



Charting Lives and Careers: Enriched Data About the Dutch East India Company's Eighteenth-Century European Workforce

DATA PAPER

LODEWIJK PETRAM

MARIJN KOOLEN

MELVIN WEVERS

JELLE VAN LOTTUM

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This dataset builds upon the ‘VOC-opvarenden’ database, which documents the Dutch East India Company’s (VOC) European personnel, primarily from the eighteenth century. We refine the 774,200 muster records containing each crew member’s name, place of origin, rank, and the specifics of the ship voyage they enlisted for, by disambiguating to identify 460,452 individuals, standardizing place names, categorizing ranks, and integrating wage information. Hosted on Zenodo, this dataset offers comprehensive insights into the careers of VOC crew and their origins from different parts of Europe. Unique for its exceptional size for early modern data, long time span, and European coverage, this extensive collection offers unprecedented opportunities for in-depth analysis and research. Its potential extends to sociological studies, network analyses, and historical research, particularly when combined with economic and social data from various European regions.

CORRESPONDING

AUTHOR:

Lodewijk Petram

Huygens Institute, Amsterdam,
The Netherlands

lodewijk.petram@huygens.knaw.nl

KEYWORDS:

digital humanities; digital history; social and economic history; Dutch East India Company (VOC); employment records

TO CITE THIS ARTICLE:

Petram, L., Koolen, M., Wevers, M., & van Lottum, J. (2024). Charting Lives and Careers: Enriched Data About the Dutch East India Company’s Eighteenth-Century European Workforce. *Journal of Open Humanities Data*, 10: 44, pp. 1–17. DOI: <https://doi.org/10.5334/johd.210>

1 CONTEXT AND MOTIVATION

From its founding in 1602 until its demise at the end of the eighteenth century, the Dutch East India Company (Verenigde Oostindische Compagnie, VOC) engaged in long-distance trade between Asia and Europe. Additionally, within Asia, it competed with local shippers and merchants, and attempted to assert its influence over a vast region surrounding the Indian Ocean, centered around modern-day Indonesia (Gaastra, 1991). Today, the company is renowned for its modern organizational structure and notorious for its brutal conduct, including active engagement in the slave trade.

All activities were meticulously recorded. Among the extensive surviving archival sources are over 3,020 pay ledgers, primarily from the eighteenth century, recording all personnel who worked on a ship voyage from the Netherlands to Asia or traveled to Asia for onshore jobs, such as soldiers. This is a true archival treasure, which sheds lights on the personnel recruited by one of the largest employers of its time: in the eighteenth century, the VOC hired around 7,000 crew members per year on average, and in peak years as many as 12,000. The recruits came from a wide area. By the mid-eighteenth century, Dutch nationals made up less than half of the crew on average; the rest came primarily from German countries, the Southern Netherlands, and Scandinavia, with additional personnel from England, France, Switzerland, and other parts of Europe (van Lottum & Petram, 2023).

For each crew member, the pay ledgers contain a record spanning two pages (see Figure 1 for an example). These records include the employee's name, rank at the time of hiring, and wage payments. VOC clerks also recorded the new personnel's place of origin, which is likely, though not necessarily, their birthplace.

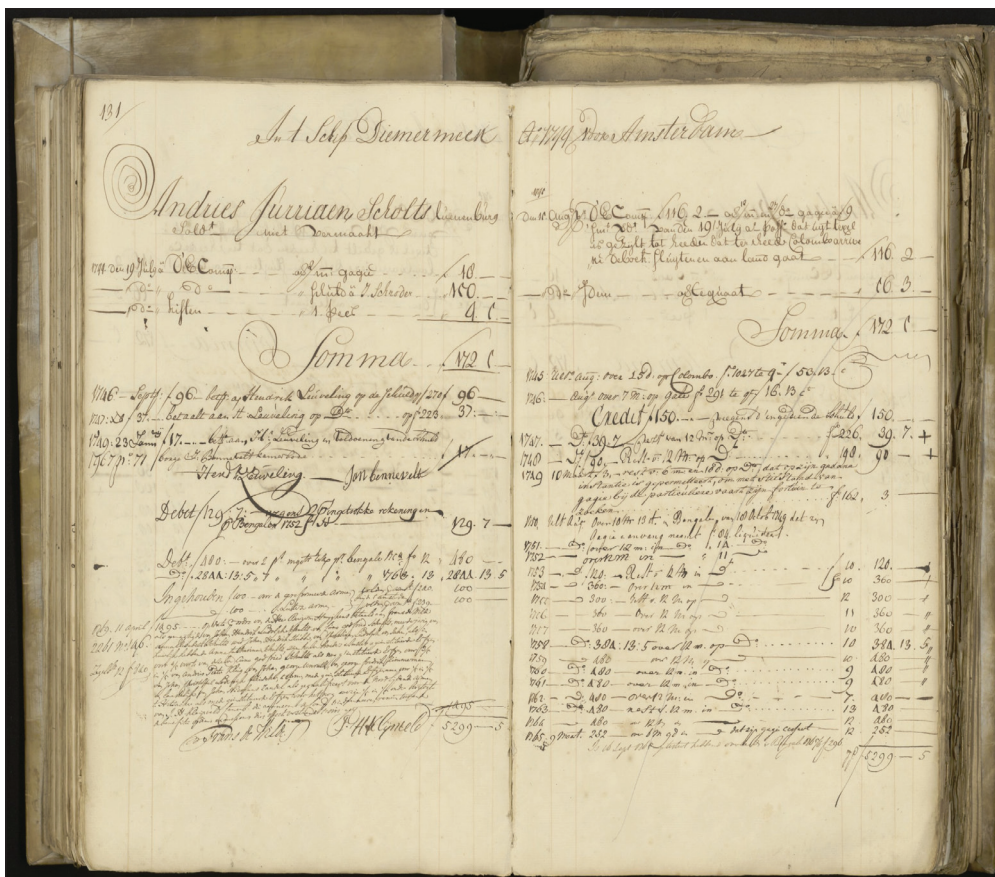


Figure 1 Original pay ledger record of Andries Jurriaen Scholts from 'Luenenburg' (Lüneburg in Germany), who signed up as a soldier on the ship 'Diemermeer' in 1744. Source: NL-HaNA, 1.04.02, inv. no. 6164, fo. 131. <http://hdl.handle.net/10648/cef964ad-a553-5c97-39e3-9aa43b1c81c6> and <http://hdl.handle.net/10648/354266c6-2546-7aea-6d27-7dad3cf2c7b5>.

