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