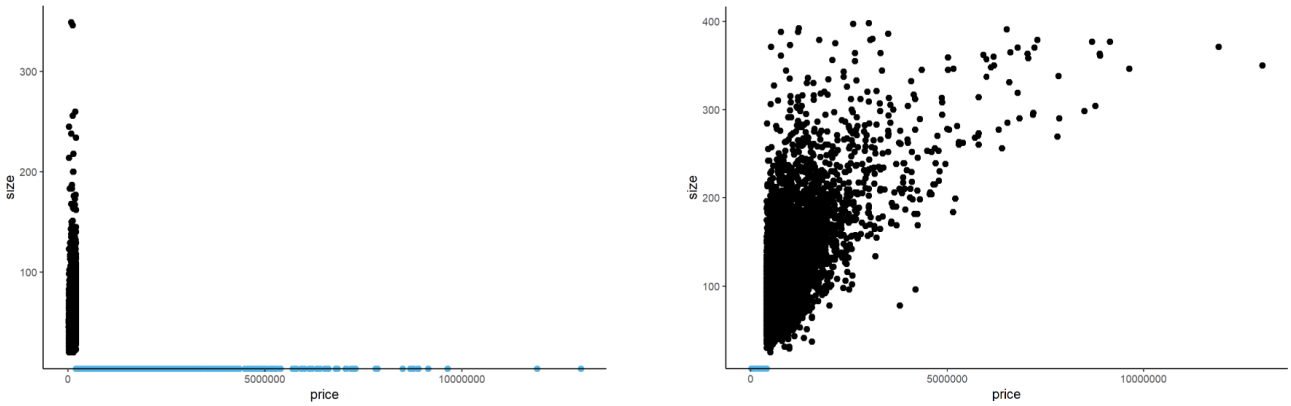
B Supplementary material for Application I: An apartment price index for Vienna, Austria

Table 6 compares the mean price and size of properties in each of the subsamples we constructed. As expected, the random samples have almost identical mean prices and square meter values. The MNAR samples show small differences in mean price and square meter values. Large differences between the means for price and size are seen only in the highly non-random *top25, top50, top75, bottom25, bottom50*, and *bottom75* samples.

