History as a data science: Missing data imputation on the the slave voyages dataset

Phillip Tran
jqt1@rice.edu
Rice University
Houston, Texas, USA

Arlei Silva
arlei@rice.edu
Rice University
Houston, Texas, USA

ABSTRACT

One could argue that Historians are far from becoming the next victims of automation and AI. But with the increasing popularization of digital history databases, we are now able to apply data science to historical data. Computational History is an emerging field that leverages recent advances in digitalization, data science, and machine learning toward a better understanding of our past. However, history databases are the result of human-intensive work analyzing very limited historical evidence, and thus, missing data is a prevalent problem in these datasets.

In this paper, we investigate the missing data imputation problem for digital history databases. Slave voyages, which is the largest collection of records of forced relocations of Africans to and within the Americas, is applied as a case study. We first characterize key properties of the dataset, including the prevalence of missing data—nearly 80% of the entries are missing and the majority of attributes have at least a 90% missing ratio. Next, we assess the potential for data imputation approaches to exploit the correlations in the data to accurately predict missing values. Finally, we apply a representative set of imputation methods to slave voyages and evaluate their performance in terms of prediction error.

Our results illustrate the challenges of imputing missing data in digital history databases. Historical data is highly heterogeneous, and the missingness in the data is far from random. However, we also show that some imputation methods achieve promising results.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; *Vagueness and fuzzy logic.*

KEYWORDS

datasets, missing data imputation, slave voyages, computational history, digital history

ACM Reference Format:
Phillip Tran and Arlei Silva. 2022. History as a data science: Missing data imputation on the the slave voyages dataset. In *Proceedings of KDD Undergraduate Consortium (KDD-UC '22).* ACM, New York, NY, USA, 7 pages.

KDD-UC '22, August 14–18, 2022, Washington, DC

1 INTRODUCTION

Data science and machine learning have gone from a narrow set of applications (e.g. market-basket analysis [1], information retrieval [30], recommender systems [34]) to impacting almost every area of knowledge. From drug discovery [6] to mathematics [23] to autonomous driving [2], data-driven methods have been advocated as a cheaper—and sometimes more effective—alternatives to experts. Key to the success of these approaches is the availability of (big) data, often collected from users or generated via simulations.

In small data settings, one can either improve data collection to enable the application of data-hungry models or develop new methods that can be effective using small datasets. However, there are many scenarios where acquiring high-quality data is extremely difficult. In such applications, domain experts might still be needed, but data-driven approaches could assist experts in human-intensive tasks. This paper focuses on one such application, which is History. In particular, *Computational History* is an emerging field that aims to leverage digital history databases and data-driven methods for the discovery of historical knowledge.

Yuval Noah Harari, in the best-seller *Sapiens: A brief history of mankind* [18], argues for the study of History *"not to know the future but to widen our horizons, to understand that our present situation is neither natural nor inevitable, and that we consequently have many more possibilities before us than we imagine."* However, information about past events is highly incomplete, and thus, History relies almost completely on the intensive work of experts.

From a data science lens, a significant part of the job of a Historian can be seen as (an extreme version of) *missing data imputation.* While developing a narrative for a past event is a complex process, such narratives can be built upon existing historical evidence, which has been increasingly made available to the History community in digital form. Unsurprisingly, such digital history databases also suffer from missing values, which is an obstacle to historical discovery. To address this problem, this paper investigates missing data imputation for historical databases.

As a case study, we consider data imputation in the *slave voyages* dataset[1]. Slave voyages contain records of forced relocations of more than 12 million African people to and within the Americas between 1514 and 1866 [12–14]. Covering over 36,000 voyages, each with up to 274 attributes, the project is a result of the continuous effort of many historians, librarians, curriculum specialists, cartographers, computer programmers, and web designers over more than two decades. However, 78% of the values in the database is missing, and 135 attributes have a missing rate of at least 90%. Data imputation for slave voyages can improve the accuracy of the statistics and visualizations of the dataset.

