What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model

Melvin Wevers[1,*,†], Folgert Karsdorp[2,†] and Jelle van Lottum[3]

[1]University of Amsterdam, Amsterdam, the Netherlands
[2]KNAW Meertens Institute, Amsterdam, the Netherlands
[3]KNAW Huygens Institute, Amsterdam, the Netherlands

Abstract

Historians base their inquiries on the sources that are available to them. However, not all sources that are relevant to the historian's inquiry may have survived the test of time. Consequently, the resulting data can be biased in unknown ways, possibly skewing analyses. This paper deals with the Dutch East India Company its digitized ledgers of contracts. We apply an unseen species model, a method from ecology, to estimate the *actual* number of unique seafarers contracted. We find that the lower bound of actual seafarers is much higher than what the remaining contracts indicate: at least, thirty-six percent of the seafarers is unknown. Moreover, we find that even in periods when few records survived, we can still credibly estimate a lower bound on the unique number of seafarers.

Keywords

Computational History, Survivor Bias, Unseen Species Model, Sampling Without Replacement

1. Introduction: Historical Records and Survivor Bias

Historians can only rely on the archival records that have survived the test of time. That a substantial share of historical records has not survived may be due to natural causes, such as fires, decisions on the level of archival policy making, but also content production biases [17]. For instance, whether or not particular sources were retained can depend on socio-economical factors [20], as data representative of lower classes were long deemed less relevant by archivists [21].

As historians are working with data that is hampered by many possible types of bias, they need to critically evaluate to what extent the remaining data is representative of the collection or historical period from which its stems [14]. Put differently, historians need to reflect on how

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium
*Corresponding author.
†Both MW and FK contributed equally. MW and FK have the right to list their name first in their CV.
melvin.wevers@uva.nl (M. Wevers); folgert@karsdorp.io (F. Karsdorp); jelle.van.lottum@huygens.knaw.nl (J. v. Lottum)
https://www.melvinwevers.nl (M. Wevers); https://www.karsdorp.io (F. Karsdorp); https://www.huygens.knaw.nl/medewerkers/jelle-van-lottum/ (J. v. Lottum)
0000-0001-8177-4582 (M. Wevers); 0000-0002-5958-0551 (F. Karsdorp); 0000-0003-0534-4745 (J. v. Lottum)

transmitted archival records relate to the actual historical past. Especially now when historical records are rapidly and continuously being digitized, we have to be even more cautious of potential biases in archives. The speed with which we can analyze data combined with the increased distance between the researcher and the source material makes it easier to overlook how bias impacts the historical inferences we make [1]. Evidently, we can only digitize the records that have survived. Even if all surviving records would be digitized, biases will remain to exist.

Yet, at the same time, the fact that data is digitized also facilitates the use of statistical and computational methods which help chart and possibly deal with the blind spots of the data. Studies have already applied statistical methods to expose how bias can lead to overestimating the impact of historical events [18], misrepresentations of the standard of living [20], or the underestimation of wartime causalities [13].

In this paper, we focus on estimating the number of unique seafarers employed by the Dutch East India Company (*Verenigde Oost-Indisch Compagnie, VOC*). The company kept detailed personnel administration records (pay ledgers), which have been digitized in the VOCOP dataset.[1] While the dataset, which has been used extensively by historians and social scientists in the context of financial [22] and maritime history [23], contains a sizeable amount of data, we also know that records have been lost and thus have not been digitized. This can impact the qualitative and quantitative historical study of the VOC. More specifically, we do not know how the loss of records is distributed temporally, and whether the surviving records can give a credible estimate of the number of unique individuals hired by the VOC. Having better information on the representativeness of this data allows us to better study aspects such as career mobility and the financial position of VOC employees.

For the estimation, we draw on unseen species models [7, 6], which aim to estimate the number of unique species living in a given environment. Beyond ecology, these models have been successfully applied to a wide array of cases, ranging from estimating the number of classes of stone tools in archaeology [12], the number of bugs in software code [5], the number of stars in the Pleiades [3], the size of an author's vocabulary [11], and, more recently, to estimate the number of lost medieval literary works [16]. Here, we apply a modification of the model for samples without replacement [9], which has not yet been applied in the context of humanities research.[2]

2. Data and preprocessing

This paper uses two different data sources: *VOC: CAREERS (VOCCAR)* and *Dutch Asiatic Shipping in the 17th and 18th centuries (DAS)*.