Table 6: Sub-sample descriptions

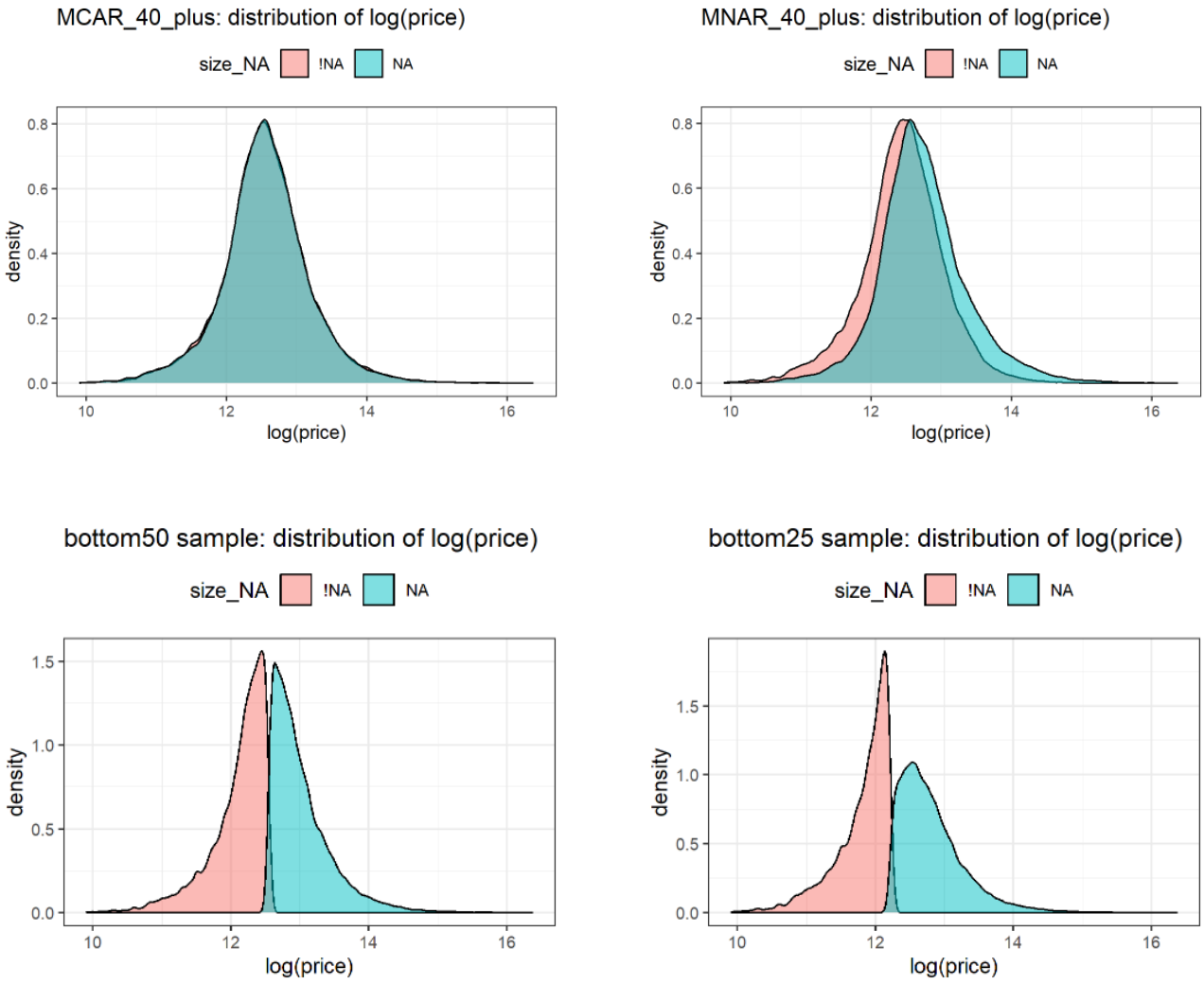
	MCAR_10	**MCAR_20**	**MCAR_30**	**MCAR_40**	**MCAR_50**
N	26236	8946	2701	594	123
mean Price	355979	355792	356651	355999	355985
sd(Price)	351994	349425	351770	341872	355932
mean(sqm)	72.00	72.00	72.00	72.00	72.00
sd(sqm)	33.63	33.58	33.36	33.44	33.72
MCAR-test	not reject	not reject	not reject	not reject	not reject
	nonMCAR_10	**nonMCAR_20**	**nonMCAR_30**	**nonMCAR_40**	**nonMCAR_50**
N	34310	18251	9083	4024	1559
mean Price	327069	316640	304952	299024	291880
sd(Price)	246000	232306	216691	224566	215839
mean(sqm)	70.0	69.00	68.00	68.00	67.00
sd(sqm)	30.6	29.83	28.95	28.88	28.83
MCAR-test	reject	reject	reject	reject	reject
	top75	**top50**	**top25**		
N	50829	33657	16823		
mean Price	425248	519545	703867		
sd(Price)	375165	431222	550366		
mean(sqm)	80.00	89.0	105.00		
sd(sqm)	34.25	36.4	41.38		
MCAR-test	reject	reject	reject		
	bottom75	**bottom50**	**bottom25**		
N	50469	33700	16891		
mean Price	39520	191736	142097		
sd(Price)	86420	60971	44022		
mean(sqm)	61.0	55.00	50.00		
sd(sqm)	21.5	18.76	17.75		
MCAR-test	reject	reject	reject		

