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

Miriam Steurer and Sabrina Spiegel

Discussant: Christopher R. Bollinger

This paper proposes an approach for imputing missing characteristic data in real estate data used for constructing price indices. Price indices attempt to capture changes in the price structure while holding constant the "market basket of goods" purchased. In cases like the CPI, an artificial market basket is constructed (based on past consumption) and the cost of the market basket is computed and compared to prior years. In the real estate market, the consumption "basket" are the characteristics of the structures. These characteristics are not traded in the market specifically, only the composite good (a house or apartment, for example) is traded. To isolate the changes in price from different "market baskets", a typical approach is to run hedonic regressions on the available transactions controlling for characteristics and time period. The time period coefficient is then used to construct the price index.

However, in many real estate data sets, key characteristics are missing. The example here is the office market in Austria where both size and age are often missing. This presents significant issues for estimation.

The paper begins with a complete data set - apartment transactions in Vienna. These data are apparently of high quality with no missing characteristics. From these data they construct a set of example data with various missing patterns and structures including Missing Completely at Random (MCAR) and two types of Missing Not At Random (MNAR).

The first type of MNAR is where the value of the missing characteristic is related to the missingness pattern, choosing low values of the characteristics to have higher missing rates. The second type induces missingness in characteristics when price is either higher or lower (two different designs). Each of the missing patterns is done for a variety of base missingness rates ranging from 30% to 90%.

The authors then use these data sets to investigate the performance of their proposed MICE-RF approach relative to complete case analysis first and then two typical simple imputation approaches second. The authors find that the MICE-RF approach works quite well in MCAR and when missingness is only related directly to the value of the missing variable. It performs less well - as would be expected - when missingness is related to the price (complete case does not perform well either).

The MICE-RF algorithm has two parts. MICE is multiple imputation chained questions, and RF is random forests. The MICE algorithm is a relatively straightforward imputation procedure as described in the appendix, where imputations are based on a predictive model, conditional on all other variables. The imputation is iterated (as some conditioning variables may be missing) until convergence. The RF is used for the predictive model for characteristics. A model is initially trained and then applied.

The second part of the paper uses the approach to investigate the implications of actual missing data in office unit transactions throughout Austria. Two key variables are missing: size (space) and age. They find that the missing corrected index is lower and turns more quickly during covid than the complete case.

I think this is a very well done and useful paper. Some of the crucial positives I see:

- They do a very nice job using real data (which is complete) and then simulating missing data patterns. This is probably the most important part of this paper, evaluating the approach under different scenarios.
- The patterns they choose are quite sensible, both missing at random, but also not at random, often overlooked.
- The resulting comparisons to complete case and some comparison to two other approaches demonstrate the advantages of their approach.
- The main application - office rents - is interesting. And they make a great case for why it's interesting.
- The combination of using random forest within the imputation is valuable.
- The real estate markets are an interesting and important special case of imputations: we need the full data to properly compute a market price index.
- They propose a nice approach to combining the imputation data sets to constructing the index. Rather than simply averaging the estimated coefficient of interest and then constructing the index, they construct the index for each imputation after the main regression, and then geometrically average the indices. I think this is right.
- Both in my own work (Bollinger and Hirsch, 2005), and in Little and Rubin (2002) we note that it is crucial for imputation to be done in the context of the model. These authors do indeed follow this approach.

Initially, the choice of competing approaches was rather limited. The authors have experimented with further options. I suggested using a random forest approach but for a single equation imputation. The authors have tried some additional approaches. In some sense the results were slightly disappointing in that the other approaches do almost as well. I think highlighting alternative approaches is useful, in that different applications may require different approaches. There are three approaches I might suggest could be examined more closely.

The first approach is a "hot deck" type approach (often used by U.S. Census). Rather than predict, simply draw an observation from the observed data which match on other characteristics and use that to fill in the missing value. This can easily be done in a

multiple imputation with multiple draws. Census breaks the population in many cells based on the observed characteristics and thus the underlying prediction model is quite non-parametric. An advantage of the approach is no first stage estimation is necessary.

A second approach is that of Abrevaya and Donald (2017), which uses a GMM approach to dealing with missing data. The approach can be used for estimating a main specification when regressors are missing. It uses all observations. The moment conditions are familiar in general, for a linear model (like that fit for the main equation here), the OLS normal equations are the starting point. When an X variable is missing, the equations are modified for those observations but still enter the estimation. This makes use of all variances and covariances in the observed data.