[1] https://www.slavevoyages.org

Our work investigates data imputation approaches for digital history databases. We characterize the slave voyages dataset based on statistics and distributions of variables (e.g. slaves/voyage, mortality rate) to illustrate key properties of the data. We also analyze the missingness patterns in the data, demonstrating the challenges and the potential of data imputation methods in our dataset. Finally, we apply existing data imputation methods to the slave voyages dataset and compare their results.

We summarize our contributions as follows:

- Our work is—to the best of our knowledge—the first formal exploration of data imputation for digital history databases;
- We propose a methodology for data imputation in digital history databases, which can support further research on this problem;
- Our preliminary results demonstrate the potential of data imputation approaches for digital history data.

2 RELATED WORK

Digital History is the use of communication and information technologies to facilitate the access to historical knowledge [7]. In the last decades, there has been a great effort to digitize historical records in many formats (e.g., text, images, videos). Moreover, today, new historic records are born already in digital format [39]. This has motivated recent research on visualization, information retrieval, and database technology for historical data. Examples of digital history databases include the Texas Slavery [38], Gilded Age Plains City [27], Spatial History [40], Railroads and the Making of Modern America [37], and Slave Voyages [36].

Our work applies the slave voyages dataset as a case study. It is a result of the work started by H. S. Klein and others [20, 21] based on an idea proposed by D. Eltis and S. Behrendt [9, 12–14]. The data was initially shared as a CD-ROM until it became a website in 2008. Today, slave voyages is an important source of information for Historians studying the transatlantic slave trade and the history of slavery [11, 21, 33]. More information about the database can be found in the project website.

The advent of digital history has motivated new computational approaches that go beyond managing and accessing information. In particular, *Computational History*—also known as *Histoinformatics*—aims to apply machine learning and other data-driven techniques in the discovery of historical knowledge. Viewing History as a data science creates many opportunities to assist historians in validating and exploring new hypotheses about past events [29]. As an example, in [32], the authors apply network science—clustering and centrality measures—to better understand the relationships between different actors and their role during the Byzantine Empire.

Our work focuses on missing value imputation, which is a key step in many machine learning and data analytics pipelines [3, 15, 25]. For instance, in Bioinformatics, DNA microarray data is often missing due to image corruption or resolution issues. In clinical studies, many subjects leave during the experiment (e.g. due to death) [28]. There are several approaches for data imputation [24, 25], which can be divided into univariate and multivariate methods. A representative set of imputation methods is described in the next section. A key question in the study of missing data is the reason why the data is missing, which can be completely random, random, or not random [25]. The source of missingness affects the expected performance of imputation methods. Thus, model-based solutions attempt to model the missingness mechanism [25, 35].

Digital history databases [7] are also affected by missing data, mostly due to the lack of reliable historical evidence regarding a particular historical event. Moreover, many historical databases integrate multiple sources of data that might overlap only for a subset of attributes [13]. In the case of the slave voyages dataset, roughly 80% of the values is missing from the original dataset. Out of the 36,108 voyages, only 22,883 have a starting date, and 751 have a total number of adults embarked. This work is focused on data imputation methods to alleviate the missing data problem in the slave voyages dataset.

3 PROBLEM AND METHODOLOGY

In this section, we formalize the missing value imputation problem and describe a representative set of solutions to the problem.

3.1 Missing value imputation problem

The missing value imputation problem consists of predicting missing values based on observed ones. Missing value imputation is often a better alternative to completely removing objects that are not fully observed from the database. This is particularly true in the case where many values are missing. For instance, in the case of the slave voyages dataset, every voyage has at least one missing attribute. Instead, imputation methods exploit correlations in the data to fill the missing values.

The main motivation for imputation methods is enabling the application of techniques designed for complete data (e.g., classification, clustering, regression) to datasets with missing values. As a consequence, missing value imputation is a key component of many data science pipelines.

3.2 Missing value imputation approaches

Here, we will briefly describe missing value imputation approaches proposed in the literature. For a more detailed overview of these approaches, we refer to [24, 25].

3.2.1 Univariate imputation. Univariate imputation methods only consider the target attribute as an input. Their main advantages are their simplicity and efficiency. However, different from multivariate approaches, they do not leverage correlations between the target and other attributes in the dataset. We consider *Mean Imputation* as an example of a univariate imputation method.