Figure 12: Distribution of available price and size values



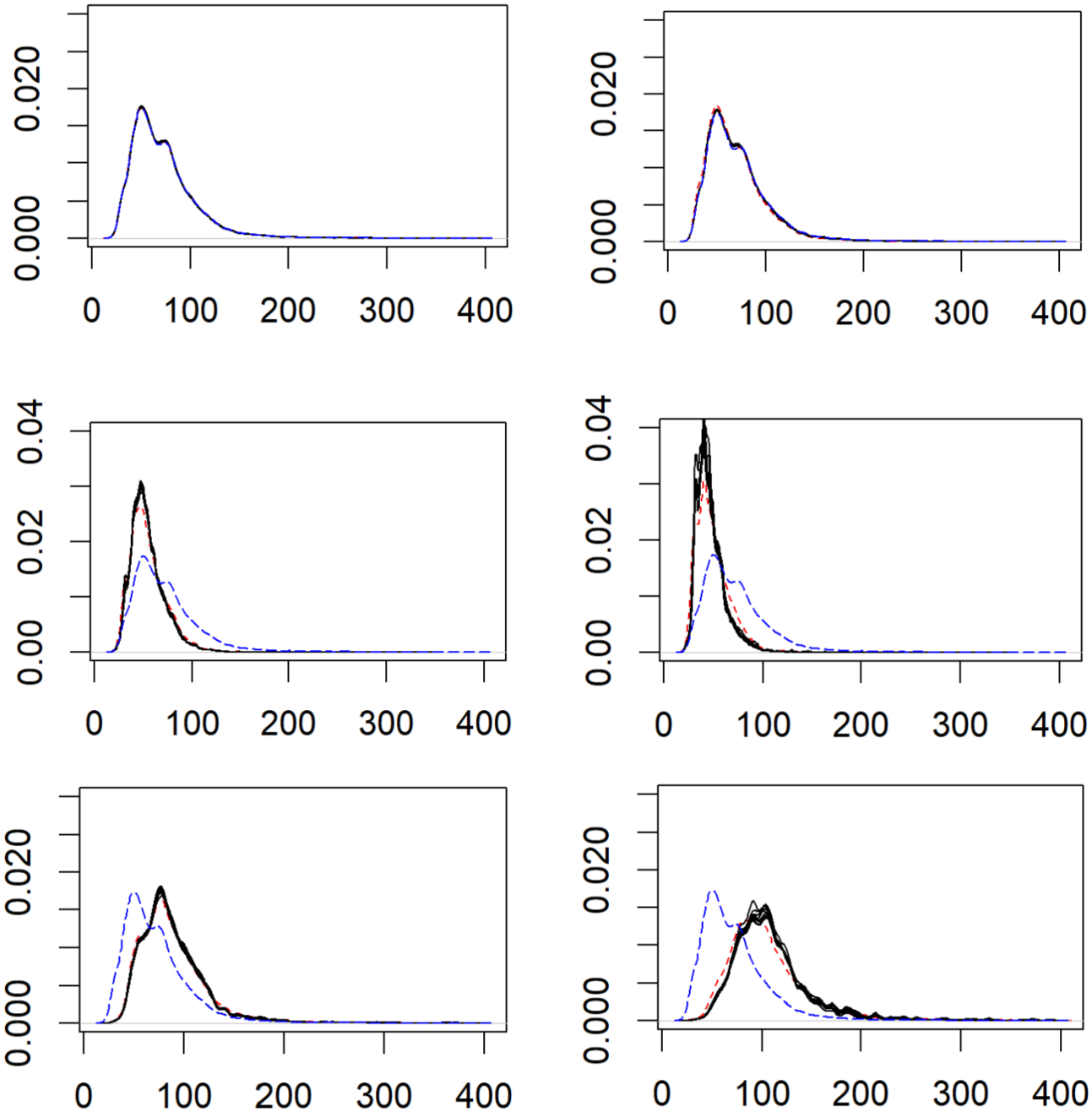
Note: The figures above illustrate the available price-size combinations for two data subsets: (a) the subset where size information is available only for the bottom 25 percent of the market (in terms of total price), and (b) the subset where size information is available only for the top 25 percent of the market. Observed price-size combinations are represented by black dots, while prices without corresponding size information are marked by blue dots along the x-axis.

Figure 13: Distribution of apartment price depending on whether size is available or not: four different data samples