Between 2000 and 2012, a selection of fields from all records in the surviving VOC pay ledgers was manually entered into the 'VOC-opvarenden' database (van Velzen, 2006).¹ Although fragmentary for the seventeenth century, this database, capturing 774,200 muster records of European crew, offers a near-complete overview of the company's personnel on voyages between Europe and Asia in the eighteenth century (Wevers, Karsdorp, & van Lottum, 2022). This historical resource is unparalleled in its size and scope. The closest equivalents to 'VOC-

¹ <https://www.nationaalarchief.nl/onderzoeken/index/nt00444>.

opvarenden' are the databases 'Staten van oorlog',² which contains over 400,000 entries of military personnel employed by the armies of the Dutch States General, and 'Mémoire des hommes', which is still under development and currently includes nearly 150,000 records of crew and passengers on ships of the French Compagnie des Indes from the eighteenth century.³ The records in these databases are searchable, but they cannot be easily downloaded or referenced. The database containing information on more than 10,000 crew members from the eighteenth-century Middelburgsche Commercie Compagnie (MCC), a slave trading company based in the Dutch province of Zeeland, and records of crews of Zeeland naval warships is rich but comparatively much smaller.⁴ Records of the Dutch West India Company have been lost, but a team of digital historians is currently mining notarial deeds to partially reconstruct crew lists.⁵ While many eighteenth-century records in British archives have been digitized and, in some cases, made searchable, no comprehensive database exists. There are also Danish and Swedish databases, but they are very small in comparison to 'VOC-opvarenden', with each containing only a couple of thousand records.⁶

Researchers have frequently used the 'VOC-opvarenden' database (Bruijn & Gaastra, 2012; Hoyer, 2020; Ketelaars, 2014; Odegard, 2022; Pustai & Teszelszky, 2016; Sgourev & van Lent, 2017; van Bochove, 2018; van Bochove & van Velzen, 2014; van Lottum & Petram, 2023; van Rossum, 2014; Wevers et al., 2022; Wezel & Ruef, 2017, 2020, 2024; Zijlstra, 2021), yet its form lends itself poorly to large-scale, structured analyses. The first challenge with the database lies in its focus on employment records rather than on individuals. The original VOC pay ledgers, designed solely for calculating salaries, terminated records at the end of a crew member's service. Subsequent re-employments were recorded separately, leading to fragmented data for individuals who served multiple times. Consequently, research aimed at tracing individual histories requires an intricate process of clustering related observations. Furthermore, the dataset's varied and non-standardized origin notations severely limit analyses of the origins of the VOC workforce.

Shortly after completion of the dataset, de Boer, van Rossum, Leinenga, and Hoekstra (2014) made attempts to enhance its usability by standardizing place names and integrating related datasets. While their endeavor was commendable, it had limitations due to their exclusive reliance on automated processes, resulting in quality shortcomings. In contrast, Wezel and Ruef (2017, 2020, 2024) undertook a more rigorous process. For their analysis of labor control mechanisms aboard ships of the VOC (Wezel & Ruef, 2017), they identified 561,454 seafarers with one or multiple employments records. Further, they write that they were able to 'unambiguously' assign the place of origin in 82% of the records to a region. However, their methods and data remain undisclosed, despite its potential value to other researchers.

We followed a similar approach to Wezel and Ruef in enhancing the 'VOC-opvarenden' dataset: by disambiguating muster records (linking individuals to their respective records) and standardizing places of origin. Additionally, we organized the ranks in the records and integrated wage data. Dataset metadata is provided in Section 2, our approach detailed in Section 3, and potential applications of the enriched dataset described in Section 4.

2 DATASET DESCRIPTION

REPOSITORY LOCATION

<https://doi.org/10.5281/zenodo.10599528>

OBJECT NAME

The Dutch East India Company's Eighteenth-Century Workforce: an Enriched Data Collection

2 <https://www.bhic.nl/onderzoeken/staten-van-oorlog>.

3 https://www.memoiredeshommes.sga.defense.gouv.fr/fr/arkotheque/client/mdh/compagnie_des_indes/equipages_et_passagers.php.

4 <https://www.zeeuwsarchief.nl/zeeuwen-gezocht/>.

5 <https://www.amsterdam.nl/stadsarchief/organisatie/projecten/wic-opvarenden/>.

6 Danish database: <https://www.ddd.dda.dk/dop/sogeside.asp>; Swedish database: <https://www2.ub.gu.se/samlingar/handskrift/ostindie/personer/>.

CSV for enriched files; CSV and Excel for source files; Jupyter notebooks for data transformation documentation, PDF for comprehensive data paper, Markdown for codebooks.

CREATION DATES

Start date: 2017-02-01; end date: 2024-02-02

DATASET CREATORS

- Lodewijk Petram (Huygens Institute, Amsterdam, The Netherlands): Conceptualization, Data curation, Validation.
- Marijn Koolen (DHLab, KNAW Humanities Cluster, Amsterdam, The Netherlands): Conceptualization, Methodology, Software, Validation, Visualization.
- Melvin Wevers (Department of History, University of Amsterdam, Amsterdam, The Netherlands): Conceptualization, Methodology, Software, Visualization.
- Rutger van Koert (Digital Infrastructure, KNAW Humanities Cluster, Amsterdam, The Netherlands): Conceptualization, Methodology, Software.
- Jelle van Lottum (Huygens Institute, Amsterdam, The Netherlands; Radboud Institute for Culture and History, Radboud University, Nijmegen, The Netherlands): Conceptualization, Funding acquisition.
- Maarten Bosker and Dinand Webbink (Erasmus School of Economics, Erasmus University, Rotterdam, The Netherlands): Methodology, Software.
- Michael Baars, Ger Dijkstra, Ger Ruigrok, Mariëlle Scherer, and Daniël Tuik (Huygens Institute, Amsterdam, The Netherlands): Data curation.

LANGUAGE

English

LICENSE

Creative Commons Attribution 4.0 International

REPOSITORY NAME

Zenodo

PUBLICATION DATE

2024-02-12

DESCRIPTION OF FILES AND VARIABLES

The enriched dataset comes in seven files. The tables included in the appendix detail the variables included in the files that respectively describe the sources (4), the persons and their employment contracts (5), their names (6), their places of origin (7 and 8), their ranks (9), the voyages on which they served (10), and the beneficiaries to whom they assigned their month letters (11).

3 METHOD

The original ‘VOC-opvarenden’ database includes three files detailing the sources, muster records of VOC employees, and the beneficiaries designated to receive a portion of an employee’s yearly salary in the Netherlands. The file containing the muster records is the main collection. It comprises 774,200 muster records from 1633–1794.

Each record represents an individual enlisting in the Dutch Republic on a specific date on a VOC ship outfitted by one of the company’s six Dutch branches. Consequently, Asian personnel of the VOC, who comprised a significant portion of the crew on voyages within Asia, are not

included, except in rare cases where Asians re-entered VOC service in the Netherlands after working on a voyage from Asia to Europe. Apart from the variables mentioned in Section 1, the ‘VOC-opvarenden’ records also state whether the employee signed a transport letter and month letter. Transport letters, with a standard value of 150 guilders for sailors and higher amounts for senior ranks, were transferable salary claims, often used to clear pre-employment debts. Month letters allowed employees to designate beneficiaries (often family members) to receive a portion of their salary, typically three months’ wages annually. The dataset does include details on the month letter beneficiaries, but not on the size and payee of the transport letters. The records also provide details on when and why an employee’s service ended. Common reasons include completion of service (upon returning to Europe), death, desertion, and shipwreck. If a crew member returned to Europe, the record includes information about the return journey. However, details of time spent in Asia and wage information, though present in the original pay ledgers, are not included in the ‘VOC-opvarenden’ dataset. The last piece of information that was taken from the original source documents reveals if a crew member changed ships during a stop at the Cape of Good Hope, a routine break for VOC ships en route between Europe and Asia since the mid-seventeenth century. If a crew member transferred from one ship to another during the outbound journey, a new entry was made in the receiving ship’s pay ledger. This entry included the name of the previous ship and was usually closed again after arrival in Asia. The original pay ledger record remained open until the crew member’s service ended.

