Filling the Gaps with MICE: Addressing Missing Data in Real Estate Indices

Miriam Steurer Sabrina Spiegel

University of Graz, Austria

ZTdatenforum Austria

ESCoE Conference 2026
May 19-21, 2026

Missing characteristics in official property price statistics

Most countries construct residential and commercial property price indices using hedonic methods applied to transaction data.

In these datasets:

- transaction price, date, and location are almost always observed
- missingness mainly affects quality characteristics:
  - floor area
  - age
  - parking
  - building characteristics

Common practice

Statistical agencies commonly:

1. drop variables with many missing values
2. use complete-case estimation

Our paper shows:

> Complete-case estimation is not always harmless.
> It can implicitly redefine the target population.

Research questions

1. When does missingness distort hedonic property price indices?
2. Can multiple imputation reduce these distortions?
3. How should we go about it? $\Rightarrow$ MI, but how?

Why multiple imputation?

Single imputation fills missing values as if they were known with certainty.

Multiple imputation instead generates several plausible completed datasets.

The idea:

1. impute missing values multiple times
2. estimate the model on each completed dataset
3. aggregate results across imputations

Multiple imputation preserves uncertainty about missing values.
$\Rightarrow$ More uncertainty about the missing data leads to greater dispersion across the resulting estimates.

Multiple Imputation by Chained Equations (MICE)

We implement multiple imputation using the **MICE framework** (van Buuren 2007, 2018).

Basic idea:

- each variable with missing values is imputed using a separate model
- models are estimated sequentially
- the process iterates until convergence (or max iteration reached)

Advantages:

- flexible model choice for each variable
- easy to implement using the `mice` package in R

What matters more: algorithm or sample restoration?

We compare:

- MICE with random forests
- default MICE (PMM)
- linear regression imputation
- mean imputation

Findings:

- MICE-RF and default MICE perform very similarly
- linear models become unstable under collinearity
- mean imputation performs clearly worse

The main gain comes from restoring sample composition, not from optimizing the prediction algorithm.

Why Rubin's Rules are not appropriate for chained indices

Rubin pooling combines parameter estimates across imputations.

But chained price indices are built from adjacent-period growth rates.
Pooling index levels directly would distort this structure.

Our approach:

1. generate m completed datasets
2. compute adjacent-period growth rates
3. pool growth rates across imputations
4. chain pooled growth rates into an index

The pooled index is then constructed by chaining pooled growth rates:

$$P_t = \prod_{s=1}^{t} (1 + \bar{g}_s)$$

Application 1: Vienna apartment market

Controlled benchmark setting:

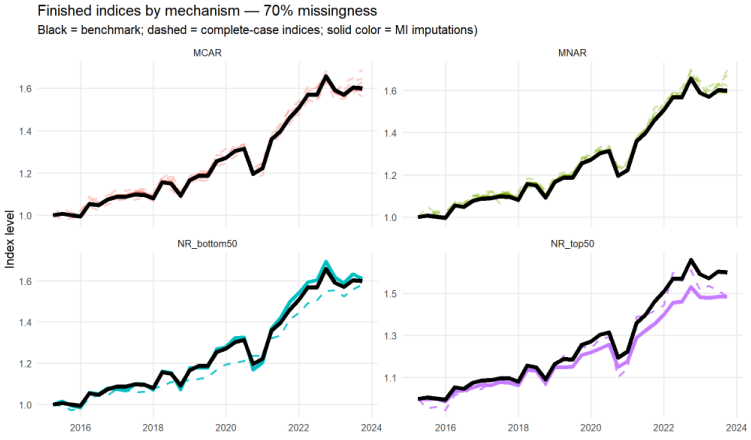
- Vienna apartment transactions
- 2015–2023
- more than 67,000 observations
- originally complete data

Design:

- introduce synthetic missingness
- compare complete-case and imputation-based indices
- evaluate both against known full-sample benchmark

This application separates variance effects from systematic composition effects.

Vienna result: thick markets are surprisingly robust



⇒ In a large, homogeneous residential market, missingness mostly increases dispersion. It does not necessarily induce large index bias.

What does the Vienna apartment example show?

We imposed severe missingness:

- up to 90% missing values,
- non-random missingness,
- and truncation designs.

⇒ In large homogeneous markets, missingness mainly increases dispersion.

Bias appears when missingness removes entire parts of the price-quality distribution.
For example:

- delete the upper half of transactions by price,
- then impute high-end properties using only lower-end observations.

  ⇒ In this case, the imputation model must extrapolate beyond the observed support, and bias remains.

Application 2: Austrian office market

Now consider a much harder setting:

- Austrian office-unit transactions
- 2015–2024
- about 3,000 observations
- roughly one twentieth of the Vienna apartment sample

This market is:

- thin
- heterogeneous
- spatially uneven
- characterized by substantial missingness in key quality variables

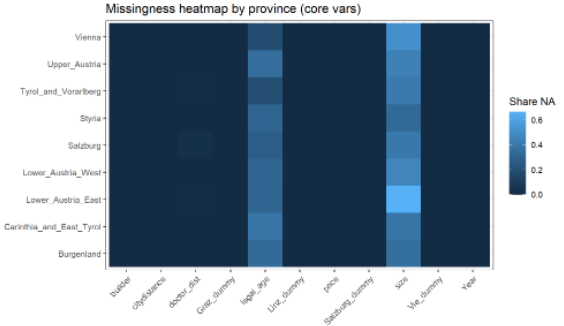
Missingness in the Austrian office market