Density distribution of size variable before and after imputation

Figure 14: Density plots for size before and after imputation of missing values

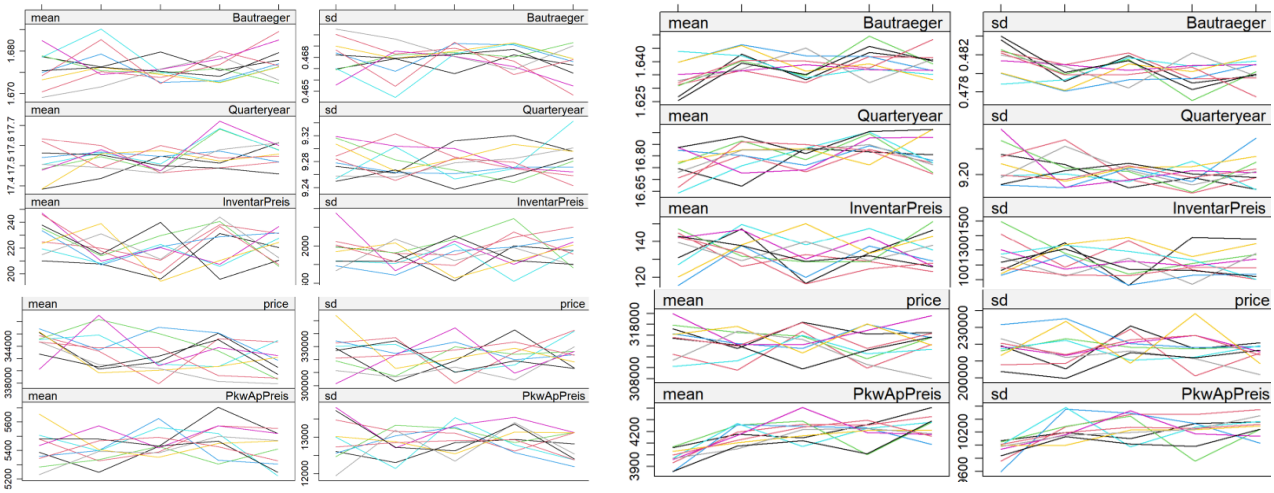


Note: The figures illustrate density plots for size before and after imputing missing values for the following six subsamples. From top left the figures illustrate the cases of the MCAR40, MNAR40, bottom50, bottom25, top50, and top25 subsets. The density plots for the 10 imputed datasets are shown as the black solid lines. The pre-imputation densities are shown as red dotted lines, and the original full-case data as the blue dotted line. As expected, the univariate distributions of the observed and imputed data are identical when missingness is of the MCAR type (figure top left). The non-random deletion of values via our so-called MNAR process did not lead to a shift in the size distribution either (figure top right). The middle row (for bottom50 and bottom25) and final row (top50, top25) illustrate that starting from a severely censured sample, the imputation process is

notable to recover the original size distribution (the dashed blue line). Howecer, if missing values are generated via a MAR process, the univariate distributions may differ in location and dispersion (even though it is assumed that their multivariate distribution is identical) van Buuren and Groothuis-Oudshoorn (2011).

Implementation results

Figure 15: Imputation with mice: stream conversion of MCAR-40 and MNAR-40 samples



Note: There is no completely certain way to check the convergence of the imputations, but comparing the parameters with the number of iterations is a good way to see if the samples are similar. The figures above show the mean and variance of the imputations for each of the ten parallel streams for the MCAR-40 and MNAR-40 samples. Both samples were imputed using the mice algorithm (van Buuren and Groothuis-Oudshoorn, 2011). We used the Random Forest method (specifications: m=10, maxit=10, meth='rf', ntree = 10, and visitSequence = "monotone"). We can conclude that convergence is present as the different streams mix freely and without clear trends. Generally, convergence is diagnosed when the variance between the different streams is not greater than the variance within each individual stream. Both samples were imputed within the mice algorithm

B.0.1 Illustration of MICE-RF hedonic regression output

Table 7: Adjusted R^2 Values

Model Type	1	2	3	4	5	6	7	8	9	10
Adjusted R^2 before Imputation										
Original Model	0.66									
Random 50%	0.66									
MCAR40	0.65									
MNAR40	0.55									
Top25	0.67									
Bottom25	0.18									
Adjusted R^2 for each Imputation										
MNAR40	0.42	0.44	0.39	0.43	0.43	0.46	0.44	0.48	0.43	0.45
MCAR40	0.44	0.45	0.48	0.50	0.51	0.47	0.46	0.47	0.44	0.47
Top25	0.26	0.26	0.25	0.27	0.27	0.26	0.26	0.26	0.25	0.26
Bottom25	0.22	0.23	0.22	0.22	0.22	0.22	0.22	0.23	0.23	0.23

Note: The complete-case analysis tends to give higher adjusted R^2 values than imputation models as the missingness is not taken into account.

