Modeling the Fragmented Archive: A Missing Data Case Study from Provenance Research

2017-11-27

Historians grapple with missing information constantly. While there are many statistical tools for gauging the impact of missing source data on quantitative results and conclusions, DH researchers have rarely deployed these tools in their work. This paper presents one implementation of data imputation used in the study of the New York City art dealer M. Knoedler & Co. Demonstrating the significant contribution imputation had on our study and its conclusions, this paper will discuss specific, practical rhetorical strategies, including static and interactive visualization, for explaining this methodology to an audience that does not specialize in quantitative methods.

Missing Data in the Digital Humanities

Miriam Posner has argued that both data structures and rhetorical conventions for computing with missing information, uncertainty, and highly subjective/viewpoint-contingent knowledge remains a key desideratum of DH scholarship. (Posner 2015) Several attempts have been made by the information science community to express uncertainty in a structured format, ranging from generalized ontologies for reasoning in a networked world (K. J. Laskey et al. 2008), as well as more specific projects such as the *Topotime* library for reasoning about temporal uncertainty. (Grossner and Meeks 2013)

However, many DH projects have sidestepped these approaches. Matthew Jockers, for example, has asserted that the availability of full text is becoming such that literary historians will no longer have to be concerned about drawing a representative sample. (Jockers 2013, 7–8) More commonly, though, scholars have attempted to carefully constrain their conclusions based on what they know to be missing from their data. Theorizing and documenting the difference between one's data set and one's subject has become a genre of DH work unto itself. Katherine Bode has argued that such documented datasets should be understood as *the* object of DH inquiry. (Bode 2017)

While statistical literature on the problem of missing-data imputation is quite mature (see Gelman and Hill 2006 for a valuable review), few DH research projects have openly explored the use of statistical procedures for reckoning with missing data, nor have they grappled with how to theorize and present such imputation in the context of their home disciplines. (An important exception includes Brosens et al. 2016) Bode, for one, has explicitly rejected such approaches, arguing (without specific evidence) that quantitative error assessment

cannot be usefully performed in historical analysis. (Bode 2017, 101)

We argue that such methods should be *central* to data-based digital humanities practice. Simulation and imputation allow us to realize multiple, sometimes conflicting assumptions about the nature of missing data. In doing so, these affordances allow us to evaluate how certain assertions may propagate their assumptions through the transformations we perform on our sources.

Case Study: Modeling M. Knoedler & Co.'s Business from Sparse Stock Books

As part of a research initiative into data-based approaches to the study of the art market, we are investigating the changing strategies of the New York City art dealer M. Knoedler & Co., whose stock books have been encoded by the Getty Research Institute (`http://www.getty.edu/research/tools/provenance/search.html`). Based on these transaction data, we have built a predictive model that classifies whether a given artwork would result in a profit or a loss, using a host of variables such as how much money the work of art originally cost, the genre and size of the work, their prior relationships with buyers and sellers, and the time the work remained in stock before it was sold, to name but a few. Predictive modeling illuminates complex relationships between these variables and highlights unusual sales for further archival research.

As informative as these stock books are, however, many of their notations are partial: Knoedler's staff may have neglected to record the date of sale; there may be a listed purchase without a description of the type of work (i.e. portrait, landscape, etc.); the identity of the buyer, and whether they were a first-time customer or a well-known shopper, may also have gone unrecorded. Because our random-forest-based model (Liaw and Wiener 2002) does not allow missing values, we must either discard incomplete records (and thus eliminate nearly half of the records from consideration), or we must find ways to impute values for our predictor variables.

While it is impossible to perfectly reconstruct these missing records, it *is* possible to operationalize educated guesses about their possible values. (Figure 1) Purchase and sale dates for artworks, for example, can be predicted with some accuracy based on their location in the roughly-chronological series of stock books. Likewise, unknown genres can also be imputed as a function of the existing distribution of genres across stock books, with, e.g. abstract paintings being far less common in the pre-20th c. books than in the later ones. By defining an informed range of possibilities for these missing data, and then sampling from that range, we can produce ensemble

