

Beyond the Register: Demographic Modeling of Arrest Patterns in 1879-1880 Brussels

Folger Karsdorp^{1,*}, Mike Kestemont² and Margo de Koster³

¹*KNAW Meertens Institute, Amsterdam, the Netherlands*

²*University of Antwerp, Antwerp, Belgium*

³*Ghent University, Ghent, Belgium*

Abstract

Unseen species models from ecology have recently been applied to censored historical cultural datasets to estimate unobserved populations. We extend this approach to historical criminology, analyzing the police registers of Brussels' Amigo prison (1879-1880) using the Generalized Chao estimator. Our study aims to quantify the 'dark number' of unarrested perpetrators and model demographic biases in policing efforts. We investigate how factors such as age, gender, and origin influence arrest vulnerability. While all examined covariates contribute positively to our model, their small effect sizes limit the model's predictive performance. Our findings largely align with prior historical scholarship but suggest that demographic factors alone may insufficiently explain arrest patterns. The Generalized Chao estimator modestly improves population size estimates compared to simpler models. However, our results indicate that more refined models or additional data may be necessary for robust estimates in historical criminological studies. This work contributes to the growing field of computational methods in humanities research and offers insights into the challenges of quantifying hidden populations in historical datasets.

Keywords

unseen species model, Generalized Chao, survivorship bias, police history, historical criminology

1. Introduction: dark numbers in history

'Dark numbers' are a well-established concept in criminology and sociology, referring to unobserved or latent criminality not captured in official crime statistics [21]. This concept has gained traction in popular discourse, highlighting the gap between reported and actual crime rates. Contemporary examples of phenomena with significant 'dark numbers' include drug use in urban areas, undocumented migration, and domestic violence – all areas with strong evidence of underreporting [13]. These examples underscore the challenges in accurately quantifying certain types of criminal activity.

Researchers have developed various methods to estimate these undocumented numbers, ranging from comparing self-reported arrests with official statistics [19] to more sophisticated

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

✉ folgert.karsdorp@meertens.knaw.nl (F. Karsdorp); mike.kestemont@uantwerpen.be (M. Kestemont); margo.dekoster@ugent.be (M. d. Koster)

🌐 <https://www.karsdorp.io> (F. Karsdorp); <http://mikekestemont.github.io/> (M. Kestemont)

🆔 0000-0002-5958-0551 (F. Karsdorp); 0000-0003-3590-693X (M. Kestemont); 0000-0002-5566-9137 (M. d. Koster)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



statistical modeling approaches. While empirical validation of these estimates remains challenging, they provide valuable insights that can inform policy decisions and guide law enforcement strategies. In recent years, criminology has increasingly turned to advanced quantitative methods to address the negative bias in available data. Although these models have produced valuable insights into contemporary crime patterns, their application to historical data remains limited.

Our study aims to bridge this gap by applying one such modern technique, the Generalized Chao method [5], to a historical case study: the register of the Amigo prison in Brussels for the period 1879-1880. This unique source documents individuals imprisoned overnight in the municipal police's central detention institution. Given the complex nature of urban crime in a major 19th-century city like Brussels, it is reasonable to assume that police arrest efforts were imperfect, leaving a considerable amount of undetected crime. Our analysis aims to quantify this 'dark number' by estimating a lower bound on undetected crime in late-nineteenth century Brussels.

Furthermore, we leverage the demographic information available in the Amigo register to construct a model that predicts crime detection rates across different social groups. This approach allows us to explore historical vulnerability to arrest, potentially revealing differential treatment by the police and broader societal biases of the era. By applying contemporary statistical methods to historical data, we aim to provide new insights into the dynamics of crime and law enforcement in 19th-century urban environments.¹

Our research aligns with a growing scholarly interest in measuring the amount of under-detection in historical data, known as *Überlieferungschance* [10]. This issue of under-sampling is pervasive across disciplines studying the past, where available datasets often represent only a fraction of the original historical population, be it artifacts or individuals. Imperfect registration and survival rates result in incomplete and potentially biased historical data, often skewed towards categories with higher survival ratios due to their material nature (e.g., stone inscriptions versus papyrus writings). This situation exemplifies 'survivorship bias', a concept that has gained popular attention, partly due to Abraham Wald's seminal work on R.A.F. bombers during World War II [17, 20].² Humanities scholars have long recognized this precarious situation but often lacked a rigorous framework to address it systematically.