VOCCAR is an enriched version of the VOCOP dataset, which contains digitized pay ledgers of the VOC. [19] The dataset contains 774,200 contracts between 1633 and 1795, with the majority of records stemming from the 18$^{\text{th}}$ century. The contracts specify, among other

[1]These records have been digitized by volunteers working for the National Archives of the Netherlands, and can be accessed here: https://www.nationaalarchief.nl/onderzoeken/index/nt00444?activeTab=nt

[2]The data and code used in this paper have been registered under: https://doi.org/10.5281/zenodo.7268250

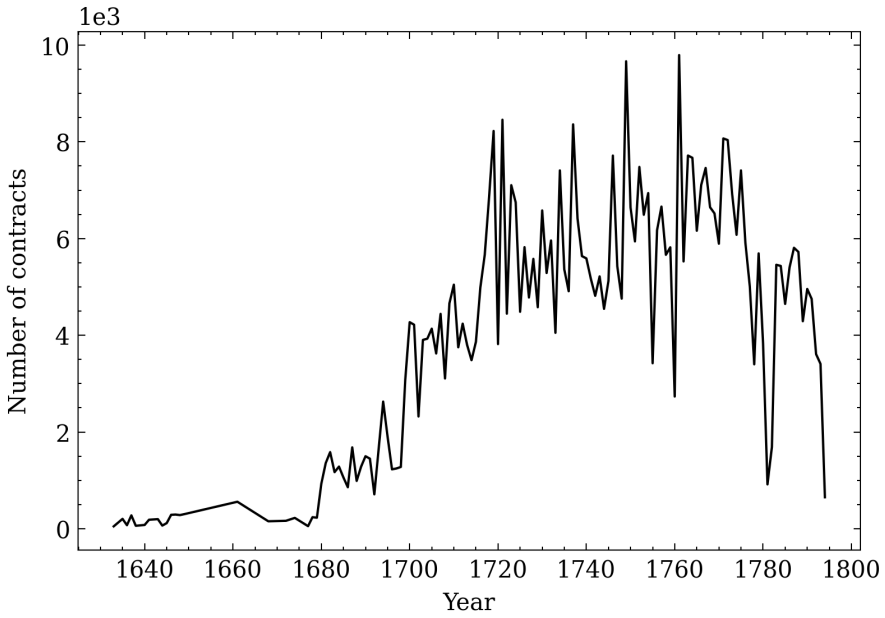

Figure 1: Overview of clustered contracts in the VOCCAR dataset.

things, the name, rank, place of birth of the contractee, the date of sailing, and the ship on which they sailed. The original ledgers from which these records have been digitized could contain multiple contracts belonging to the same person. In VOCCAR, the records have been clustered around unique individuals, which allows us to count how often they appeared in the records.

We only focus on the records that have been clustered, resulting in 546,973 records (N), of which 460,274 are unique seafarers (V). Figure 1 provides an overview of the clustered records in the VOCCAR dataset. We see a sharp increase in records during the 17th century as well as noticeable gaps in the data for the 18th century.

DAS provides an overview of the number of ships that sailed out from the Dutch republic.[3] This data is almost complete, with only a few voyages missing from the data.[4]

We learn from DAS that the VOC sailed out 4,352 times between 1633 and 1795. For about 91 percent of these voyages, DAS provides information on the number of people that boarded the vessel. However, because this data contains some noticeable outliers, we decided to calculate the mean voyagers per ship for periods of twenty-five years rather than imputing merely the missing values. Next, for each period, we multiplied this mean by the number of voyages, including those that lack information on the number of voyagers. This provides us with an estimation of the total number of records: 952,147.[5]

[3] http://resources.huygens.knaw.nl/das/EnglishIntro

[4] The data quality is discussed here: http://resources.huygens.knaw.nl/retroboeken/das/

[5] We use DAS to calculate the total number of records rather than VOCCAR because DAS is much more complete.

3. Method: Estimating the Number of Unique Individuals under Sampling without Replacement

To estimate the number of unique seafarers of the VOC, we employ an unseen species model. The model was originally developed in ecology, where researchers are often confronted with incomplete data as a result of undersampling. Due to such data incompleteness, it is possible that important statistics such as biodiversity are estimated to be much lower than they actually are. To combat such estimation biases, it is an important research question in ecology how the resulting difference between the number of observed and the true number of unique species can be reliably estimated. A potential solution is given by the Chao1 estimator, developed by Anne Chao [7, 6].

