# Filling the Gaps with MICE-RF: Missing Data in Real Estate Price Indices

Miriam Steurer     Sabrina Spiegel

Department of Economics, University of Graz

ZTdatenforum Graz

NBER CRIW Pre-Conference, July 16, 2025

# Motivation

- Real estate price indices: vital for monetary and macroprudential policy
- Internationally, hedonic regression based on transaction data is the primary method for constructing property price indices (Eurostat, 2013, 2017; Hill et al., 2018; International Monetary Fund, 2020)
- EU countries are required to compile hedonic residential property price indices. Hedonic CRE price indices from 2027 onwards.
- Problem: registry data developed for taxation, not for price index construction –> missing values in “non-essential" fields.
- Particularly a problem for CRE price indices (more heterogeneous properties, lower number of transactions, and more missingness).

# Motivation cont.

- Missing data: a problem for hedonic real estate price indices.
- Traditional methods:
  Most NSIs ignore missingness and do the best with remaining data.
  (i) leaving important variable out, or
  (ii) performing complete-case analysis
  –> these strategies can bias index
- Existing literature on imputation in real estate indices is sparse, despite its extensive use in fields like medical research starting with Rubin (1976).
- IMF or Eurostat do not provide guidance on how to deal with missingnessin RE data.

# Research Questions & Contributions

**Research Questions:**

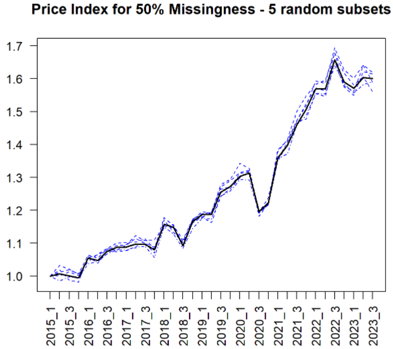
1. How do RE price indices behave under different missingness patterns?
2. Can multiple imputation (with MICE + RF) reduce bias in real estate indices?
3. Can the optimal imputation model be "automated" so to ease use by NSIs?

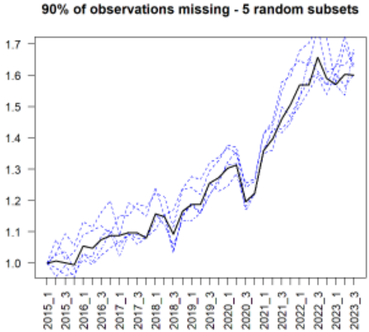
**Contributions of our paper:**

- Introduce MICE-RF for price index compilation
- Simulation study on Vienna apartments (2015–23)
- Application to Austrian office market (2015–24)

# Illustrating the impact of missing data

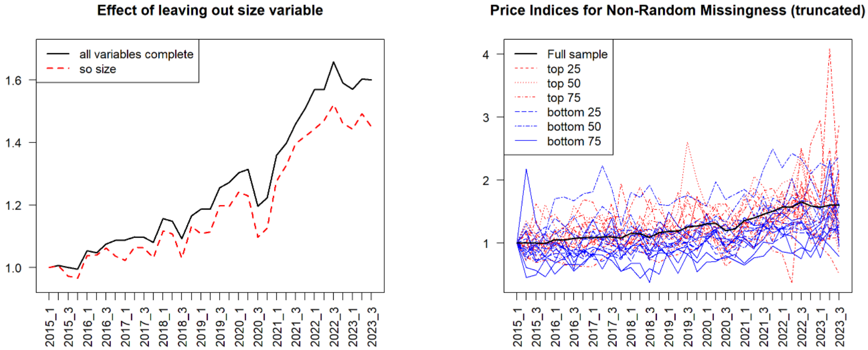
# The good news – hedonic indices are quite robust against randomly missing data





- Randomly missing data for Vienna apartment transactions: little bias, more volatility

# The bad news: systematic missingness does produce biases



(a) leaving a key variable out

(b) truncation of dataset

- Leaving out key variable –> bias
- Truncation (e.g. missing high- or low-end of market) –> bias