To adapt the ‘VOC-opvarenden’ data for large-scale analysis, we performed three main enrichment steps, outlined below. Descriptive statistics of the resulting enriched dataset are provided in [Table 1](#).

ASPECT		NUMBER
Records	Total	774,200
	Disambiguated	547,132
	Ambiguous	227,068
Person names	Raw	554,810
	Normalized	493,653
Persons		460,452
Place names	Raw	152,315
	Normalized	143,101
Ranks		197
Outward voyages		3,268

Table 1 Descriptive statistics of the enriched dataset. Note that the number of outward voyages is higher than the number of pay ledgers in the original VOC-soldijboeken dataset (3,020). Despite the absence of pay ledgers for certain outbound voyages, details about some crew members and their outbound voyages were retrieved from other sources.

STEPS

Identifying individuals

There is a large literature on linking historical records (see Abramitzky, Boustan, Eriksson, Feigenbaum, and Pérez (2021); Bailey, Cole, Henderson, and Massey (2020) for overviews), with a strong focus on linking census records (Abramitzky, Mill, & Pérez, 2020; Bailey et al., 2020; Ruggles, Fitch, & Roberts, 2018). The approaches featured in the literature share many commonalities, such as 1) using logical criteria to rule out certain matches, like linking records where the date of death in one record is before the date of birth in another, 2) using edit distance, a measure of how many changes are needed to convert one string into another, to allow for small variations in names, and 3) using weights or probabilities to rank or filter candidate links. But there are also many differences, primarily due to variations in the data being linked and the reliability of different data elements (Abramitzky et al., 2020). The muster records in ‘VOC-opvarenden’ can only be linked by name and place of origin, with other data elements enabling the assessment of the likelihood or feasibility of matches. In our approach to creating accurate clusters of records linked to individuals, we prioritized quality over quantity. We applied limited normalization rules for names, including removing punctuation and standardizing suffixes. We then matched records based on name and origin similarity, allowing for small differences by permitting a maximum edit distance of 1 (meaning one change, such as a single letter substitution, addition, or deletion, is allowed), ensuring

no temporal overlap in voyages and no gaps of more than seven years between voyages. We based the maximum interval between voyages on a review of batches of clusters where varying interval criteria were applied. Additionally, there could be no records of deaths within a cluster, except at the end.

Abramitzky et al. (2020) used a probabilistic framework based on Dempster, Laird, and Rubin (1977) to estimate the likelihood that two records with the same name but different ages correspond to a true match. This method assumes that correct matches are more likely to have smaller age differences than incorrect matches. In our case, the period between voyages could be used in the same way, but we have no reason to assume that zero years between subsequent voyages is more likely than one, two or three. Therefore, we chose not to use probabilistic weighting of differences. In the case of multiple candidate records following a given record, with different gaps within the maximum of seven years, but overlapping with each other in muster period, we assumed the cluster to be ambiguous. Our linking is thus conservative to reduce Type I errors, accepting as a consequence an increase in Type II errors. We thus assume that the career dataset provides a lower bound for actual career lengths.⁷

Using the criteria above, we linked each individual record, except those records that describe crew members who boarded a ship at the Cape of Good Hope, to any other record (again, except Cape boarders) that could represent the same person. We thus turned the list of records into a graph of connected records. Once a record was linked to another, all linked name variants were used to link additional records. Finally, we integrated records of crew members who changed ship at the Cape of Good Hope, using similar criteria for person name and origin similarity, and ensuring ship name consistency in the original and Cape boarder records. The linking process resulted in two record categories:

Disambiguated

Any cluster of records that does not violate any of our criteria for assuming that records represent the same person, is considered to represent an individual who re-enlisted one or multiple times. Any record that could not be linked to another record according to the criteria for matching person names and places of origin described above, is considered to represent an individual who went on a single voyage for the VOC.

Ambiguous

Any record linked to another based on the criteria for matching person names and places of origin, but violating at least one of the other criteria.

This process identified 547,132 disambiguated entries, representing 460,452 unique individuals.

Place name standardization

The original ‘VOC-opvarenden’ database contains 152,315 unique toponyms for the origins of the crew members. Clerks at the VOC recorded the place names based on what they thought they heard a new recruit say and all had their own style of abbreviation, resulting in a wide range of spellings for the same locations. In the digitization project, no standardization of place names was undertaken. The town of The Hague, for example, is represented in 333 different forms in the original dataset. In our place name standardization work, we leveraged existing databases and did manual work. Drawing from the ErfGeo project,⁸ we identified modern equivalents for many historical Dutch toponyms, covering 5,407 origin attestations. Additionally, we incorporated data from the Sound Toll Registers online project,⁹ which offered standardized names for early modern attestations of port towns, especially in Scandinavia and the Baltic region, and from a dataset compiled by Ton van Velzen, who led the digitization of the VOC pay ledgers, which provided standardized places for an additional 2,100 original attestations. Using the bigram clustering function of OpenRefine,¹⁰ we matched as yet unstandardized toponym attestations to attestations for which we had a standardized equivalent from one of the datasets described above.

⁷ For further evaluation details, see Petram, Koolen, Wevers, van Koert, and van Lottum (2024, sec. 3.1).

⁸ <https://linkeddata.cultureelerfgoed.nl/rce/erfgeo>.

⁹ <http://www.soundtoll.nl/>.

¹⁰ <https://openrefine.org/>.

We then manually reviewed the data and used Google Maps to find additional places based on their phonetic Dutch spelling. Faced with numerous unstandardized and ambiguous toponym attestations, our goal was not to standardize every attestation but to accurately identify as many VOC crew members' places of origin as possible. We prioritized toponyms based on their frequency of occurrence, addressing the most common ones first. When a single clear modern place name candidate for a historical toponym emerged using the aforementioned resources, we recorded the modern toponym and its respective modern country's ISO 3166-1 alpha 2 code. If there were multiple candidates, we based our standardization on 1) size—if one place clearly has more inhabitants than the other(s), and it is likely that this was also the case in the eighteenth century, we argue that it is plausible that the VOC employee came from the larger place; 2) location—if one place is located in a region where many VOC employees came from (for example, the Netherlands, coastal and central Germany, the Baltic region, and coastal regions of Scandinavia), and the other(s) not, we argue that it is plausible that the VOC employee came from the place located in the core area. If there were multiple plausible candidates, we left the place of origin undefined. Once a likely place of origin was identified, curators also standardized similar toponyms from the list.

Finally, we applied automated matching against 37 historical European GIS maps for all remaining origin attestations, resulting in standardized places for an additional 7,053 origins. Combined, our approaches led to the standardization of 44,152 original place attestations to 8,591 modern toponyms, ensuring over 82% of muster records now have identified origins. The map in [Figure 3](#) illustrates the European origins of disambiguated individuals and the delineation of nine regions to which we assigned the places.