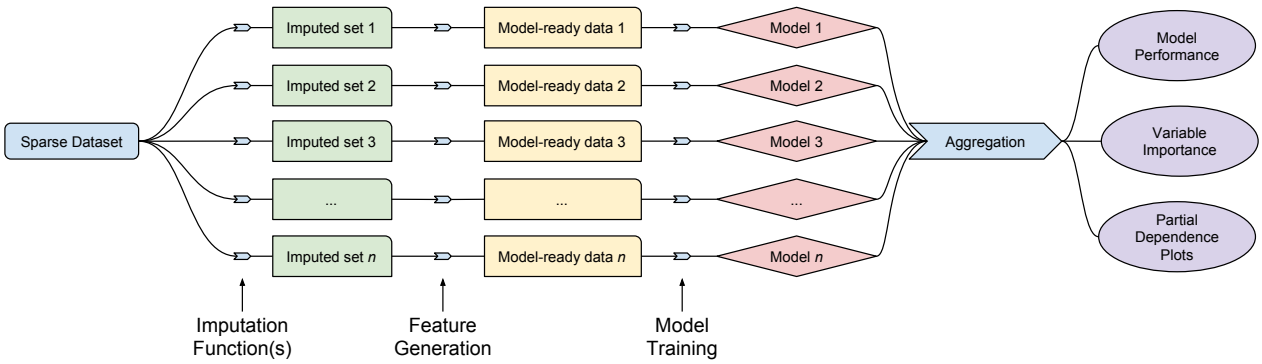

Figure 1: Schematic workflow for imputing missing data, producing derivative features, building models, and then aggregating statistics from the multiple models produced.

models and results that provide a more nuanced representation.

Figure 2(a) shows the marginal effect of artwork genre on Knoedler's chance of turning a profit across three periods of their business, only considering around 20,000 "complete" cases from the Knoedler transaction records (approximately 60% of the known transactions they made.) A first glance suggests that history paintings were markedly less profitable between 1900 and 1935, while still lifes became significantly more profitable than other genres after 1935.

However, 2(b) shows the results not from 1 model, but from 500 models, each trained on a slightly different set of stochastically-imputed data. By visualizing one bar for each model, this plot drives home the effect of increased uncertainty on these measurements, while visually foregrounding the crucial methodological decision - 500 models instead of 1 - in a way that a box plot or other summary visualization method does not (at least, not in the eyes of a reader unused to reading such idioms.) The apparent advantage of still life in Knoedler's post-1935 business has evaporated, although the notably-lower value of history paintings between 1900-1935 may have withstood this simulation of uncertainty. While this model affirms that genre is largely an anachronistic construct that has little effect on prices, these results complicate a simplistic reading by indicating that, in some cases, there *is* a significant relationship that must be reckoned with.

Figure 3 shows a similar comparison of complete case vs. imputed data for a continuous variable: the time a painting spent in stock. Both 3(a) and 3(b) support the conclusion that not only did a longer time in stock contribute to lower chances of turning a profit, but that Knoedler's window for making a profitable sale grew throughout the lifetime of the firm, from around 2 years before 1900, to more than 5 years after 1935. The increased uncertainty added by the multiplicity of models in 3(b) discourages the kind of over-interpretation that