The Chao1 estimator is a non-parametric unseen species model that estimates a universally valid lower bound on the number of unseen entities (e.g., seafarers; call that f_0), based on entities that have been observed once or twice (call those f_1 and f_2). Theoretically, we can calculate the number of unseen entities by taking the product of the average relative frequency of unseen entities (α_0) and the number of unseen entities (f_0), divided by α_0. However, α_0 cannot be calculated directly. What we do know is that the average relative frequency of unseen entities (α_0) is probably lower than that of entities occurring once (α_1), i.e., $\alpha_0 \leq \alpha_1$. It then follows that $\frac{\alpha_0 f_0}{\alpha_0}$ must be *at least* equal to or greater than $\frac{\alpha_0 f_0}{\alpha_1}$ (hence, f_0 is a lower bound). That latter expression is computable and can be rewritten into the Chao1 estimator [cf. 8]:

$$\hat{V} = V + \frac{\alpha_0 f_0}{\alpha_1} = V + \frac{\frac{f_1}{n}}{\frac{2f_2}{(n-1)f_1}} \equiv V + \frac{(n-1)}{n}\frac{f_1^2}{2f_2}, \tag{1}$$

where V refers to the observed number of unique entities, n to the sum of their occurrences, and $\hat{V}$ to the bias-corrected lower bound. It is important to note that when $\alpha_0 \approx \alpha_1$, that is, when unseen entities have approximately the same average relative frequency as entities occurring once, Chao1 becomes an unbiased point estimator [8].

The Chao1 estimator was developed assuming that samples are formed with replacement. This means that during each sampling moment, the same individuals can be observed multiple times. It also means that observations are independent of each other, and that the observation of one individual does not affect the observation of the next. In other words, the covariation between successive observations is zero. Thus, sampling with replacement essentially assumes an infinite population. For example, a snippet of text can be seen as a sample of an author's infinite stream of words. And if we apply the Chao1 estimator to this snippet, a lower bound on the vocabulary of the author is also exactly what is estimated [11].

We can also think of the snippet as a sample of the finite space of the snippet's encompassing book. Treating the snippet as such would imply that the sample was created *without* replacement. In such samples, observations are not independent, nor is the covariation between successive observations zero. Crucially, however, because of its assumption that samples are formed with replacement and are thus drawn from an infinite population, the Chao1 estimator does *not* estimate the number of unique words in the book encompassing the snippet. Thus, even though we know a given sample to come from a finite population, Chao1 always treats it

as coming from an infinite one.

The VOC records of this study should be conceptualized as samples created without replacement. There has been a finite population of seafarers with the VOC of which the records show a sample without replacement.[6] The problem, however, is that when we apply the Chao1 estimator to this sample, we do not obtain an estimate of the number of unique individuals in the total, finite population, but rather that of a *potential* population of seafarers, which is not what we are after. To estimate the number of unique individuals in the finite population of employees of the VOC, we employ a modified Chao1 estimator developed by Chao and Lin for samples without replacement [9]. This modified estimator assumes we know the size N of the total population, and thus know the ratio q of the observed sample size to the total population:

$$\hat{V}_{\text{wor}} = V + \frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{q}{1-q}f_1} f_1 \equiv V + \frac{f_1^2}{2wf_2 + rf_1}, \tag{2}$$

where $w = n/(n-1)$ and $r = q/(1-q)$. Note that when q approaches zero, Eq. 2 reduces to the standard Chao1 estimator in Eq. 1. We refer to the modified estimator as $\text{Chao1}_{\text{wor}}$. Confidence intervals for $\text{Chao1}_{\text{wor}}$ can be computed based on the variance estimator [9]:

$$\text{var}(\hat{V}_{\text{wor}}) = \hat{f}_0 + \frac{(2wf_2\hat{f}_0^{\,2} + f_1^2\hat{f}_0)^2}{f_1^5} + 4w^2 f_2 \left(\frac{\hat{f}_0}{f_1}\right)^4 \tag{3}$$

Based on the total number of records we derived from DAS ($\hat{N} = 952,147$, see above), we calculate the sample fraction q by dividing N by $\hat{N}$. For the complete dataset $q \approx 0.57$. Note that $\hat{N} \neq \hat{V}$, since individuals may have been shipped out multiple times.

4. Results: There are many more unique seafarers than the records show

At least thirty-six percent of the seafarers is unknown Based on the observed abundances, i.e. how many times each unique individual was "*sighted*" in the data, and the sample fraction q estimated from $\hat{N}$, we calculate with $\text{Chao1}_{\text{wor}}$ the lower bound on the number of *actual* unique individuals in the VOC population ($\hat{V}$) to be 716,818 (95%CI: 715,439 to 718,203). This number suggests that we should account for a survival rate of $V/\hat{V} \approx 64\%$, or conversely, that of the original VOC population, at least 36% of the individuals is unknown.