Mean Imputation [10]: A missing value is simply predicted as the mean observed value of the corresponding attribute.

3.2.2 Iterative multivariate imputation. Iterative multivariate imputation methods exploit the correlations between observed values to predict missing values iteratively. Let $a_1, a_2, \ldots a_D$ be the set of attributes in the database. The method selects an attribute a_i and learns to predict a_i as a function of other attributes a_j ($j \neq i$). The learned predictor is then applied to the missing values of a_i. Next, a new attribute is selected and the process is repeated (using predicted values for a_i). At the end of the process, all missing values in the database are imputed. The different methods listed below

apply this general approach with different predictors. Because we will focus on numeric values, our predictors are regression models.

Bayesian Ridge Regressor [19]: Ridge regression learns regression weights $\mathbf{w}$ by minimizing $||A\mathbf{w} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{w}||_2^2$, where A is a matrix with inputs as rows, $\mathbf{b}$ are outputs in the training data, and Γ is a a diagonal matrix. In Bayesian Ridge Regression, weights $\mathbf{w}$ are assumed to follow a zero-mean Gaussian distribution $N(0, \Sigma)$.

Decision Tree Regressor [26]: The *Decision Tree Regressor* is a non-parametric supervised learning approach based on decision trees. Branches in the tree represent tests (or decisions), and leaves are associated with continuous value predictions. Learning the optimal decision tree for a given input is NP-hard. Here, we apply CART (binary trees) as an efficient heuristic for decision tree learning.

Extra Trees Regressor [17]: The Extra Trees Regressor is an efficient tree ensemble approach. It constructs a collection of decision trees for random subsets of features in the data. Moreover, the decision rules are selected based on randomly drawn thresholds for each feature—as opposed to the most discriminative one. Predictions are computed as the mean prediction over the trees.

3.2.3 KNN-based multivariate imputation [4]. KNN-based approach for missing value imputation. Nearest neighbors are computed based on observed values. Missing values are imputed as the mean value of the k-nearest neighbors with observed values.

While our list of missing value imputation approaches is representative, it is not exhaustive. Classical imputation approaches can be found in [24, 25]. Examples of approaches not discussed in this section include matrix factorization [5], expectation-maximization [25, 35], and autoencoders [31].

4 RESULTS

In this section, we characterize key properties of the slave voyages dataset and present some preliminary results on missing data imputation methods for slave voyages.

4.1 Dataset

The slave voyages dataset is composed of 36,108 voyages, where each can contain up to 274 attributes. Key statistics of the database are shown in Table 1. The dataset, with associated descriptions and visualizations, is publicly available in project website[2].

# of voyages	36,108
# of attributes	274
# of ships	9,440
avg of slaves/voyage	329.9
avg mortality rate	12%
avg voyage duration (days)	435.12
avg crew size	30.12
Time period	1526-1844

Table 1: Dataset statistics.

Variables (or columns) in the database are divided into the following types: technical, data, and imputed. The technical variables assign unique identification numbers for each of the voyages, document the changes to the database since publication, and provide voyage groupings for computing the total numbers of slaves embarked (SLAXIMP) and disembarked (SLAMIMP). The data variables consist of features that come directly from historical records. Lastly, the imputed variables are features either computed from the data variables such as slave mortality rates or infer data not found in existing documentation on the basis of patterns observed in data variables such as embarkations for voyages on ships of similar tonnage and rigging in the same period of time [12]. Each variable in the database has one of the following formats: F (numeric), A (alphanumeric), or DATE (date).

Figure 1 exhibits a characterization of the slave voyage dataset based on distributions of key variables: slaves per voyage, mortality rate, voyage duration, crew size, voyages per ship, and voyages per year. Understanding the data distribution is a key step for the application of data imputation approaches. We notice that slaves per voyage and voyage duration could be well approximated by a normal or log-normal distribution. Variables mortality rate, crew size, and voyages per ship are biased toward smaller values—these variables are constrained by physical or logical limits. Moreover, the variable voyages per year has a clear peak within a 50-year window (1750-1800).