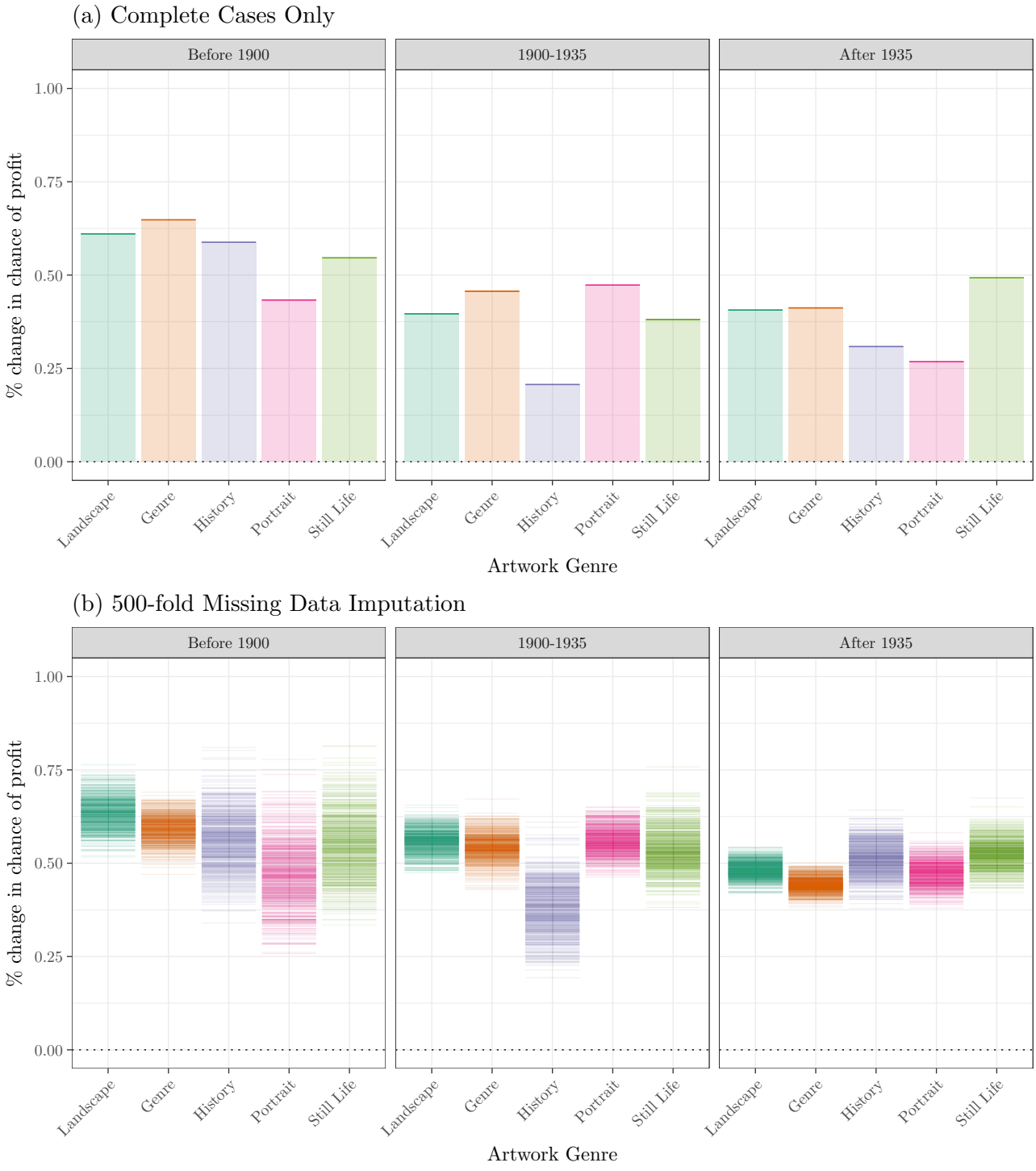

Figure 2: Partial dependence plots illustrating the marginal effect of artwork genre on Knoedler's chance at profitability.