The loss of records impacted the number of unique individuals in the records To get a better understanding of the coverage of the data across time, we applied the same approach to successive periods of twenty-five years. For each period, we calculated the mean number of voyagers on a journey and multiplied this with the total number of journeys in that period, thus estimating the *actual* number of seafarers (see Table 1). Figure 2 displays the observed number of unique seafarers V against the estimated number $\hat{V}$ over time. The gray overlay

[6] This finite population can be constrained by many different things, about which we can now only speculate: the total number of ships, skills required to be enlisted, etc.

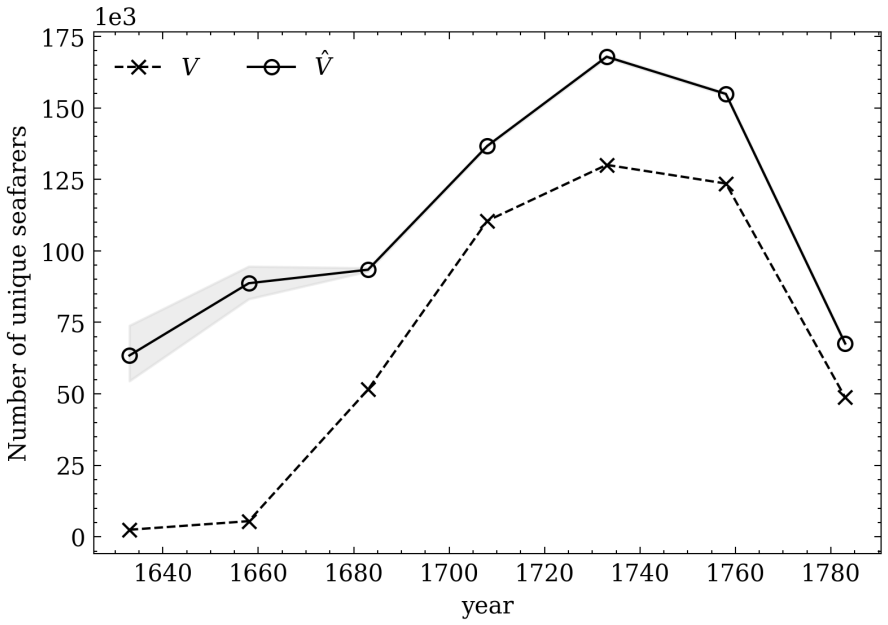

Figure 2: Plot showing the observed (V) and estimated ($\hat{V}$) number of unique seafarers computed with $\text{Chao1}_{\text{wor}}$ for time spans of 25 years.

represents the 95% confidence intervals of the estimates. The plot shows that, especially in the 17$^{\text{th}}$ century, the lack of data has led to a severe underestimation of the number of unique individuals (ranging from 45% in 1683–1708 to 96% in 1633–1658). With an average of $\approx$ 23%, the gap between the observed and the actual number of unique individuals is smaller in the 18$^{\text{th}}$ century but still considerable.

Table 1
Overview of data for twenty-five-year periods. The last period only spans twelve years.

	period	voyages	q	$\hat{N}$	V	$\hat{V}$	CI_{lower}	CI_{upper}	loss	Chao1
0	1633 - 1658	449	0.03	88,840	2,457	63,417	54,494	73,869	96.13%	215,522
1	1658 - 1683	565	0.05	106,756	5,462	88,671	83,170	94,560	93.84%	497,150
2	1683 - 1708	596	0.50	111,352	51,501	93,409	92,808	94,018	44.87%	409,425
3	1708 - 1733	856	0.77	169,588	110,517	136,699	136,337	137,065	19.15%	540,560
4	1733 - 1758	821	0.74	202,405	130,069	167,892	167,446	168,343	22.53%	725,923
5	1758 - 1783	682	0.76	193,968	123,553	154,891	154,490	155,296	20.23%	574,654
6	1783 - 1795	383	0.68	79,238	48,738	67,597	67,267	67,932	27.90%	293,352

Assuming sampling with replacement yields impossible estimates We have established empirically that assuming sampling with replacement, Chao1 gives an unrealistic lower bound of more than 2.3 million seafarers, which by far exceeds the upper limit of $\hat{N}$, which is just below 1 million. The same is true for the shorter periods of 25 years. Here too, Chao1 systematically produces impossible estimates. As we explained above, the Chao1 estimate might

be considered the *potential* rather than the *actual* number of seafarers that could have worked for the VOC. By contrast, the estimate of the $\text{Chao1}_{\text{wor}}$ estimator is compatible with the upper limit of $\hat{N}$ and thus supports our approach of conceptualizing these sightings as samples without replacement. More generally, these results emphasize the need to understand the sampling process underlying the data, and to exercise caution when applying the estimators. When data are sampled without replacement, but the estimator assumes otherwise, Chao1 is not guaranteed to produce a lower bound, which puts any reliable interpretation of the results into question.