# What is MICE-RF?

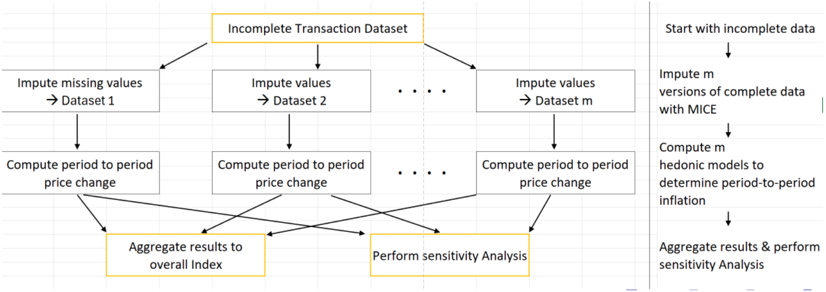
# Multiple Imputation by Chained Equations (MICE)

- Multiple imputation (MI): significant advantages over single imputation and complete-case analysis (Rubin, 1987).
- Multiple Imputation by Chained Equations (MICE) framework developed by van Buuren (2007), van Buuren and Groothuis-Oudshoorn (2011).
- MICE is gold standard MI approach in medical research.
- Advances in computing power and statistical software → MI possible for large real estate datasets.
- Within MICE framework different imputation algorithms can be used → we need to make two decisions:
  1. Choice of Multiple Imputation framework (i.e., MICE)
  2. Choice of algorithm/model to use for imputing each of the variables with missing data.

# Multiple Imputation by Chained Equations (MICE)

How it works:

- Start with incomplete dataset.
- Create *m* versions of complete datasets by iteratively imputing each variable until convergence.
- The variability across imputations captures the uncertainty involved.
- Perform separate analysis on each of the *m* datasets.
- Then pool results using Rubin's Rules



Incomplete Transaction Dataset

Impute missing values → Dataset 1

Impute values → Dataset 2

. . . .

Impute values → Dataset m

Compute period to period price change

Compute period to period price change

. . . .

Compute period to period price change

Aggregate results to overall Index

Perform sensitivity Analysis

Start with incomplete data

Impute m versions of complete data with MICE

Compute m hedonic models to determine period-to-period inflation

Aggregate results & perform sensitivity Analysis

# The MICE algorithm according to van Buuren and Groothuis-Oudshoorn (2011)

- Let $Y = \{Y_{i,j}\}$ be a data matrix, where $Y_{i,j}$ represents the $j$-th variable for the $i$-th observation. $Y$ is partitioned into observed ($Y_{obs}$) and missing ($Y_{mis}$) components.
- Start with an initial guess for $Y_{mis}$.
- For each iteration, impute missing values of $Y_j$ using the observed values of other variables (and previously imputed values):

$$P(Y_j|X_j, \theta_j), \quad X_j = Y \setminus Y_j.$$

- Update predictors: $X_j^{(t)} = (X_j^{(t-1)}, \tilde{Y}_j^{mis,(t)})$ and $\theta_j$'s.
- Repeat until convergence (imputed values stabilize).

# Choice of algorithm to use within MICE

- We could specify a separate model for each variable.
- However, we use Random Forest algorithm to impute missingness across all variables (Breiman 2001, Breiman and Cutler, 2003).
- Reasons:
  - Tree-based ML methods best at value estimates for RE (AVMs, Kaggle competitions).
  - RF can automatically capture non-linearities and interactions → easy to use
  - Automates model selection within MICE
  - Subsampling and feature randomness reduce overfitting

# Hedonic Time Dummy Price Index

# Hedonic Time Dummy Price Index

- Once we have created $m$ complete versions of the original dataset, we perform standard hedonic price index compilation on each of them.
- For the time dummy hedonic model we estimate:

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

  - $D_n^\tau$: time dummy for period $\tau$
  - $z_{nk}^t$: property characteristics
  - Bilateral index: $P^{t-1,t} = \exp(\hat{\delta}^t - \hat{\delta}^{t-1})$
- Then aggregate the $m$ bilateral price indices into final price index (each analysis gets equal weight).