4.2 Missing values

Here, we focus on analyzing the missingness in the slave voyage datasets. More specifically, we are interested in understanding how missingness is distributed over the different variables (or columns) in the data to better understand the potential for missing data imputation approaches to be effective in practice.

Figure 2 shows the distribution of the missingness rate for the variables. We notice that the large majority of the variables (135) are missing at a rate of 90% or higher. Some attributes, such as the number of adults embarked at second port of purchase (ADULT4), the number of children embarked at third port of purchase (CHILD5), and the number of infants disembarked at second place of landing (INFANT6) are missing for all voyages. On the other hand, the outcome of the voyages and the slaves on board as well as the period when the voyages took place all have no missing data. It is a common trend to see that the attributes pertaining to the number of slaves (e.g. number of children, number of dead adults) are mostly missing in the dataset.

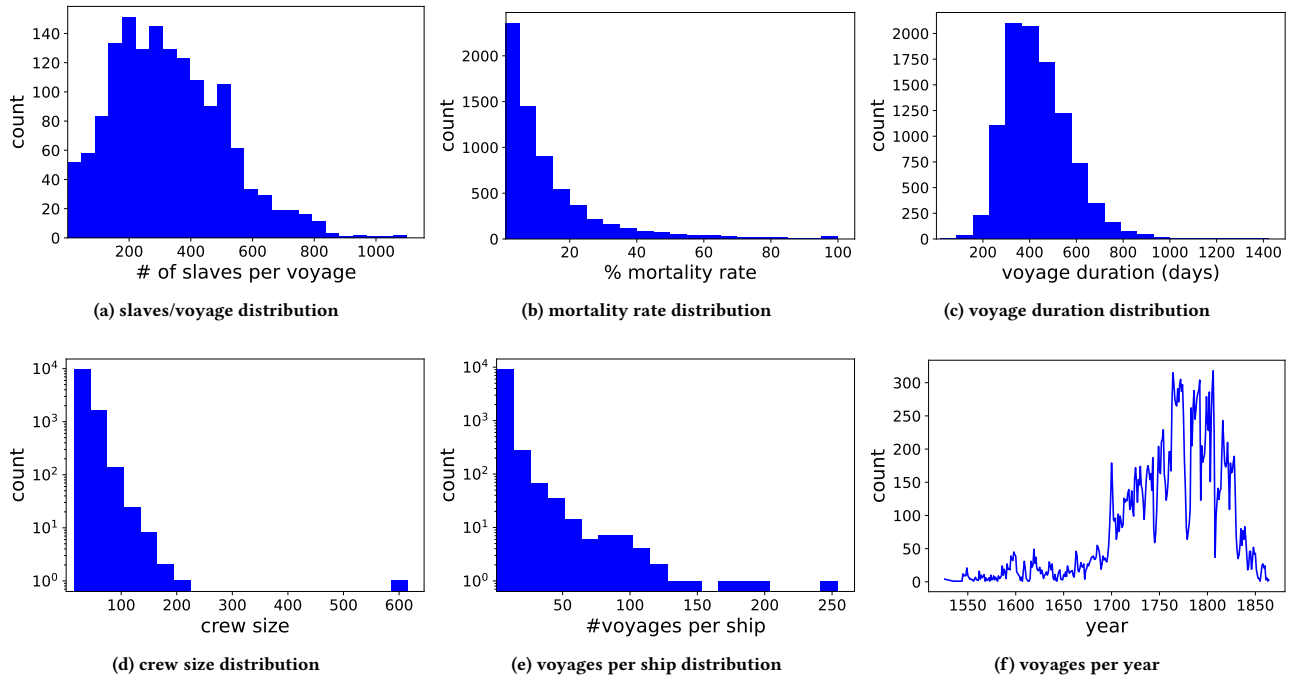
Figure 3 shows the missing values for the data variables. Positions in white are missing values. We notice that most entries (around 80%) are missing. This illustrates both the challenge and the need for missing value imputation in the slave voyages dataset.

4.3 Correlation and Missingness

In this section, we analyze how correlation in the slave voyages dataset can be exploited for missing value imputation. We are interested in pairs of variables that are correlated while not missing for the same rows. Intuitively, these are pairs of variables that can be used to predict each other.