C Supplementary material for Application II: CRE price index for Austria

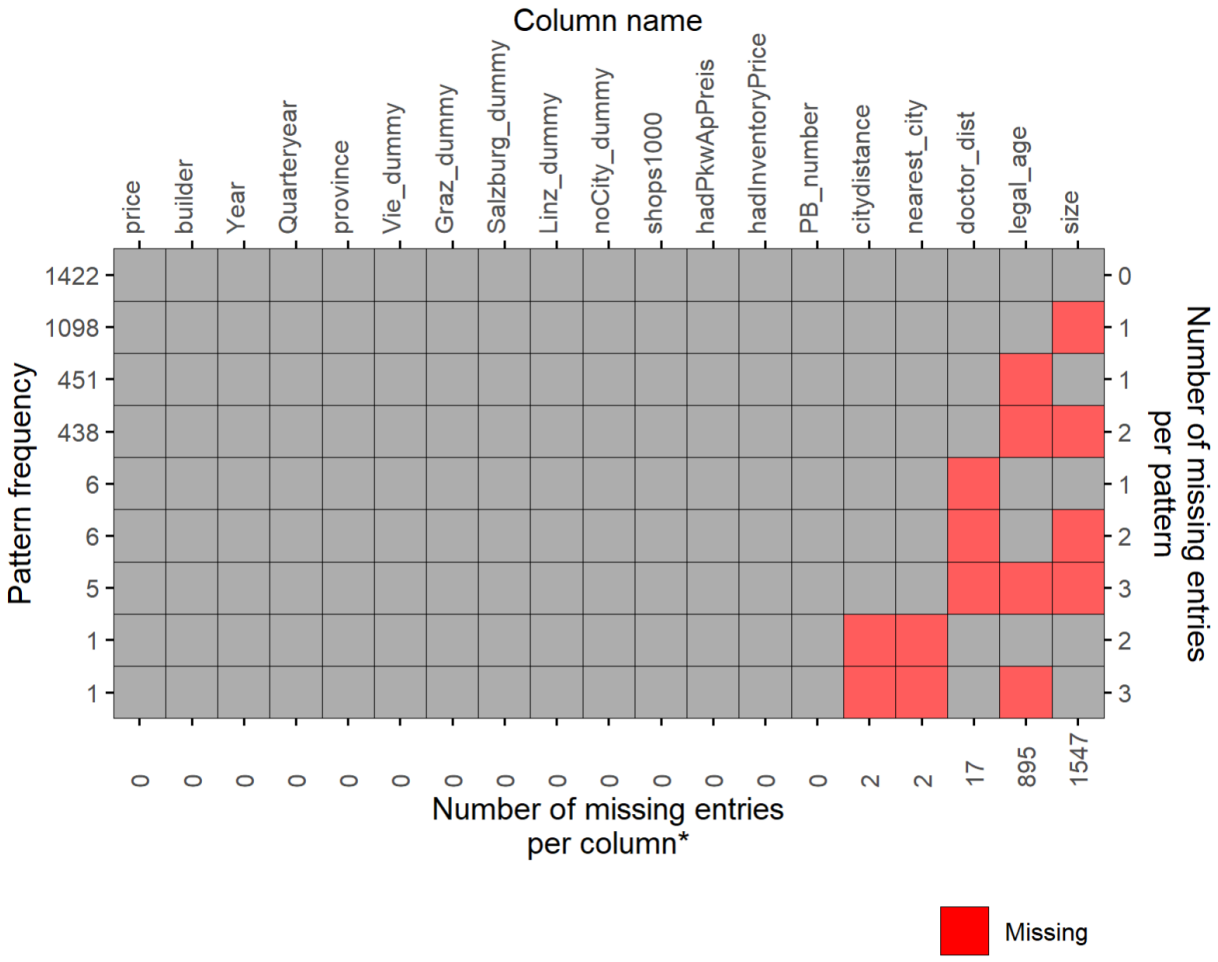
Table 8: Average property price by Province and whether or not size is missing in contract

Size	Province	Price_Mean	Price_SD	N_Obs
yes	B	**186935**	173128	32
no	B	291348	276468	14
yes	K	**341814**	480776	155
no	K	228776	202422	100
yes	N	**535656**	607209	119
no	N	325206	422596	162
yes	O	**449772**	530203	251
no	O	348504	352252	197
yes	S	**653977**	972900	301
no	S	483321	610366	198
yes	ST	**481101**	562498	168
no	ST	317856	379304	80
yes	T	**463155**	476619	221
no	T	374125	322176	141
no	V	**385578**	450323	70
yes	V	353795	300624	53
no	W	**817847**	1964160	585
yes	W	637430	940795	581

Note: In most provinces, the mean price (*Price_Mean*) is higher for properties without missing size values, but for the provinces Vorarlberg (V) and Vienna (W) this trend is reversed. Overall, there is no large difference in price between those properties that have size information given and those that do not.