# Application 1: Vienna Apartment market 2015-2023

# Application 1: Vienna Apartments

- Data: 67,292 transactions (2015–Q3/2023)
- Descriptive variables: price, size, postcode, distance to urban center and amenities
- Simulated missingness:
  - MCAR 10–50%
  - MNAR 10–50%
  - Truncation: top/bottom quantiles (missing size)
- We found it very hard to construct MNAR missingness. Reason: market forces establish a connection between the characteristics of a property and its price (Rosen, 1974)
  → MNAR cases exhibit MAR-like structures
  → allows imputations to work.
- Note: multicolinearity is great for imputation
- Mice-RF imputation setting: $m = 10$, $maxit = 5$, $ntree = 10$

# Application 1 Results: Vienna

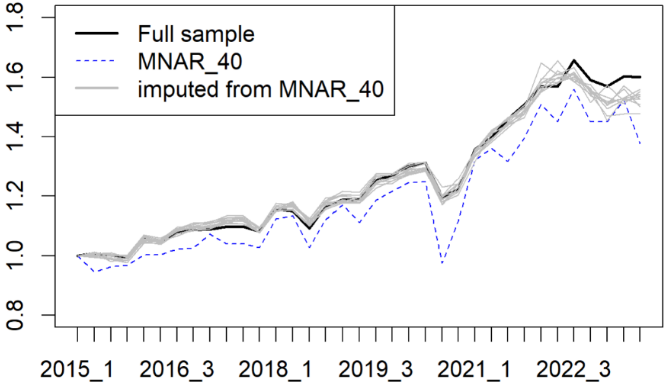


Figure: Imputation for MNAR_40 Sample

# Application 1 Results: Vienna

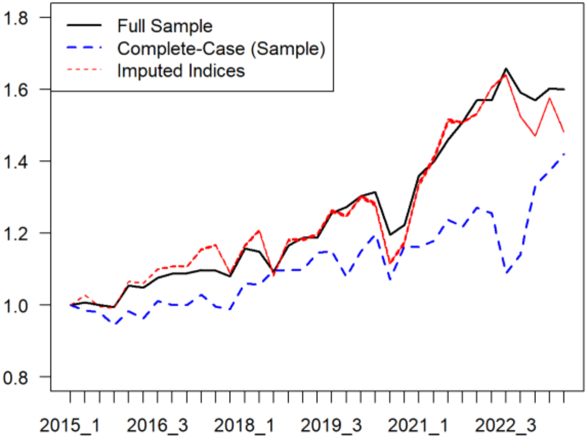


Figure: Imputation result for truncated dataset (top25)

# Imputation Results with different missingness patterns

| **Missingness Type** | **Little's MCAR Test** | | **N** | | **Bias** | | **Volatility** | |
|---|---|---|---|---|---|---|---|---|
| | **Statistic** | **p-value** | **Before** | **After** | **Before** | **After** | **Before** | **After** |
| **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** | | | | | | | | |
| MNAR10 | 41,417 | 0 | 36,427 | 67,292 | 0.952 | 0.997 | 0.043 | 0.036 |
| MNAR20 | 38,371 | 0 | 22,830 | 67,292 | 0.946 | 0.997 | 0.052 | 0.038 |
| MNAR30 | 35,916 | 0 | 12,697 | 67,292 | 0.907 | 0.963 | 0.059 | 0.038 |
| MNAR40 | 32,025 | 0 | 6,462 | 67,292 | 0.881 | 0.956 | 0.078 | 0.038 |
| MNAR50 | 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 |

# Application 2: Austrian Office Market 2015-2024

# Application 2: Austrian Office Market

- 3,448 transactions (2015–9/2024)
- Key missing: size (45%), legal age (26%)
- Locational  physical characteristics only have little missingness
- MICE-RF: $m = 100$, $maxit = 5$, $ntree = 10$