Ranks and wages enrichment

To address ambiguities and digitization inconsistencies, we categorized the 197 ranks in the original datasets under 107 parent categories. This process was aided by descriptions from the original dataset, and our wage data that we collected for a 5,335-record sample, spanning various ranks, origins, and periods. We also manually checked common mistakes with the scan of the original entry, resulting in 725 corrections. Ranks and their parent categories were classified into six broader categories (Trade, Sea, Ship, Medical, Military, and Other) based on their specific duties. Each rank was assigned a Historical International Standard Classification of Occupations (HISCO) code to reflect its hierarchy and job type, facilitating historical occupational classification ([van Leeuwen, Edvinsson, Maas, & Miles, 2002](#)). [Table 2](#) summarizes our work on ranks. It shows the most frequently occurring parent ranks, hierarchically organized within their categories, and data from our wage sample. The high wage variance among captains and some other ranks is due to skill premiums. Additionally, the variance in wages for ship carpenters and junior ship carpenters is high because of inconsistencies in how the various ship carpenter ranks were transcribed during digitization.

CATEGORY	PARENT RANK	MEDIAN WAGE	AVERAGE WAGE	STANDARD DEVIATION	n WAGE SAMPLE	# RECORDS IN DATASET
TRADE	senior merchant	135	115.09	41.01	44	146
TRADE	merchant	40	41.19	5.50	42	1,322
TRADE	merchant's assistant	24	40.95	28.06	55	1,629
SEA	captain	72	73.36	11.85	315	3,134
SEA	first mate	48	47.92	2.06	294	3,239
SEA	second mate	32	32.18	2.78	234	3,726
SEA	third mate	26	26.05	0.70	202	5,074
SEA	boatswain	22	19.92	3.84	134	6,909
SEA	junior boatswain	14	14.01	0.44	67	5,801
SEA	quartermaster	14	13.98	0.65	60	10,845
SEA	sailor	11	10.56	1.35	540	274,455
SEA	junior sailor	7	6.68	1.16	244	96,036
SHIP	ship carpenter	48	45.37	4.19	144	2,412
SHIP	junior ship carpenter	28	28.20	7.93	451	16,488
SHIP	senior sail maker	20	20.46	2.55	48	2,683

(Contd.)

Table 2 Wage sample statistics for top 43 parent ranks, organized by category. All wages are nominal and expressed in guilders per month. As in the Dutch Republic and other parts of the world, VOC wages remained stable throughout the entire eighteenth century ([Allen et al., 2011](#)), the period from which the bulk of the data originates. Consequently, we did not differentiate the data by time period in the table. Note that the SHIP and MILITARY categories have one or more sub-hierarchies. Note also that the OTHER category and certain less frequently occurring parent ranks from the five main categories (together representing 31,714 records or 4% of the total) were omitted from this table.

CATEGORY	PARENT RANK	MEDIAN WAGE	AVERAGE WAGE	STANDARD DEVIATION	n WAGE SAMPLE	# RECORDS IN DATASET
SHIP	sail maker	16	16.67	2.94	27	814
SHIP	junior sail maker	14	13.75	1.83	32	2,902
SHIP	weapons master	22	21.92	0.52	74	3,228
SHIP	junior weapons master	14	14.09	1.53	75	7,074
SHIP	master-at-arms	13	12.57	0.50	35	2,800
SHIP	cook	20	20.07	0.86	92	3,263
SHIP	junior cook	14	14.07	0.47	41	3,098
SHIP	water maker	18	17.61	1.50	31	215
SHIP	junior water maker	12	12.00	0.00	9	111
SHIP	steward	20	19.50	1.71	60	3,235
SHIP	junior steward	14	11.61	2.86	49	3,177
SHIP	senior cooper	16	16.69	1.22	91	2,775
SHIP	cooper	14	14.51	2.09	55	827
SHIP	junior cooper	14	13.82	0.80	98	3,902
MEDICAL	senior barber-surgeon	36	36.74	2.51	113	3,173
MEDICAL	barber-surgeon	26	24.99	3.64	162	4,092
MEDICAL	junior barber-surgeon	14	14.15	1.64	104	2,768
MILITARY	military captain	80	81.56	17.31	9	120
MILITARY	lieutenant	50	47.58	8.47	38	193
MILITARY	ensign	40	40.00	0.00	11	91
MILITARY	bombardier	24	24.67	5.16	6	72
MILITARY	sergeant	20	20.00	0.00	61	3,124
MILITARY	corporal	14	14.00	0.00	46	6,227
MILITARY	lance-corporal	12	12.79	1.03	86	6,038
MILITARY	soldier	9	9.21	0.47	197	241,102
MILITARY	junior soldier	7	7.00	0.00	23	1,844
MILITARY	weapon maker	14	13.97	0.80	87	2,230
MILITARY	junior weapon maker	9	9.00	0.00	2	92
Total					4,588	742,486

DATA QUALITY

95% (737,467) of all muster records in the ‘VOC-opvarenden’ database come from the 3,020 VOC pay ledgers that have survived until today. So-called ‘verzoekboeken’ (request books, documenting salary payment requests by VOC personnel in Asia to individuals in the Netherlands) and regiment books (detailing foreign soldiers hired by the VOC) are the source for the remaining 5%. Utilizing data from the Dutch-Asiatic Shipping (DAS) database,¹¹ which provides a comprehensive list of VOC voyages between the Dutch Republic and Asia, we assessed the completeness of the employment records.

Table 3 presents statistical insights into source availability by decade. It shows that records prior to 1680 are fragmentary. For the final two decades of the seventeenth century, pay ledgers are available for nearly half of the outgoing voyages, supplemented by data from request books. Starting from the early eighteenth century until 1780, an almost complete collection of pay ledgers is available. The decrease in the number of pay ledgers compared to outward voyages in the final two decades of the VOC is mainly due to changes in the VOC’s hiring policies: during this period, the VOC engaged foreign regiments of soldiers (for which data are accessible from regiment books) and chartered foreign ships with their crews (van Velzen, 2006). Although these chartered ship voyages are included in the DAS database, they did not involve seafarers mustering with the VOC, leading to the absence of pay ledgers. In summary, structural analyses of the VOC’s seafaring workforce are best supported by data from the eighteenth century.

¹¹ <https://resources.huygens.knaw.nl/das>.

DECADE	# OUTWARD VOYAGES (DAS)	# PAY LEDGERS	# REQUEST BOOKS	# REGIMENT BOOKS	% OF VOYAGES FOR WHICH PAY LEDGER IS AVAILABLE
1600–1609	96	0	0	0	0.0%
1610–1619	112	0	0	0	0.0%
1620–1629	145	0	0	0	0.0%
1630–1639	142	7	0	0	4.9%
1640–1649	181	10	0	0	5.5%
1650–1659	205	0	0	0	0.0%
1660–1669	241	4	0	0	1.7%
1670–1679	225	5	29	0	2.2%
1680–1689	204	88	49	0	43.1%
1690–1699	233	108	46	0	46.4%
1700–1709	272	255	2	0	93.8%
1710–1719	318	307	2	0	96.5%
1720–1729	377	360	6	0	95.5%
1730–1739	381	373	2	0	97.9%
1740–1749	314	307	1	0	97.8%
1750–1759	292	287	1	0	98.3%
1760–1769	295	294	0	0	99.7%
1770–1779	293	292	0	0	99.7%
1780–1789	290	218	3	43	75.2%
1790–1794	133	102	1	42	76.7%
Total	4787	3017	142	85	

Table 3 Statistics on the availability of sources per decade. The table excludes 3 pay ledgers and 1 regiment book that could not be linked to a DAS voyage.