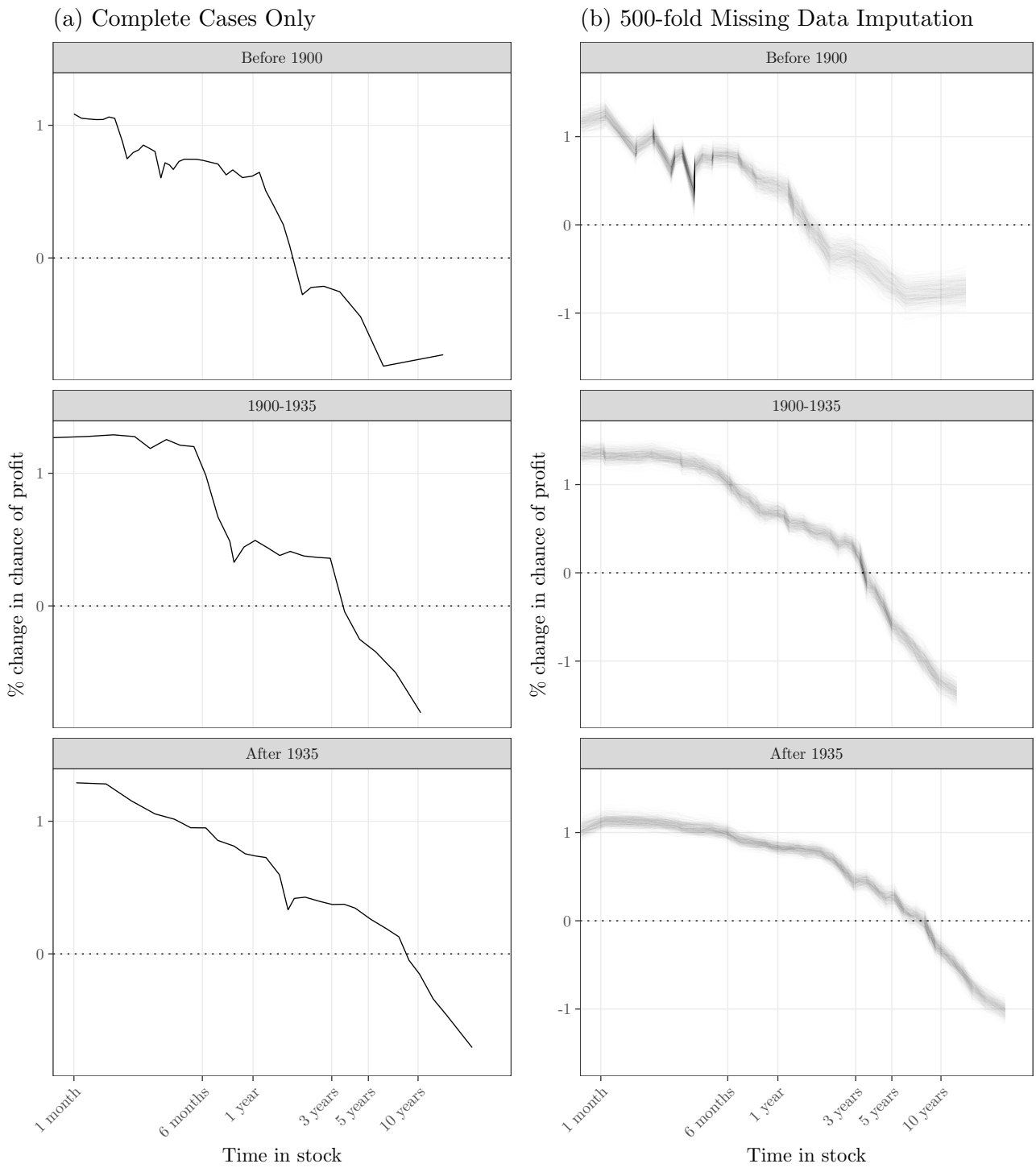

Figure 3: Partial dependence plots illustrating the marginal effect of time in stock on Knoedler's chance at profitability.

the seeming-precision of 3(a) allows. However it also demonstrates that, even in the face of so many missing or imprecise dates in the Knoedler stock books, we can still recover meaningful quantitative conclusions.

Figure 4: Screenshot of an interactive application allowing users to modify imputation assumptions and see the effect on modeling and analysis results.

These static visualizations are easily enhanced through animation that shows the buildup of individual model characteristics into aggregate confidence intervals. (Lincoln 2015) We have also experimented with interactive applications (Figure 4) that allow the user to specify different imputation assumptions, and then immediately see the downstream results on our predictive models, reinforcing the close relationship between starting assumptions and modeled conclusions. (An early demo of this work: `https:`

//mdlincoln.shinyapps.io/missingness/)

Computationally, these imputations are simple, perhaps even simplistic. More complex approaches, such as iteratively modeling every missing variable (Buuren and Groothuis-Oudshoorn 2011), might lead to more accurate modeling. However, these less parsimonious methods are more opaque to humanities scholars. Operationalizing the historian's habit of educated guessing and thoughtful assumptions, and visualizing those operations straightforwardly, may allow missing data imputation to work its way into the accepted suite of DH methodologies.

Works Cited

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78 (1): 77–106. doi:10.1215/00267929-3699787.

Brosens, Koenraad, Klara Alen, Astrid Slegten, and Fred Truyen. 2016. "MapTap and Cornelia: Slow Digital Art History and Formal Art Historical Social Network Research." *Zeitschrift Für Kunstgeschichte* 79: 1–14.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in *R*." *Journal of Statistical Software* 45 (3). doi:10.18637/jss.v045.i03.

Gelman, Andrew, and Jennifer Hill. 2006. "Missing-Data Imputation." In *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 529–43. Oxford: Cambridge University Press.

Grossner, Karl, and Elijah Meeks. 2013. "Temporal Geometry in Topotime." Stanford University Libraries. `http://dh.stanford.edu/topotime/docs/TemporalGeometry.pdf`.

Jockers, Matthew Lee. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Laskey, Kenneth J., Kathryn Blackmond Laskey, Paulo C. G. Costa, Mieczyslaw M. Kokar, Trevor Martin, and Thomas Lukasiewicz. 2008. "Uncertainty Reasoning for the World Wide Web." W3C Incubator Group Report. World Wide Web Consortium (W3C). `https://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/`.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. `https://www.stat.berkeley.edu/~breiman/RandomForests/`.

Lincoln, Matthew D. 2015. "DataGIFs: Animate Your Visualizations for Fun and Clarity." *Matthew Lincoln, PhD*. `https://matthewlincoln.net/2015/12/18/datagifs-animate-your-visualizations-for-fun-and-clarity.html`.

Posner, Miriam. 2015. "What's Next: The Radical, Unrealized Potential of Digital Humanities." *Miriam Posner's Blog*. `http://miriamposner.com/blog/whats-next-the-radical-unrealized-potential-of-digital-humanities/`.