5. Conclusion

This short paper is the first to quantify the scale and extent of the assumed data loss and lack of representativeness of the archives of the the Dutch East India Company (the VOC). We applied the $\text{Chao1}_{\text{wor}}$ estimator to a database of employees of the VOC, and found that we can make credible predictions on the lower bound of the number of unique seafarers that have been employed by the company. Moreover, even when relatively small fractions of the records have survived, the estimates appear to be robust. For the entire archival period, we estimate that at least forty percent of unique seafarers are not recorded in the archives. Put differently, the *actual* number of unique seafarers was much higher than the surviving records indicate. Moreover, the estimated increase in the number of unique seafarers in the 17^{th} century is not as steep in actuality as the empirical, observed records suggest. Finally from the 18^{th} century onward, the difference between the observed and the actual number of unique seafarers is smaller but still considerable. More generally, our results show how unseen species models from ecology can be used to obtain a clearer perspective on the parts of historical archives that are lost.

This paper adds to a series of recent studies exploring the applicability of unseen species models to cultural data [12, 15, 10, 16]. While these prior studies primarily investigate samples from infinite populations, the present paper explored the applicability of Chao1 without replacement [9] in the context of cultural data sampled from finite populations. The case study of the VOC underscored the importance of a proper conceptualization and understanding of the sampling process underlying the data. Without such understanding, or when the assumptions about the sampling process of the model do not correspond to the actual sampling process underlying the observed data, the estimates may no longer be reliable – or, more precisely – they do not match what we hope to estimate. For example, when data are sampled without replacement, Chao1 is no longer guaranteed to estimate a lower bound. The records of the VOC should be conceptualized as a sample without replacement, for which the modified $\text{Chao1}_{\text{wor}}$ estimator can, by contrast, adequately estimate a credible lower bound. An important remaining issue with the application of unseen species models to cultural data (whether they are sampled with or without replacement) is that the data are assumed to be homogeneous and thus that all entities (e.g., seafarers) are equally likely to be observed. The consequence of this simplifying assumption is that the unseen species estimators reduce from a point-estimate to a lower bound of the actual population size. In a series of studies, Böhning and colleagues present generalized unseen species models that show how adding information about the origins of heterogeneity

of the data can reduce some of the bias of the estimates [4, 2]. In future work, we aim to refine our estimates by incorporating such covariate information in these generalized unseen species models. For example, the current analysis offers no information on whether factors such as rank or origin impacted the loss of certain records. It is quite conceivable, however, that the scrupulousness of the log files may vary between records of high-ranking officials from the Dutch republic and those of seafarers from further away. One may also wonder whether the effect of rank or origin fluctuates over time, possibly relating to periods of social unrest.

References

[1] R. Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. 1st edition. Medford, MA: Polity, June 17, 2019. 172 pp.

[2] D. Bohning, A. Vidal-Diez, R. Lerdsuwansri, C. Viwatwongkasem, and M. Arnold. "A Generalization of Chao's Estimator for Covariate Information". In: *Biometrics* 69 (2013), pp. 1033–1042. DOI: 10.1111/biom.12082.

[3] D. Böhning. "Chao's Lower Bound Estimator and the Size of the Pleiades". In: *Environmental and Ecological Statistics* 27.1 (2020), pp. 171–173. DOI: 10.1007/s10651-020-00440-w.

[4] D. Böhning and P. G. M. van der Heijden. "A Covariate Adjustment for Zero-Truncated Approaches to Estimating the Size of Hidden and Elusive Populations". In: *The Annals of Applied Statistics* 3.2 (2009). DOI: 10.1214/08-aoas214.

[5] L. Briand, K. El Emam, B. Freimut, and O. Laitenberger. "A Comprehensive Evaluation of Capture-Recapture Models for Estimating Software Defect Content". In: *IEEE Transactions on Software Engineering* 26.6 (2000), pp. 518–540. DOI: 10.1109/32.852741.