The quality of our data enrichments cannot be validated by external sources. However, we do have a key piece of external information for evaluation of the clusters of an individual's records: month letter beneficiaries. When clustered individuals consistently nominate the same beneficiary across multiple voyages, it strengthens the likelihood that the cluster represents a single person. Of the 3,007 clusters suitable for this evaluation method, we found that beneficiary relations support our clustering in 92% of cases. We are confident that the rate of valid clusters is even higher, since name variations, transcription errors, and alternating use of initials and full names hamper the analysis of beneficiary relations.

Further, indirect validation of our clustering method comes from the distribution in the frequency of individuals signing up with the VOC. Considering that employment periods spanned several years and had a high mortality rate, we expect a highly skewed distribution. The observed power-law distribution (bottom chart in [Figure 2](#)) suggests there are no data artifacts and provides confidence that the clustering is reliable. However, our automatic clustering of individual records may still have occasionally conflated records of distinct individuals or fragment records of the same person.

As for the place name standardization, the distribution of disambiguated records over regions in the chart in [Figure 3](#) confirms that most records are from the Dutch Republic and Dutch-speaking Southern Netherlands. This is in line with previous research ([Bruijn, Gaastra, & Vermeulen, 1987](#); [Bruijn & Lucassen, 1980](#)). However, a potential bias towards Dutch-speaking European places might exist for two reasons. First, VOC clerks were likely more familiar with Dutch places and had easier communication with prospective employees from these regions, probably resulting in less spelling variation for origin notations. The people who transcribed the places of origins and entered them into the original 'VOC-opvarenden' database were also of Dutch origin, again probably leading to fewer errors and lower spelling variation for Dutch toponyms. Second, the standardization process, which used some resources with a strong focus on the Netherlands and involved Dutch data curators, might again have favored Dutch places.

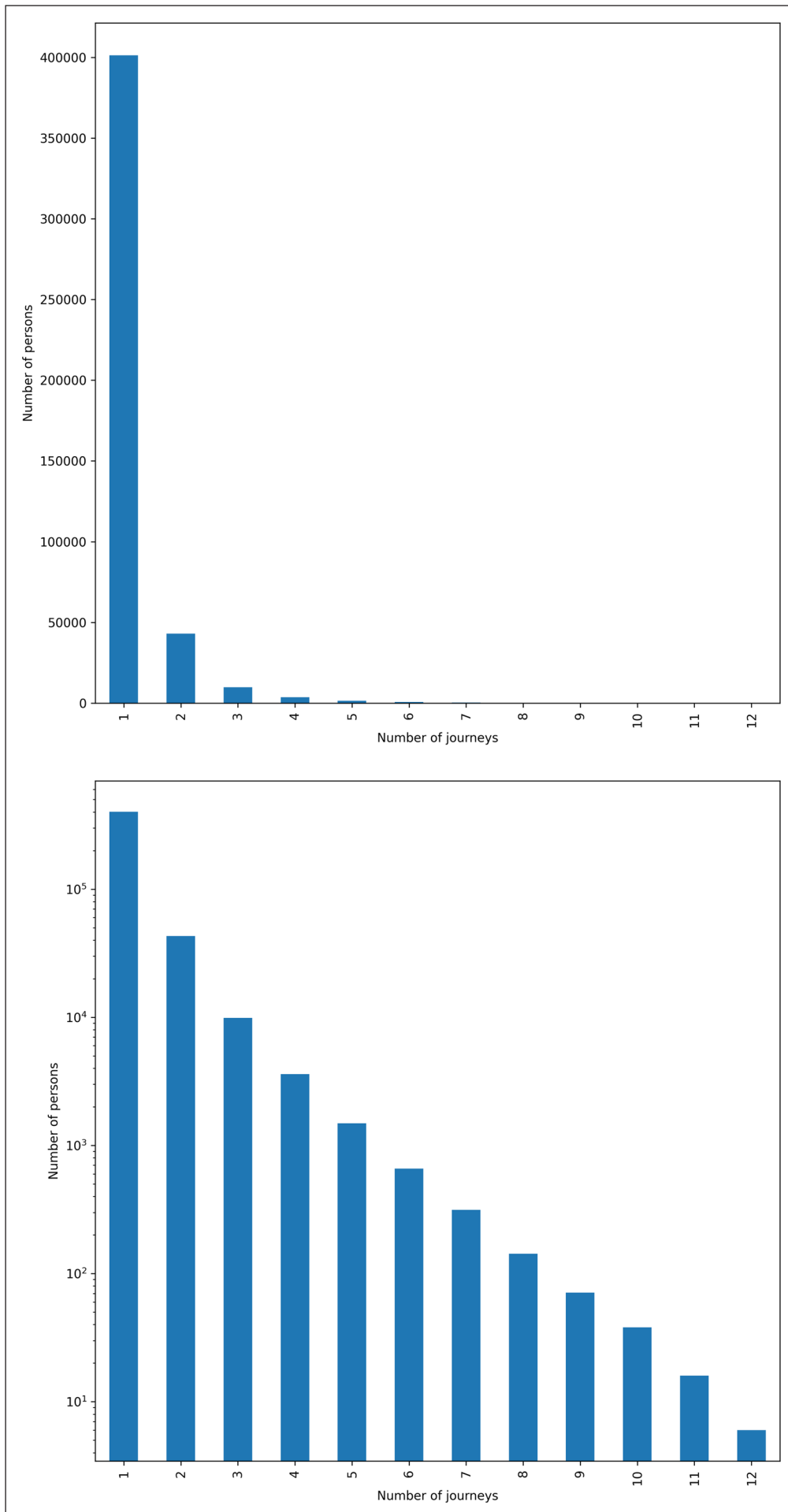


Figure 2 The distribution of the number of VOC employments per (disambiguated) individual, with the Y axis scaled normally on the top and logarithmically on the bottom.