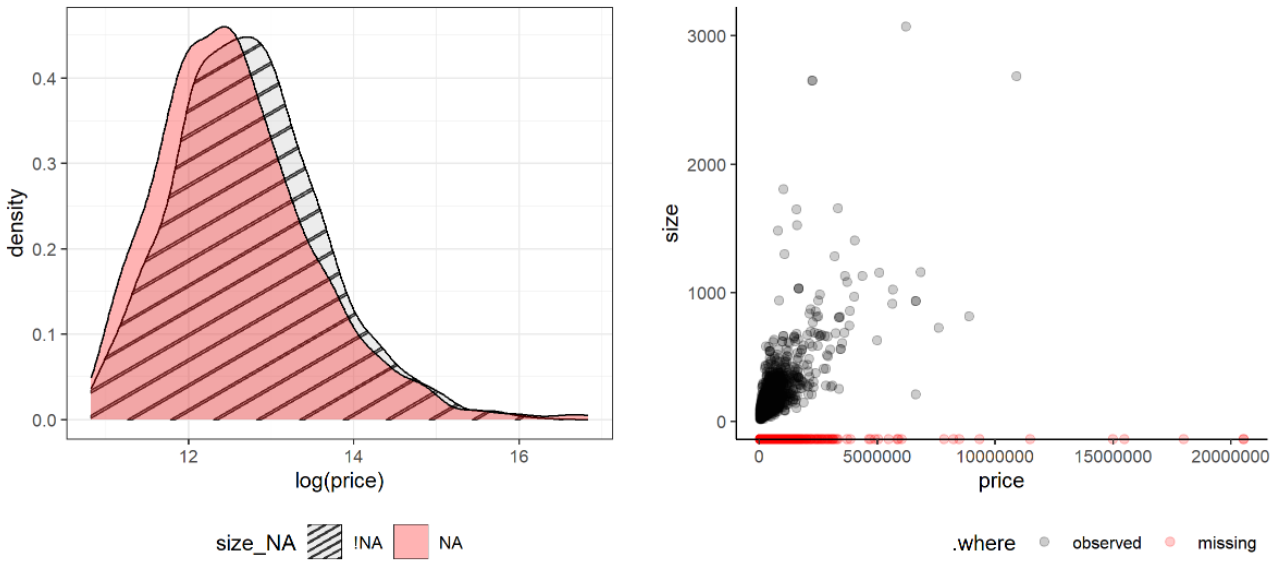
Missingness patterns in office dataset

Figure 16: Missingness pattern



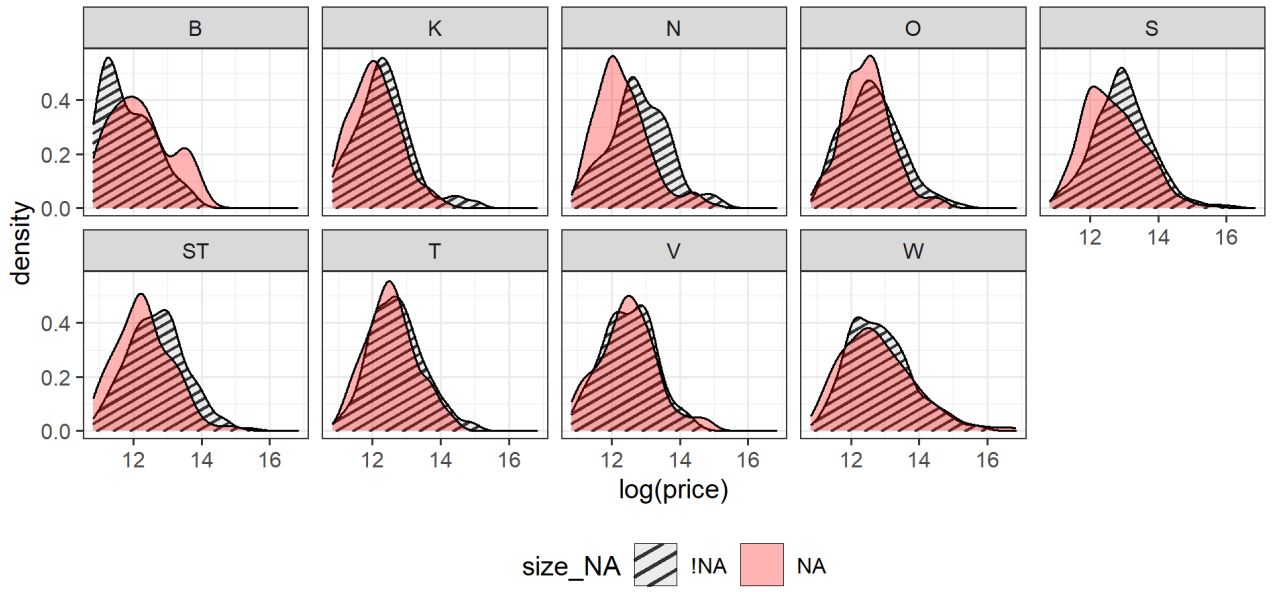
Note: Missing values are indicated in red.

Figure 17: Log-price distributions by size data availability.



Note: Most missing size values fall within frequently observed price ranges, except for four high-value properties (red dots). As no complete-case examples exist in this range, imputation is unlikely to yield reliable estimates. Figure generated using R-package *ggmice* (Oberman, 2024).

Figure 18: Missingness and log(price) by region.



Note: Regional price distributions are largely unaffected by missing size data, except in Lower Austria (N) and Styria (ST), where properties with size entries tend to be more expensive.

Figure 19 illustrates postcode-level missingness patterns, where larger cities show mixed missingness due to higher notary density, while rural areas tend to have either complete or no size data.