[6] A. Chao. "Estimating Population Size for Sparse Data in Capture-Recapture Experiments". In: *Biometrics* 45.2 (1989), pp. 427–438. DOI: 10.2307/2531487.

[7] A. Chao. "Nonparametric Estimation of the Number of Classes in a Population". In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 265–270.

[8] A. Chao, C.-H. Chiu, R. K. Colwell, L. F. S. Magnago, R. L. Chazdon, and N. J. Gotelli. "Deciphering the Enigma of Undetected Species, Phylogenetic, and Functional Diversity Based on Good-Turing Theory". In: *Ecology* 98.11 (2017), pp. 2914–2929. DOI: 10.1002/ecy.2000.

[9] A. Chao and C.-W. Lin. "Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement". In: *Biometrics* 68.3 (2012), pp. 912–921. DOI: 10.1111/j.1541-0420.2011.01739.x.

[10] R. K. Colwell and A. Chao. "Measuring and comparing class diversity in archaeological assemblages: A brief guide to the history and state-of-the-art in diversity statistics." In: *Defining and Measuring Diversity in Archaeology. Another Step Toward an Evolutionary Synthesis of Culture*. Ed. by M. I. Eren and B. Buchanan. New York – Oxford: Berghahn, 2020, pp. 263–294.

[11] B. Efron and R. Thisted. “Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?” In: *Biometrika* 63.3 (1976), p. 435. DOI: 10.2307/2335721.

[12] M. I. Eren, A. Chao, W.-H. Hwang, and R. K. Colwell. “Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed: A Nonparametric Solution to an Archaeological Problem”. In: *PLoS ONE* 7.5 (2012). Ed. by A. Mesoudi, e34179. DOI: 10.1371/journal.pone.0034179.

[13] C. S. Gillespie. “Estimating the Number of Casualties in the American Indian War: A Bayesian Analysis Using the Power Law Distribution”. In: *The Annals of Applied Statistics* 11.4 (2017), pp. 2357–2374.

[14] K. Inwood and H. Maxwell-Stewart. “Selection Bias and Social Science History”. In: *Social Science History* 44.3 (2020), pp. 411–416.

[15] M. Kestemont and F. Karsdorp. “Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity”. In: *Workshop on Computational Humanities Research*. Amsterdam: Ceur-ws, 2020, pp. 44–55.

[16] M. Kestemont, F. Karsdorp, E. de Bruijn, M. Driscoll, K. A. Kapitan, P. Ó Macháin, D. Sawyer, R. Sleiderink, and A. Chao. “Forgotten Books: The Application of Unseen Species Models to the Survival of Culture”. In: *Science* 375.6582 (2022), pp. 765–769. DOI: 10.1126/science.abl7655.

[17] A. Lee. “The Library of Babel: How (and How Not) to Use Archival Sources in Political Science”. In: *Journal of Historical Political Economy* 2.3 (2022), pp. 1–39.

[18] L. Mordechai, M. Eisenberg, T. P. Newfield, A. Izdebski, J. E. Kay, and H. Poinar. “The Justinianic Plague: An Inconsequential Pandemic?” In: *Proceedings of the National Academy of Sciences* 116.51 (2019), pp. 25546–25554.

[19] L. Petram, M. Koolen, M. Wevers, R. van Koert, and J. van Lottum. “Data on the Maritime Workforce of the Dutch East India Company in the 18th Century”. In: (Forthcoming).

[20] B. Quanjer and J. Kok. “Drafting the Dutch: Selection Biases in Dutch Conscript Records in the Second Half of the Nineteenth Century”. In: *Social Science History* 44.3 (2020), pp. 501–524.

[21] M.-R. Trouillot and H. V. Carby. *Silencing the Past: Power and the Production of History*. Boston, Massachusetts: Beacon Press, 2015. 190 pp.

[22] C. Van Bochove and T. Van Velzen. “Loans to Salaried Employees: The Case of the Dutch East India Company, 1602–1794”. In: *European Review of Economic History* 18.1 (2014), pp. 19–38.

[23] J. van Lottum and L. Petram. “In Search of Strayed Englishmen. English Seamen Employed in the Dutch East India Company in the Late Seventeenth and Eighteenth Centuries”. In: *Anglo-Dutch Connections in the Early Modern World*. Ed. by E. van Raamsdonk, S. Levelt, and M. Rose. London: Taylor & Francis, Forthcoming. Forthcoming.