To measure the difference in missingness for a pair of variables a, b, we apply the Jaccard similarity between the sets of missing rows and compute 1 minus this similarity to measure the mismatch,

[2]https://www.slavevoyages.org



(a) slaves/voyage distribution (b) mortality rate distribution (c) voyage duration distribution

(d) crew size distribution (e) voyages per ship distribution (f) voyages per year

Figure 1: Characterization of the slave voyage dataset based on distributions of key variables (e.g. slaves/voyage, mortality rate) over the set of voyages in the data. While some variables appear to have normal or log-normal distributions (e.g. slaves/voyage) others are skewed towards smaller values (e.g. voyages/ship). The number of voyages peaked from 1750-1800.

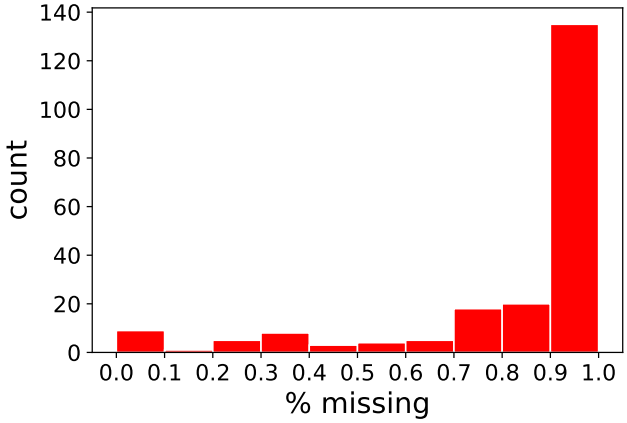


Figure 2: Distribution of missingness (in %) for all non-imputed variables in the slave voyages dataset. Most variables are missing for 90% or more entries (or rows).

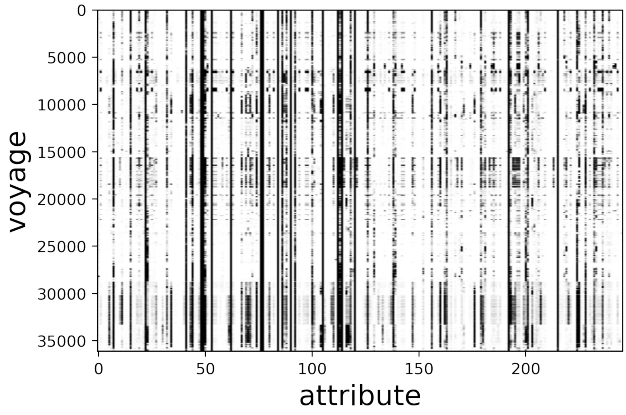


Figure 3: Missingness of attributes in a binary matrix where white represents missing.

$1 - (A \cap B)/(A \cup B)$—where A and B are the sets of missing row identifiers for variables a and b, respectively. To measure correlation, we apply different metrics according to the type of the variables involved. More specifically, we apply *Pearson's correlation* for pairs of numerical values [16], the *correlation ratio*—square root of the intraclass correlation—for numerical vs. nominal values [22], and *Cramer's V* for nominal values [8].

Figure 4 shows the results. As expected, most correlated pairs also present a high overlap in missingness. However, we can identify the following pairs of variables in the top right area of the plot: (TSLAVESD, SLAMIMP), (SLAVESP, TONNAGE), (SLAARRIV, SLAVMAX1), and (SLAS32, SLAXIMP). The values of 1-Jaccard and correlation for each of these pairs are given in Table 2. TSLAVESD is the total slaves on board at departure from last slaving port. SLAMIMP is the imputed total slaves disembarked. SLAVESP is