# Key Variables in the Dataset

**Physical Property Characteristics:**

- Size
- Legal Age
- Builder
- hadPkwApPrice
- hadInventoryPrice

**Locational Attributes:**

- Province (9)
- PB Number (50)
- City Dummies (Vienna, Graz, Salzburg, Linz)
- Postcodes
- Longitude and Latitude

**Distance-Related Variables:**

- Shops1000
- Doctor Dist
- Citydistance

# Bilateral Index Values for 100 Imputations

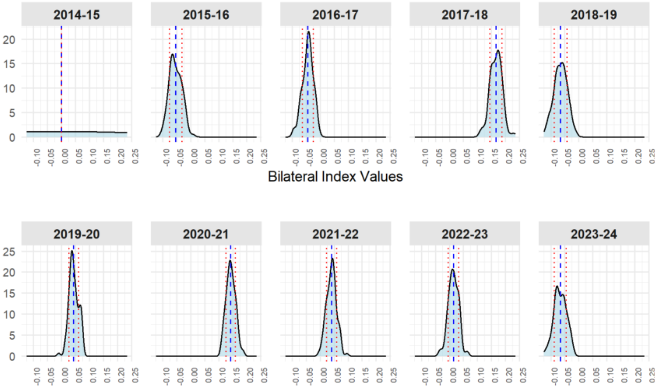
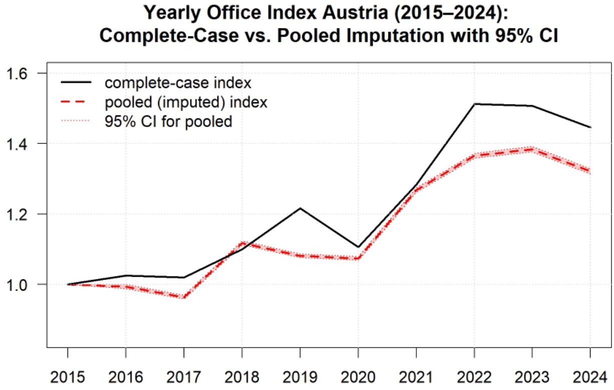


Figure: The figure illustrates the variation in bilateral price index values for the 100 imputed datasets. The mean corresponds to the value of the bilateral index according to Rubin's Rules Rubin (1976, 1987). The red dotted lines indicate the standard deviation values. The first panel is empty as data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

# Application 2 Results: Office Index Results



- Imputed index has flatter post-2021 trajectory
  → more realistic according to experts.
- m=100 → confidence bands very narrow

# Same method – but only 10 Imputations

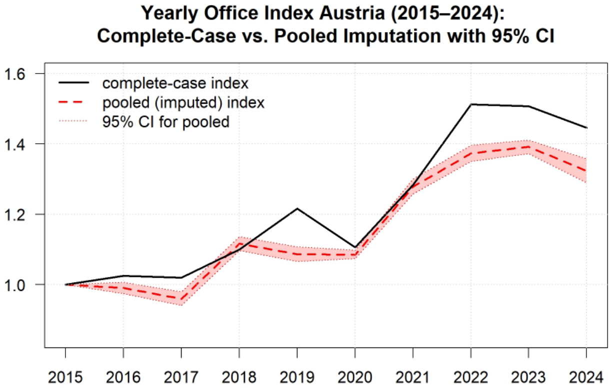


Figure: Result for m=10

# Implications and Conclusions

# Implications and Conclusion

- MICE-RF can effectively correct bias under systematic missingness
  –> Enhances reliability of real estate price indices
  –> We encourage broader adoption by NSIs

- Next steps:
  –> Broader discussion on how both imputing and not-imputing missing data can bias results.
  –> Development of guidelines on imputation methods in NSI production workflow.

# Acknowledgments

- Thank you for listening. Are there any Questions?
- If you think of questions or suggestions later, please contact me via miriam.steurer@uni-graz.at

- We thank ZTdatenforum for providing the data.