A final approach is to use weights for the complete case data: reweighting complete cases to match key characteristics from the known population. Ideally, the weights would be based upon some kind of office or housing census which provided the population characteristics. Then the complete cases are weighed up to be representative of the population distribution on those characteristics (e.g. size, bedrooms, age, etc). Weighting the prices appropriately.

Two approaches that have been considered for missing data when it may not be missing at random are sample selection and inverse probability weighting. Both of these could be considered here, to use complete case estimation and still control for the missing data. Bollinger and Hirsch (2013) and Bollinger et al (2019) consider these. However, this may not work well here, as it is not clear what kinds of exclusion restrictions might be available which predict missingness, but do not affect the main equation.

The next comments go beyond the scope of the paper. These concern how one might construct a price index in general.

The paper uses a very standard approach to estimating a price index for real estate. While there are variations on this, using a hedonic model to "normalize" the price growth. However, another approach would be to follow something closer to a CPI: fix the market basket. If there were – and there may be in administrative records, such as building permits – a census of all properties, we could use that as the basis for a market basket. Then, like the CPI, weight the observed transaction prices in each year up to that census level. As I note above, weighting can be a solution to missing data, and this would eliminate the step of imputation. The price index then doesn't the hedonic model, just time averages.

I have a concern about how the index is constructed from the regression, please note the authors follow a very typical approach. The approach taken is to take the δ coefficient on time from the hedonic pooled time regression and form $P_t = e^{\delta_t}$. This is a typical approach where $\mathrm{E}[\ln \mathrm{P}\_\{\mathrm{it}\}|\mathrm{X}]$ is estimated with OLS (or similar). However, $\mathrm{E}[\mathrm{P_t}|\mathrm{X}] \neq e^{E[lnP_t|X]}$ in general. Even if we believe in a log-normal assumption $\mathrm{E}[\mathrm{P}_{it}|\mathrm{X}] = e^{\{\mathrm{E}[\ln \mathrm{P_{it}}|\mathrm{X}]+ \mathrm{v}[\ln \mathrm{P_i}|\mathrm{X}]/2\}}$. I think, in particular in real estate, that variance term is changing over time. If not, if the variance of the conditional log price series is constant, then it will drop out of the index. However, that seems unlikely, although it is an easily testable hypothesis. This is a general criticism of trying "unlog" a log-linear specification. Another approach is to estimate medians or means of the actual price series, with possible non-linear components. The random Forrest approach could be very nice for this.

The key assumption in these hedonic models is that the prices of the characteristics are constant over time. If, for example, the price of one characteristic is rising, this is going to be imperfectly captured because that may lead to (or be caused by) changes in consumption of that characteristic. I've always been skeptical of this for any "long run." My suggestion here is again a weighted average price series, keeping the characteristics constant.

I want to stress that the paper written by Miriam Steurer and Sabrina Spiegel is quite good. It takes a series empirical issue in an important empirical exercise that is used widely by researchers, practitioners and policy makes and offers a well thought through approach. I especially commend the authors for carefully evaluating their approach on a data set drawn from the real world, but where the missing patterns were controlled by them. This allows their approach to be evaluated and for those who use it in other settings to see both its advantages and disadvantages. I think this paper is an excellent example of how to implement a methodological improvement, and I should be carefully read by researchers constructing or using housing price indices.

Abrevaya, Jason, and Stephen G. Donald. (2017), "A GMM Approach for Dealing with Missing Data on Regressors." *The Review of Economics and Statistics* 99, no. 4: 657–62.

Bollinger, Christopher R. Barry T. Hirsch, Charles Hokayem and James P. Ziliak (2019), "Trouble in the Tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch." *Journal of Political Economy*, vol. 127, no. 5, pp. 2143-2185.

Bollinger, Christopher R. and Barry T. Hirsch (2013), "Is Earnings Response Ignorable?" (with Barry Hirsch), *Review of Economics and Statistics*, vol. 95, no. 2, pp. 407-416.

Bollinger, Christopher R. and Barry T. Hirsch (2006), "Match Bias in the Earnings Imputations in the Current Population Survey: The Case of Imperfect Matching" *Journal of Labor Economics*, Vol. 24, no. 3, pp. 483-520.

Roderick J. A. Little and Donald B. Rubin (2002), *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: Wiley-Interscience.