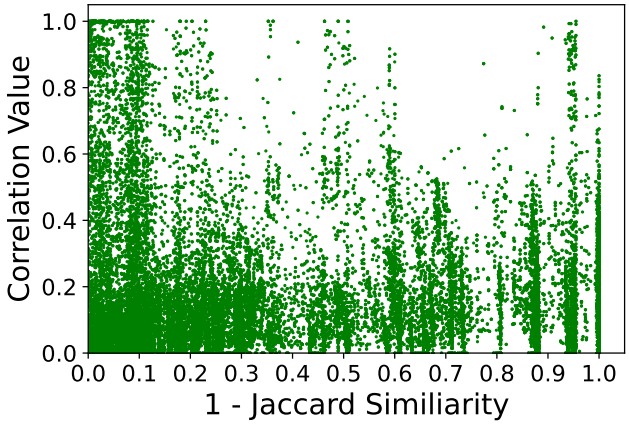


Figure 4: Difference (1-Jaccard similarity) in missingness vs correlation for every pair of attributes (or columns) in the database. Pairs at top right have high correlation but are not often missing at the same time.

the total slaves purchased. TONNAGE is the tonnage of the vessel. SLAARRIV is the total slaves arrived at first port of disembarkation. SLAVMAX1 is the total slaves embarked with age and gender identified. SLAS32 is the total slaves disembarked at first place. SLAXIMP is the imputed total slaves embarked.

Variables	1-Jaccard	Correlation
TSLAVESD, SLAMIMP	0.50	0.93
TSLAVESP, TONNAGE	0.57	0.57
SLAARRIV, SLAVMAX1	0.50	0.92
SLAS32, SLAXIMP	0.95	0.68

Table 2: Pairs of variables with both high difference in missingness (1-Jaccard similarity) and high correlation.

4.4 Missing value imputation

We will now focus on missing value imputation, which is the key problem addressed in this paper. We will consider only the imputation of numeric variables and leave the case of nominal variables and dates as future work. The missing value imputation approaches applied are the ones described in Section 3.2.

Implementations and dataset: The imputation approaches were implemented in Python using the *scikit-learn* library[3]. The source code for reproducing our results are available on github[4]. We also provide instructions on how to download the slave voyages dataset, which is publicly available in the project website[5].

Parameter settings: For *Bayesian Ridge*, we set the maximum number of imputation rounds to 10. For *Decision Tree*, we adjusted the number of features for the best split to be the square root of the number of attributes. For *Extra Trees*, we set the number of trees in the forest to be 10. For *K-Nearest Neighbor*, we assigned the value of k to 5, assuming uniform weights for each data point.

Target variables: The imputation approaches will be applied to the following set of variables: TSLAVESD, TSLAVESP, SLAARRIV, and SLAS32. A description of these variables is provided in the previous section. We apply correlation and 1-Jaccard similarity (see previous section) with the target variables to select other variables to be used for imputation. More specifically, threshold values of 0.50 and 0.25 were considered for correlation and 1-Jaccard, respectively.

Evaluation: We evaluate the imputation methods by predicting observed values in the database. For each variable, we select 90% of observed values for training and 10% for testing. The following metrics were applied in our evaluation: **MSE** (Mean Squared Error), **MAE** (Mean Absolute Error), and **MAX** (Max Error).

Results: Table 3 shows the results, in terms of missing value imputation error, achieved by different imputation methods applied to the variables TSLAVESD, TSLAVESP, SLAARRIV, and SLAS32 after normalizing the data between the values 0 and 1. Alongside the errors, the mean is also given. The lowest errors for each metric are underlined. The results show that no single method outperforms all the alternatives. *Extra Trees* achieve the best results for most of the variables, with the exception of SLAARRIV. *Decision Tree* also achieves good results, with exception of the variables TSLAVESP and TSLAVESD. *Mean Imputation*, which is the simplest approach, achieves the worst results. More sophisticated methods, such as *Bayesian Ridge* and *K-Nearest Neighbor* also achieve poor performance for most of the variables.

5 DISCUSSION

In this paper, we have investigated data imputation approaches for digital history databases. We have applied slave voyages, the largest record of forced relocations of Africans to America and within America, as a case study. In our experiments, we have characterized key properties of the dataset, especially regarding the missing values and the potential effectiveness of imputation methods. A representative list of imputation methods was applied to slave voyages, demonstrating the potential of these methods to improve digital history databases and support the work of Historians towards better understanding major events of the past.