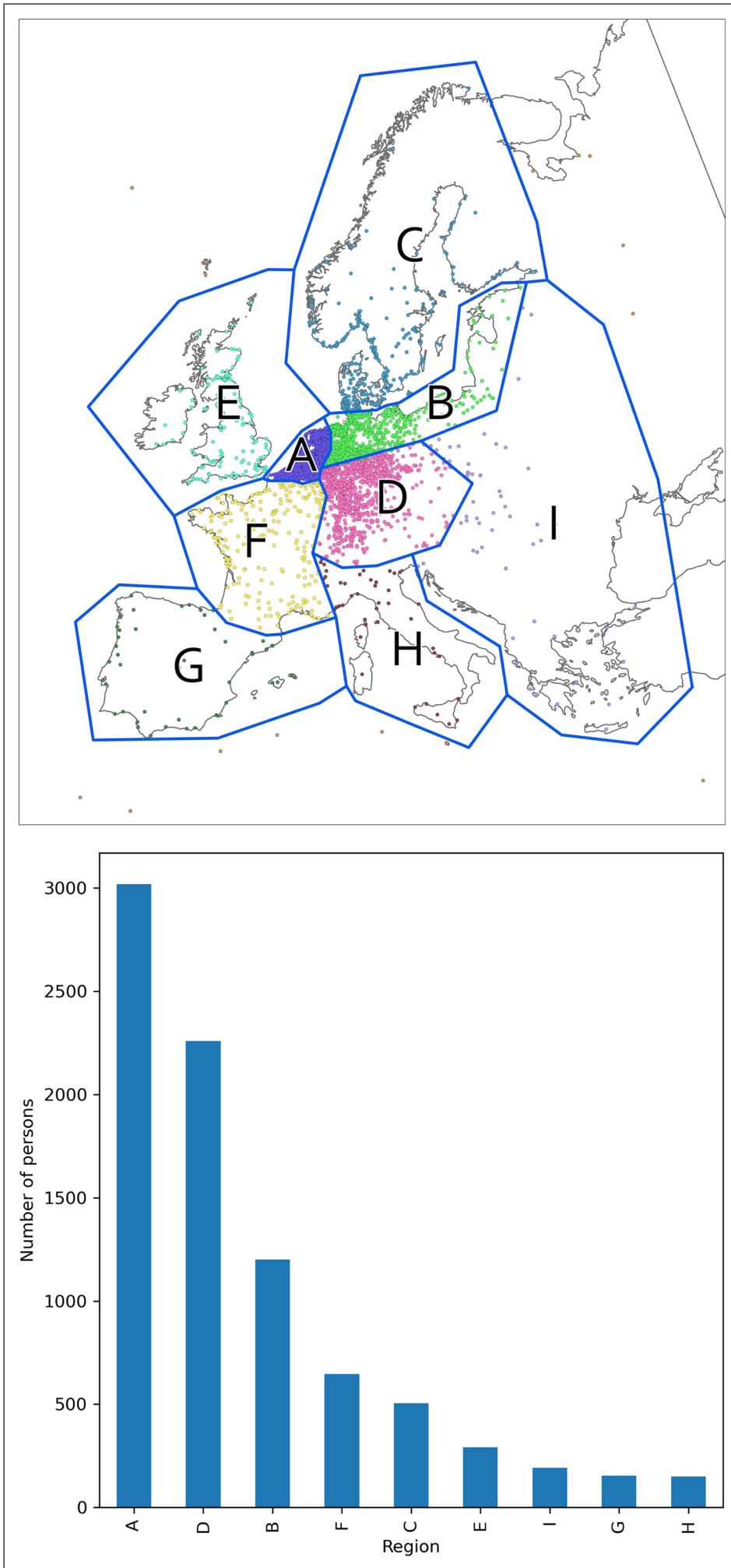


Figure 3 Map showing the region categorization (top) and distribution of disambiguated records over the regions. We distinguish regions based on language, maritime orientation, and proximity to the Netherlands. They can broadly be defined as follows: A. The Dutch Republic and Dutch-speaking area of the Southern Netherlands; B. Low German-speaking area; C. Scandinavia; D. German interior lands; E. British Isles; F. France; G. Iberian Peninsula; H. Modern-day Italy and Corsica; I. Eastern and Southeastern Europe.

Although the semi-automated linkage with GIS collections will have reduced this bias, non-Dutch places could be over-represented in the non-standardized toponyms. It is important to stress, moreover, that standardizing toponyms from an early modern source with limited context is imprecise. The process, partly based on assumptions and human judgment, likely produced incorrect matches. In short, the standardization of place names, though extensive, and significantly enhancing the dataset for geographic and demographic analysis, is not comprehensive and error-free, which may affect detailed geographic analyses.

4 DISCUSSION AND REUSE POTENTIAL

Our dataset is unique for the early modern period and unlocks many opportunities for broad, large-scale research and statistical analyses. Two forthcoming papers by the authors of this paper are leveraging the data to explore VOC career paths and the influence of networks on international recruitment. However, the dataset's utility spans further, offering great potential for sociological studies akin to those by Sgourev and van Lent (2017); Wezel and Ruef (2017, 2020, 2024). It also facilitates network analysis and the comparison of VOC crew characteristics with those of, for example, the French *Compagnie des Indes*. By integrating economic and social data from European regions into our enriched dataset, researchers could uncover fascinating insights on early modern history such as migration trends and their determinants.

There are also numerous opportunities to further enrich our dataset. The GLOBALISE project¹² plans to augment it with annual records of the VOC workforce in Asia, further enriching the data on the European crew members' lives and careers, and adding information on non-European personnel. It also plans to release the integrated collection as Linked Open Data, enhancing its accessibility and potential for integration with other datasets, such as those mentioned in section 1. Linking the VOC records to, for example, the databases of the MCC, the French *Compagnie des Indes*, or data from single crew lists preserved in the Prize Papers Collection¹³ and many other archival collections, would allow researchers to explore whether the VOC crew members also sailed for other companies.

Additional information about the life courses of the VOC crew members can be obtained from baptism, marriage, and burial records from across Europe, which are often digitized and searchable. For qualitative insights into the lives of Dutch VOC personnel, and the time spent in the Netherlands by foreign crew members, researchers can explore various digitized Dutch source collections, such as notarial deeds from Amsterdam,¹⁴ and the GLOBALISE source corpus,¹⁵ featuring nearly 5 million scans of VOC-related records, letters, and other documents, that were sent over from Asia to the Netherlands in the seventeenth and eighteenth centuries. Both resources support full-text searches.

The dataset also appeals to audiences beyond academic and genealogical researchers. The 2021 VOC Data Experience,¹⁶ an augmented reality-based museum installation, draws from an earlier version of this enriched dataset to offer immersive visualizations that explore the complex and often challenging history of the VOC.

While we do not anticipate major further enhancements to the dataset, we encourage the academic community and the public to contribute to its accuracy by submitting corrections or additions, especially regarding places of origin. If you have substantial corrections or additions, please reach out to the first author.

While using this dataset, researchers should be aware of its limitations: it only covers VOC personnel who signed on in the Netherlands, might include inaccuracies in clustering, and lacks precise locations for 18% of records, which may affect detailed geographic analyses. Despite these challenges, the dataset significantly enhances research capabilities into the lives and careers of about half a million Europeans who decided to sign on to a ship bound for Asia in the eighteenth century.

¹² <https://globalise.huygens.knaw.nl/>.

¹³ <https://www.prizepapers.de/> and <https://prizepapers.huygens.knaw.nl/>.

¹⁴ <https://amsterdam-city-archives.transkribus.eu/>.

¹⁵ <https://transcriptions.globalise.huygens.knaw.nl/>.

¹⁶ <https://vocdataexperience.nl/>.

COLUMN NAME	DESCRIPTION
source_id	record id; each record describes a physical archival document that contains data on the crew of a particular voyage of a VOC ship
ship_name	name of ship for which the source document provides data on the crew of a particular voyage
chamber	VOC chamber that managed the ship voyage and its crew
remarks	general remarks on record (in Dutch)
archival_reference	reference to inventory number in VOC archives at NA
uid	unique record id in database of Dutch National Archives
source_type	specification of source type (pay ledger, request book or regiment book)
das_voyage_id	unique id of voyage, refers to file VOC_voyages.csv

Table 4 Description of variables included in file `voc_sources.csv`.