Figure 19: Missingness of size at the postcode level.



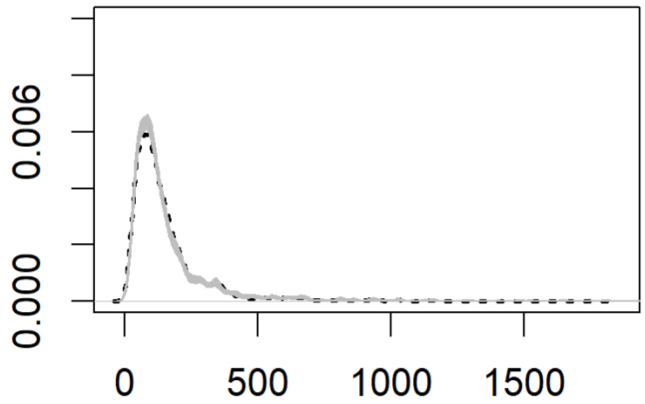
Note: The percentage of missing size values varies by postcode. In Vienna, size is missing in about 50% of cases, similar to other major cities. In rural areas, size data is either consistently recorded or entirely absent, likely reflecting local notarial practices.

Figure 20: Missingness relative to age, separate plots for different *legal_age*



Note: Price distributions for properties with and without size data are largely similar across Austria's nine regions, except in Lower Austria. However, missingness patterns vary significantly at finer geographic levels, likely due to differences in notarial recording practices across Austria's 405 notaries. The facet plots indicate that price distributions vary by age category but not by the presence of missing internal size data. Notably, properties missing both size and age data exhibit price distributions nearly identical to those with complete information.

Figure 21: Distribution of office sizes within each dataset



Note: Size is the variable with the most missingness. The figure above compares the distribution of office sizes between the full-case sample (dashed line) and each of 100 imputation datasets (gray solid lines).

Density plots for period-by-period index results after imputation

Table - Pooled regression results of 100 imputed datasets.

Term	Estimate	Std. Error	Statistic	DF	P-value	2.5%	97.5%
(Intercept)	10.930	0.331	33.030	1528.7	0.000	10.281	11.579
size	0.002	0.000	12.887	110.2	0.000	0.002	0.002
log(citydistance)	-0.109	0.033	-3.353	1426.9	0.001	-0.173	-0.045
builder	0.168	0.033	5.156	1659.4	0.000	0.104	0.231
legal_age1960s	0.291	0.252	1.154	1033.2	0.249	-0.204	0.785
legal_age1970s	0.295	0.243	1.213	1100.1	0.225	-0.182	0.771
legal_age1980s	0.541	0.243	2.225	1043.1	0.026	0.064	1.018
legal_age1990s	0.385	0.233	1.656	1138.6	0.098	-0.071	0.842
legal_age2000s	0.438	0.232	1.891	1141.4	0.059	-0.016	0.893
legal_age2010-15	0.537	0.233	2.300	1120.2	0.022	0.079	0.995
legal_age2015-2020	0.496	0.232	2.134	1144.1	0.033	0.040	0.952
legal_age2020+	0.501	0.239	2.094	1020.2	0.036	0.032	0.971
doctor_dist	0.000	0.000	2.470	1853.5	0.014	0.000	0.000
PB_number106	0.416	0.408	1.018	2150.1	0.309	-0.385	1.217
PB_number109	0.583	0.299	1.953	2489.5	0.051	-0.002	1.169
PB_number202	0.769	0.262	2.933	2345.2	0.003	0.255	1.283
PB_number204	0.235	0.245	0.959	2389.9	0.338	-0.246	0.717
PB_number205	0.482	0.358	1.345	2686.6	0.179	-0.220	1.184
PB_number206	0.871	0.309	2.820	2201.4	0.005	0.265	1.476
PB_number209	0.834	0.288	2.895	2561.4	0.004	0.269	1.399
PB_number210	0.682	0.311	2.192	2327.0	0.029	0.072	1.292
PB_number305	0.475	0.321	1.483	2341.9	0.138	-0.153	1.104
PB_number306	0.744	0.263	2.824	1952.4	0.005	0.227	1.261
PB_number311	1.044	0.325	3.217	2101.1	0.001	0.408	1.681
PB_number312	0.776	0.307	2.529	2004.1	0.012	0.174	1.379
PB_number313	0.910	0.272	3.348	2479.9	0.001	0.377	1.442
PB_number318	0.387	0.308	1.257	1092.5	0.209	-0.217	0.990
PB_number321	0.516	0.310	1.666	2518.7	0.096	-0.091	1.124
PB_number324	0.948	0.271	3.504	2340.1	0.000	0.418	1.479
PB_number401	1.362	0.652	2.088	1998.3	0.037	0.083	2.641
PB_number404	0.857	0.453	1.892	1826.5	0.059	-0.031	1.745
PB_number407	0.865	0.275	3.141	2524.0	0.002	0.325	1.405
PB_number408	1.207	0.284	4.254	2223.1	0.000	0.651	1.764
PB_number409	1.199	0.312	3.847	2471.4	0.000	0.588	1.811
PB_number410	0.672	0.288	2.337	2021.8	0.020	0.108	1.237
PB_number415	0.963	0.278	3.468	2274.6	0.001	0.418	1.507
PB_number416	0.860	0.344	2.498	2612.1	0.013	0.185	1.535