This is a list of main findings of our work:

(1) Missing data imputation in digital history databases is a challenging problem. It differs from other popular settings (e.g. recommender systems) where the data is mostly homogeneous and large datasets are available;
(2) Slave voyages is an interesting dataset for the evaluation of missing data imputation approaches in the small-data regime. It contains a significant amount of missing data, and its variables cover a wide range of types and distributions;
(3) While most of the correlations in the slave voyages datasets are not necessarily useful for missing data imputation—as correlated pairs are often missing together—a small number of correlations can potentially enable high-quality imputation results, as shown in Section 4.3;

[3]https://scikit-learn.org/stable/
[4]https://github.com/Jessickmon/slaves-voyages-research
[5]https://www.slavevoyages.org

Attribute	Metric	Mean Imputation	Bayesian Ridge	Decision Tree	Extra Trees	K-Nearest Neighbor
TSLAVESD	MSE	0.0165	6.634e-5	0.0001	8.755e-6	0.0003
	MAE	0.1018	0.0014	0.0031	0.0005	0.0098
	MAX	0.5719	0.1419	0.1808	0.0671	0.1674
TSLAVESP	MSE	0.0406	0.0028	0.0011	0.0006	0.0046
	MAE	0.1751	0.0239	0.0118	0.0087	0.0394
	MAX	0.4873	0.4992	0.1800	0.1869	0.4529
SLAARRIV	MSE	0.0129	0.0002	4.241e-10	2.021e-7	0.0003
	MAE	0.0893	0.0035	8.010e-7	1.125e-5	0.0045
	MAX	0.5131	0.3288	0.0006	0.0192	0.3282
SLAS32	MSE	0.0155	0.0054	0.0004	0.0004	0.0047
	MAE	0.1003	0.0566	0.0036	0.0036	0.0446
	MAX	0.3882	0.3655	0.2100	0.2100	0.2733

Table 3: Imputation error for different methods applied to TSLAVESD (total slaves on board at departure from last slaving port), TSLAVESP (total slaves purchased), SLAARRIV (total slaves arrived at first port of disembarkation), and SLAS32 (total slaves disembarked at first place). The mean values for these attributes were 0.2199, 0.2990, 0.1622, and 0.1660, respectively.

(4) Among the imputation methods considered (Mean Imputation, Bayesian Ridge, Decision Tree, Extra Trees, and K-Nearest Neighbor), Extra Trees achieve the best results in most of the settings.

Our work opens many opportunities for future research. This is a list of ongoing research directions:

(1) We will evaluate the missingness mechanisms in the slave voyages dataset using statistical methods;
(2) We will incorporate new results to our evaluation using other imputation methods, especially matrix completion, expectation-maximization, and autoencoder approaches;
(3) We will investigate how to generalize missing data imputation methods to handle heterogeneous data, including numeric, nominal, and date types;
(4) We will perform different data analytics—e.g., clustering and anomaly detection—using the imputed version of the slave voyages dataset;
(5) We will collaborate with a Historian to validate the missing data imputation results based on domain knowledge;
(6) We will apply our methodology to other digital history databases and assess whether the results found for slave voyages generalize to other databases.

The long-term goal of this research is to incorporate the missing data imputation into the slave voyages website in order to improve the statistics and visualizations provided. As a consequence, we expect to benefit Historians focused on the transatlantic slave trade.

REFERENCES

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Citeseer, 487–499.

[2] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. 2021. Self-driving cars: A survey. *Expert Systems with Applications* 165 (2021), 113816.

[3] Samuel F Buck. 1960. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)* 22, 2 (1960), 302–306.

[4] Florent Burba, Frédéric Ferraty, and Philippe Vieu. 2008. k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics* 21 (2008), 453–469.

[5] Emmanuel J Candes and Yaniv Plan. 2010. Matrix completion with noise. *Proc. IEEE* 98, 6 (2010), 925–936.

[6] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. 2018. The rise of deep learning in drug discovery. *Drug discovery today* 23, 6 (2018), 1241–1250.

[7] Dan Cohen and Roy Rosenzweig. 2006. Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. (2006).

[8] Harald Cramér. 1999. *Mathematical Methods of Statistics*. Vol. 43. Princeton University Press.