COLUMN NAME	DESCRIPTION
vocop_id	record id; each record describes a muster record of a VOC employee that was recorded in the pay ledger of a particular voyage of a VOC ship
place_of_origin	place of origin of employee
place_id	id of place of origin, refers to file VOC_places.csv
disambiguated_person	1 if record belongs to a disambiguated person, else 0
person_cluster_id	id for cluster of records belonging to a disambiguated person
person_cluster_row	ordinal rank of the voyage in a disambiguated person's career ('3' would e.g. indicate the third contract within a cluster)
rank_corrected	1 if the rank stated in the original data set has been altered (correction of wrong transcription; distinction ranks that changed after 1783), else 0
rank_id	unique id of rank, refers to file voc_ranks.csv
debt_letter	1 if employee signed a debt letter, else 0
month_letter	1 if employee signed a month letter, else 0
date_begin_contract	start date of employment contract
date_end_contract	end date of contract; incomplete dates in original data set were supplemented with arrival dates of return ships from DAS and/or completed with '-01' or '-01-01' to facilitate calculations with the dates
could_muster_again	1 if employee could muster again on a subsequent voyage, given the reason why the current employment contract ended, else 0
location_end_contract	location where the employment contract ended; when an employee returned to the Netherlands and signed off, the name of the return ship is given
outward_voyage_id	DAS voyage id of outward voyage, refers to file voc_voyages.csv
changed_ship_at_cape	1 if employee changed ship at the Cape of Good Hope, else 0
changed_ship_at_cape_voyage_id	DAS voyage id of outward voyage that employee joined from the Cape of Good Hope, refers to file voc_voyages.csv
return_voyage_id	DAS voyage id of return voyage, refers to file voc_voyages.csv
remark	general remark on record (in Dutch)
source_id	id of source record from which muster record originates, refers to file voc_sources.csv
source_reference	reference to finding place of record in the paper VOC archives (inventory number and page number)
uid	unique record id in database of Dutch National Archives
scan_permalink	permalink to scan of physical muster record

Table 5 Description of variables included in file `voc_persons_contracts.csv`.

COLUMN NAME	DESCRIPTION
vocop_id	muster record id from which name originates, refers to file voc_persons_contracts.csv
first_name_original	first name of employee
patronymic_original	patronymic of employee
family_name_prefix_original	family name prefix of employee
family_name_original	family name of employee
full_name_normalized	normalized full name of employee
first_name_normalized	normalized first name of employee
patronymic_normalized	normalized patronymic of employee
family_name_prefix_normalized	normalized family name prefix of employee
family_name_normalized	normalized family name of employee

Table 6 Description of variables included in file `voc_names.csv`.

COLUMN NAME	DESCRIPTION
place_id	record id; each record describes a place of origin mentioned in the VOC muster records
place_original	toponym attestation in original data set
place_normalized	normalized toponym attestation
place_standardized_id	unique id of standardized place, refers to file voc_places_standardized.csv; left blank if place has not been standardized
provenance	code specifying the provenance of the standardization

Table 7 Description of variables included in file `voc_places.csv`.

COLUMN NAME	DESCRIPTION
place_standardized_id	record id; each record describes a standardized place of origin of a VOC employee
place_standardized	standardized placename
country_code	ISO 3166-1 alpha 2 code of country in which standardized place is presently located
place_standardized_cc	concatenation of standardized placename and country code
authority_file_uri	GeoNames or Wikidata URI of standardized place (Wikidata used when GeoNames unavailable, e.g. for historical regions)
latitude	latitude of place of origin, as provided by authority file
longitude	longitude of place of origin, as provided by authority file
region	region in which standardized place is located (A: Dutch Republic and Dutch-speaking area of Southern Netherlands, B: Low German-speaking area, C: Scandinavia, D: German interior lands, E: British Isles, F: France, G: Iberian Peninsula, H: Italy and Corsica, I: Eastern and Southeastern Europe)

Table 8 Description of variables included in file `voc_places_standardized.csv`.

COLUMN NAME	DESCRIPTION
rank_id	record id; each record describes a rank on board of a VOC ship
rank	rank designation
parent_rank	parent rank, grouping together one or more similar ranks
category	rank categorization (trade, sea, ship, medical, military or other)
subcategory	further rank categorization, e.g. distinguishing between commissioned and non-commissioned officers
hisco	HISCO code for rank (see https://historyofwork.iisg.nl/)
hico_uri	URI for HISCO code
rank_nl	rank designation in Dutch
rank_description_nl	description of rank in Dutch
rank_description_eng	description of rank in English
median_wage	median wage of rank, calculated from wage sample

Table 9 Description of variables included in file `voc_ranks.csv`.

COLUMN NAME	DESCRIPTION
das_voyage_id	record id; each record describes a voyage of a VOC ship
outward_return	1 if outward voyage; 0 if return voyage
ship_name	name of ship
ship_tonnage	tonnage of ship
chamber	VOC chamber that managed the voyage
departure_date	date of departure
departure_place	place of departure
arrival_date_cape	date of arrival at the stopover at Cape of Good Hope
departure_date_cape	date of departure from Cape of Good Hope
arrival_date	date of arrival at destination
arrival_place	destination place

Table 10 Description of variables included in file `voc_voyages.csv`.

COLUMN NAME	DESCRIPTION
beneficiary_id	record id; each record describes the beneficiary of a month letter and their relation to the VOC employee who signed the letter
full_name	full name of beneficiary
first_name	first name of beneficiary
patronymic	patronymic of beneficiary
family_name_prefix	family name prefix of beneficiary
family_name	family name of beneficiary
place_of_origin	place of origin of beneficiary
relation	beneficiary's relation to VOC employee
remark	remark on month letter and/or beneficiary relation; in case the beneficiary was an institution, this field often mentions the name of the institution
vocop_id	id of muster record to which beneficiary is related, refers to file <code>VOC_persons_contracts.csv</code>
uid	unique record id in database of Dutch National Archives

Table 11 Description of variables included in file `voc_beneficiaries.csv`.

ACKNOWLEDGEMENTS

Ton van Velzen (unaffiliated, Amsterdam, The Netherlands): data provision and clarifications on the creation of the original dataset.

Ania Ahamed and Jessica den Oudsten (Huygens Institute, Amsterdam, The Netherlands): manual data linkage pilot.

Zhoubeini Zhang (Huygens Institute, Amsterdam, The Netherlands): descriptive statistics.

FUNDING INFORMATION

Dataset creation was funded in the 2016 round of CLARIAH¹⁷ research pilots.

COMPETING INTERESTS


The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Lodewijk Petram  orcid.org/0000-0002-7246-4176
 Huygens Institute, Amsterdam, The Netherlands

Marijn Koolen  orcid.org/0000-0002-0301-2029
 DHLab, KNAW Humanities Cluster, Amsterdam, The Netherlands

Melvin Wevers  orcid.org/0000-0001-8177-4582
 Department of History, University of Amsterdam, Amsterdam, The Netherlands

Jelle van Lottum  orcid.org/0000-0003-0534-4745
 Huygens Institute, Amsterdam, The Netherlands; Radboud Institute for Culture and History, Radboud University, Nijmegen, The Netherlands

¹⁷ <https://www.clariah.nl/>.

- Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3), 865–918. DOI: <https://doi.org/10.1257/jel.20201599>
- Abramitzky, R., Mill, R., & Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2), 94–111. DOI: <https://doi.org/10.1080/01615440.2018.1543034>
- Allen, R. C., Bassino, J.-P., Ma, D., Moll-Murata, C., & Van Zanden, J. L. (2011). Wages, prices, and living standards in China, 1738–1925: in comparison with Europe, Japan, and India. *The Economic History Review*, 64(s1), 8–38. DOI: <https://doi.org/10.1111/j.1468-0289.2010.00515.x>
- Bailey, M. J., Cole, C., Henderson, M., & Massey, C. (2020). How well do automated linking methods perform? Lessons from us historical data. *Journal of Economic Literature*, 58(4), 997–1044. DOI: <https://doi.org/10.1257/jel.20191526>
- Bruijn, J. R., & Gaastra, F. S. (2012). The career ladder to the top of the Dutch East India Company. Could foreigners also become commanders and junior merchants? In M. van der Linden & L. Lucassen (Eds.), *Working on labor* (pp. 213–235). Leiden: Brill. DOI: https://doi.org/10.1163/9789004231443_011
- Bruijn, J. R., Gaastra, F. S., & Vermeulen, A. C. F. (1987). *Dutch-Asiatic shipping in the 17th and 18th centuries/Vol. I Introductory volume*. The Hague: Nijhoff.
- Bruijn, J. R., & Lucassen, J. (1980). *Op de schepen der Oost-Indische Compagnie. Vijf artikelen van J. de Hullu ingeleid, bewerkt en voorzien van een studie over de werkgelegenheid bij de VOC*. Groningen: Wolters-Noordhoff/Bouma's Boekhuis.
- de Boer, V., van Rossum, M., Leinenga, J., & Hoekstra, R. (2014). Dutch Ships and Sailors Linked Data. In P. Mika et al. (Eds.), *The Semantic Web – ISWC 2014* (pp. 229–244). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-11964-9_15
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: series B (methodological)*, 39(1), 1–22. Retrieved from <https://www.jstor.org/stable/2984875>
- Gaastra, F. S. (1991). *De geschiedenis van de VOC*. Zutphen: Walburg Pers.
- Hoyer, F. (2020). *Relations of Absence. Germans in the East Indies and Their Families c. 1750–1820* (Unpublished doctoral dissertation). Acta Universitatis Upsaliensis. Retrieved from <https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-425763>
- Ketelaars, M. (2014). *Compagniesdochters. Vrouwen en de VOC*. Amsterdam: Uitgeverij Balans.
- Odegard, E. (2022). In search of sepoys. Indian soldiers and the Dutch East India Company in India and Sri Lanka, 1760–1795. *War in History*, 29(3), 543–562. DOI: <https://doi.org/10.1177/09683445211030301>
- Petram, L., Koolen, M., Wevers, M., van Koert, R., & van Lottum, J. (2024, February). The Dutch East India Company's Eighteenth-Century Workforce: an Enriched Data Collection. *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.10599528>
- Pusztai, G., & Teszelszky, K. (2016). In dienst van de VOC. Een voorlopige inventarisatie van Hongaren in dienst van de Verenigde Oost-Indische Compagnie (1602–1795). *Acta Neerlandica*, 12, 25–108. Retrieved from <https://hdl.handle.net/11370/dd43f111-6434-4dfe-b22b-01156cb28687>
- Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44, 19–37. DOI: <https://doi.org/10.1146/annurev-soc-073117-041447>
- Sgourev, S. V., & van Lent, W. (2017). When too many are not enough. Human resource slack and performance at the Dutch East India Company (1700–1795). *Human Relations*, 70(11), 1293–1315. DOI: <https://doi.org/10.1177/0018726717691340>
- van Bochove, C. J. (2018). From Nijmegen to Asia in the eighteenth century. In P. Puschmann & T. Riswick (Eds.), *Building Bridges. Scholars, History and Historical Demography. A Festschrift in Honor of Professor Theo Engelen* (pp. 118–134). Nijmegen: Valkhof Pers. Retrieved from <https://hdl.handle.net/2066/196015>
- van Bochove, C. J., & van Velzen, T. (2014). Loans to salaried employees. The case of the Dutch East India Company, 1602–1794. *European Review of Economic History*, 18(1), 19–38. DOI: <https://doi.org/10.1093/ereh/het021>
- van Leeuwen, M. H. D., Edvinsson, S., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven U.P.
- van Lottum, J., & Petram, L. (2023). In Search of Strayed Englishmen. English Seamen Employed in the Dutch East India Company in the Late Seventeenth and Eighteenth Centuries. In S. Levelt, E. van Raamsdonk, & M. Rose (Eds.), *Anglo-Dutch Connections in the Early Modern World* (pp. 100–111). New York: Routledge. DOI: <https://doi.org/10.4324/9781003049180>
- van Rossum, M. (2014). *Werkers van de wereld. Globalisering, arbeid en interculturele ontmoetingen tussen Aziatische en Europese zeelieden in dienst van de VOC, 1600–1800*. Hilversum: Verloren.

- van Velzen, T. (2006). Uitgevaren voor de Kamers. 700.000 mensen overzee. In J. Parmentier (Ed.), *Uitgevaren voor de Kamer Zeeland* (pp. 11–30). Zutphen: Walburg Pers.
- Wevers, M., Karsdorp, F., & van Lottum, J. (2022). What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model. *CEUR Workshop Proceedings Proceedings, 1613*, 189–197. Retrieved from https://ceur-ws.org/Vol-3290/short_paper1793.pdf
- Wezel, F. C., & Ruef, M. (2017). Agents with Principles. The Control of Labor in the Dutch East India Company, 1700 to 1796. *American Sociological Review*, 82(5), 1009–1036. DOI: <https://doi.org/10.1177/0003122417718165>
- Wezel, F. C., & Ruef, M. (2020). Learning against the wind. Diversity and performance on the ships of the Dutch East India Company. *Strategy Science*, 5(4), 330–347. DOI: <https://doi.org/10.1287/stsc.2020.0114>
- Wezel, F. C., & Ruef, M. (2024). Cracking the Deck: National Origins and Promotions in the Dutch East India Company, 1700–1796. *Organization Studies*. DOI: <https://doi.org/10.1177/01708406241248985>
- Zijlstra, S. (2021). *De voormoeders. Een verborgen Nederlands-Indische familiegeschiedenis*. Amsterdam: Ambo|Anthos.

Petram et al.
*Journal of Open
 Humanities Data*
 DOI: 10.5334/johd.210

TO CITE THIS ARTICLE:

Petram, L., Koolen, M., Wevers, M., & van Lottum, J. (2024). Charting Lives and Careers: Enriched Data About the Dutch East India Company's Eighteenth-Century European Workforce. *Journal of Open Humanities Data*, 10: 44, pp. 1–17. DOI: <https://doi.org/10.5334/johd.210>

Submitted: 16 March 2024

Accepted: 26 June 2024

Published: 15 July 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.