Variable	Missing observations	Missing share (%)
size	1,357	45.2
legal_age	783	26.1
doc_dist	24	0.8
citydistance	2	0.1
city_within_20km	2	0.1
nearest_city	2	0.1
All remaining variables	0	0.0

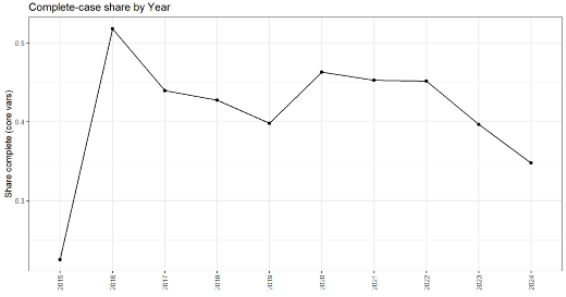
Important points:

- missingness concentrated in two key quality variables
- complete-case analysis drops 60% of observations
- missingness is not random $\rightarrow$ it varies across regions and over time

Geographic and temporal variation in missingness



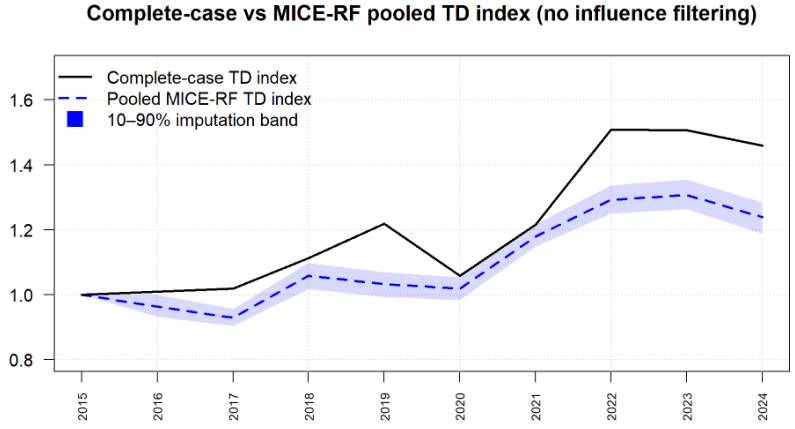
Regional variation



Temporal variation

Complete-case estimation can change the composition of the estimation sample over time $\rightarrow$ bias

Imputing missing values matters in the office market



Note: Solid line: complete-case TD index. Dashed line: pooled TD index constructed from $m = 100$ multiply imputed datasets (method: MICE-RF). Shaded areas represent the 10–90% cross-imputation dispersion of chained index levels.

Does the imputation algorithm within the Multiple imputation framework matter?

Methods compared

- MICE with random forests (MICE-RF)
- default MICE (PMM)
- linear regression imputation
- mean imputation

Main findings

- MICE-RF and default MICE (PMM) perform very similar
- linear regression becomes unstable under collinearity
- mean imputation performs clearly worse

The main gain comes from restoring observations and sample composition, not from optimizing the imputation algorithm.

What imputation can and cannot do

Imputation can:

- restore incomplete observations to the estimation sample
- reduce sensitivity to changing sample composition
- quantify missing-data sensitivity through dispersion bands

Imputation cannot:

- recover market segments with no observed support
- eliminate bias from structurally unobserved price–quality regions
- substitute for better primary data collection

Multiple imputation improves index construction when missing observations are informative but still connected to observed support.

Practical implications for index compilers

- Avoid complete-case estimation when missingness varies across regions or time
- Monitor whether missingness changes the effective estimation sample
- Flexible multiple-imputation methods are often sufficient
- Publish sensitivity bands where feasible
- Imputation cannot recover market segments with no observed support

Main takeaways

1. Thick homogeneous markets are relatively robust to missing characteristics.
2. In thin heterogeneous markets, complete-case estimation can change the effective market being indexed.
3. Multiple imputation helps mainly by restoring sample composition.
4. Chained indices should pool growth rates rather than Rubin-style coefficient estimates.

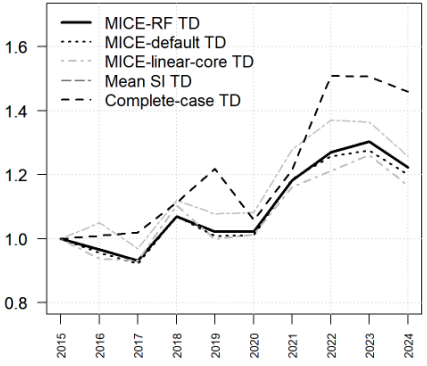
Thank you for your attention!

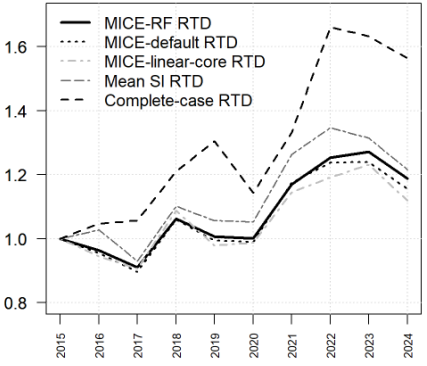
miriam.steurer@uni-graz.at

If you want to read the paper, it is available under:

https://www.nber.org/papers/w35139

Data completeness dominates index-method choice





On incomplete data, index-method differences look large. After imputation, TD and RTD indices become much more aligned.