[9] Daniel B Domingues da Silva and Philip Misevich. 2018. Atlantic Slavery and the Slave Trade: History and Historiography. In *Oxford Research Encyclopedia of African History*.

[10] A. Rogier T. Donders, Geert J.M.G. van der Heijden, Theo Stijnen, and Karel G.M. Moons. 2006. A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 59 (2006), 1087–1091.

[11] David Eltis. 2000. *The rise of African slavery in the Americas*. Cambridge University Press.

[12] David Eltis. 2007. A brief overview of the Trans-Atlantic Slave Trade. *Voyages: The trans-atlantic slave trade database* (2007), 1700–1810.

[13] David Eltis, Stephen D Behrendt, David Richardson, and Herbert S Klein. 1999. *The trans-Atlantic slave trade: a database on CD-ROM*. Cambridge University Press Cambridge.

[14] David Eltis and Paul F Lachance. 2010. Estimates of the size and direction of transatlantic slave trade. *Voyages: The Trans-Atlantic Slave Trade Database* (2010).

[15] Craig K Enders. 2010. *Applied missing data analysis*. Guilford press.

[16] David A Freedman. 2009. *Statistical models: theory and practice*. cambridge university press.

[17] P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Mach Learn* 63 (2006), 3–42.

[18] Yuval Noah Harari. 2014. *Sapiens: A brief history of humankind*. Random House.

[19] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

[20] Herbert S Klein. 1969. The Trade in African Slaves to Rio de Janeiro, 1795–1811: Estimates of Mortality and Patterns of Voyages1. *The Journal of African History* 10, 4 (1969), 533–549.

[21] Herbert S Klein. 2010. *The Atlantic slave trade*. Cambridge University Press.

[22] Gary G Koch. 2004. Intraclass correlation coefficient. *Encyclopedia of statistical sciences* (2004).

[23] Guillaume Lample and François Charton. 2019. Deep Learning For Symbolic Mathematics. In *International Conference on Learning Representations*.

[24] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 53, 2 (2020), 1487–1509.
[25] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.
[26] Wei-Yin Loh. 2011. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery* 1 (2011).
[27] Timothy R Mahoney. 2001. The Great Sheedy Murder Trial and the Booster Ethos of the Gilded Age in Lincoln. *Nebraska History* 82 (2001), 163–79.
[28] Geert Molenberghs and Michael Kenward. 2007. *Missing data in clinical studies*. John Wiley & Sons.
[29] Andrea Nanetti and Siew Ann Cheong. 2018. Computational history: from big data to big simulations. In *Big Data in Computational Social Science and Humanities*. Springer, 337–363.
[30] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.
[31] Ricardo Cardoso Pereira, Miriam Seoane Santos, Pedro Pereira Rodrigues, and Pedro Henriques Abreu. 2020. Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *Journal of Artificial Intelligence Research* 69 (2020), 1255–1285.
[32] Johannes Preiser-Kapeller. 2015. Calculating the Middle Ages? The Project" Complexities and Networks in the Medieval Mediterranean and Near East"(COMMED). *arXiv preprint arXiv:1606.03433* (2015).
[33] James A Rawley and Stephen D Behrendt. 2005. *The transatlantic slave trade: a history*. U of Nebraska Press.
[34] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.
[35] Joseph L Schafer. 1997. *Analysis of incomplete multivariate data.* CRC press.
[36] Douglas Seefeldt and William G Thomas III. 2009. What is digital history? A look at some exemplar projects. *Perspectives on History* 47, 5 (2009).
[37] William G Thomas. 2011. *The Iron Way: Railroads, the Civil War, and the Making of Modern America.* Yale University Press.
[38] Andrew J Torget. 2015. *Seeds of Empire: Cotton, Slavery, and the Transformation of the Texas Borderlands, 1800-1850.* UNC Press Books.
[39] William J Turkel, Shezan Muhammedi, and Mary Beth Start. 2014. Grounding digital history in the history of computing. *IEEE Annals of the History of Computing* 36, 2 (2014), 72–75.
[40] Richard White. 2010. What is spatial history. *Spatial History Lab* 1 (2010).