Continued on next page

Term	Estimate	Std. Error	Statistic	DF	P-value	2.5%	97.5%
PB_number418	0.797	0.250	3.191	2379.1	0.001	0.307	1.287
PB_number501	1.011	0.356	2.838	1965.5	0.005	0.312	1.709
PB_number502	0.798	0.292	2.733	1336.2	0.006	0.225	1.370
PB_number503	0.539	0.252	2.136	2279.1	0.033	0.044	1.033
PB_number504	1.141	0.265	4.312	2394.4	0.000	0.622	1.660
PB_number506	1.235	0.274	4.509	2254.0	0.000	0.698	1.772
PB_number601	0.924	0.240	3.855	2291.3	0.000	0.454	1.394
PB_number606	0.811	0.265	3.063	2487.6	0.002	0.292	1.330
PB_number612	0.697	0.304	2.292	2368.0	0.022	0.101	1.294
PB_number616	0.816	0.355	2.300	2098.9	0.022	0.120	1.512
PB_number617	1.043	0.305	3.420	2551.1	0.001	0.445	1.640
PB_number701	0.935	0.239	3.918	2306.0	0.000	0.467	1.403
PB_number702	0.848	0.270	3.146	2484.1	0.002	0.319	1.376
PB_number703	0.745	0.248	3.000	2557.7	0.003	0.258	1.232
PB_number705	1.198	0.263	4.559	2280.1	0.000	0.683	1.714
PB_number706	1.064	0.352	3.022	2733.2	0.003	0.374	1.754
PB_number707	0.975	0.304	3.207	2447.6	0.001	0.379	1.571
PB_number709	1.087	0.274	3.960	2468.1	0.000	0.549	1.625
PB_number801	0.785	0.265	2.967	2395.4	0.003	0.266	1.304
PB_number802	0.712	0.252	2.824	2430.7	0.005	0.218	1.207
PB_number803	0.991	0.252	3.928	2704.8	0.000	0.496	1.486
PB_number900_1	2.245	0.254	8.827	1703.9	0.000	1.746	2.743
PB_number900_center	1.138	0.240	4.746	2351.6	0.000	0.668	1.608
PB_number900_east	0.900	0.256	3.519	2367.2	0.000	0.399	1.402
PB_number900_west	1.003	0.246	4.070	2382.7	0.000	0.520	1.487
PB_numberOther	0.579	0.265	2.183	2337.5	0.029	0.059	1.098
Vie_dummyTRUE	-0.130	0.087	-1.501	1308.8	0.134	-0.300	0.040
Graz_dummyTRUE	-0.281	0.129	-2.178	1478.4	0.030	-0.533	-0.028
Salzburg_dummyTRUE	0.083	0.271	0.308	1626.9	0.758	-0.447	0.614
Linz_dummyTRUE	-0.494	0.600	-0.822	2000.2	0.411	-1.671	0.684
hadPkwApPreisTRUE	0.316	0.039	8.108	2261.5	0.000	0.240	0.393
Year2016	-0.062	0.062	-1.005	2134.6	0.315	-0.183	0.059
Year2017	-0.067	0.058	-1.163	2151.5	0.245	-0.181	0.046
Year2018	0.053	0.059	0.895	2214.1	0.371	-0.063	0.168
Year2019	0.026	0.060	0.438	2039.9	0.662	-0.091	0.143
Year2020	0.016	0.058	0.273	2018.2	0.785	-0.098	0.130
Year2021	0.213	0.059	3.644	1685.5	0.000	0.099	0.328
Year2022	0.217	0.058	3.734	1973.7	0.000	0.103	0.331
Year2023	0.242	0.063	3.837	2041.8	0.000	0.118	0.365
Year2024	0.216	0.083	2.604	2093.8	0.009	0.053	0.379