Recent scholarship has turned to unseen species models from ecology as an innovative approach to estimating the extent of lost or forgotten data, particularly in cultural heritage studies [15]. In the Humanities, these models have found surprising applications across diverse domains, including medieval literature, early modern collections of printed books, and studies of Dutch sailors [24, 18, 16]. Our work contributes to this emerging trend by applying these ecological models to historical criminology, offering new insights into the 'dark numbers' of 19th-century urban crime.

¹Reference to anonymized earlier conference presentation on which this paper is based.

²On social media, the same visual representation of an airplane (with bullet holes in specific areas) tends to be reproduced, but this is not a historic image: the origin of the image (the first version was created by around 2005) is traced in a blog: <https://web.archive.org/web/20240430093343/https://cameronmoll.com/journal/abraham-wald-red-bullet-holes-origin-story>. To the best of our knowledge, no published sketches by Wald himself are known. See e.g. [23, p. 58-60] but note that the illustration on p. 60 is also not historical.

2. Materials: the *Amigo* register of 1880

The primary source for our study is the *Amigo* register, a serial, handwritten source housed in the City Archives Brussels, within the records of the Brussels municipal police.³ This unique historical resource, illustrated in Fig. 1, takes a tabular form, with each row recording information about an individual who spent the night in the *Amigo* prison.⁴ Unfortunately, like many archival records of urban police forces, the *Amigo* register has suffered from poor conservation. For Brussels, only a few registers from the 1880s have survived. Our study focuses on the register covering the period from October 31, 1879, to November 1, 1880, encompassing all individuals detained in the Brussels police prison during this time.

2.1. Involuntary Arrests

A critical distinction in the *Amigo* register data is between two types of entries:

1. Involuntary detentions: Individuals forcefully arrested and detained by the police;
2. Voluntary stays (*nuit sur demande* or ‘night on request’): Individuals who voluntarily sought shelter at the *Amigo*.

The latter category primarily comprised socially vulnerable people lacking financial means or family support. These individuals requested to spend the night at the *Amigo* for shelter, rather than being brought in through active law enforcement efforts. The voluntary nature of the *nuit sure demande* entries contrasts sharply with the involuntary detentions, potentially impacting our interpretation of arrest patterns and policing efforts in late 19th-century Brussels.

Our analysis primarily focuses on involuntary detentions, specifically individuals arrested and locked up in the *Amigo* police prison on charges of ‘vagrancy’. At the end of the 19th century, the Belgian economy suffered a depression, which heightened concerns about the influx of foreign ‘vagrants’ and other perceived undesirable newcomers. In response, urban municipal authorities implemented stricter local ordinances targeting poor migrants. These included entry regulations and passport requirements at the city gates, as well as mandatory reporting by landlords and innkeepers about their lodgers’ characteristics and previous settlements. Concurrently, new anti-vagrancy legislation criminalized several conditions: (a) the inability to prove stable attachment to a local community, (b) failure to register as required by law, and (c) lack of a steady income. From 1866 onwards, a police arrest and judicial conviction for vagrancy or begging resulted in forced internment in State Vagrancy Colonies (*Rijkswel-dadigheidskolonies*). Before being transported to the vagrancy colonies, arrested vagrants were held in the police prison of the city where the arrest was made: these were the individuals who were registered as ‘vagrants’ or ‘beggars’ in the *Amigo* register, and who constitute our research population [9].

³City Archives Brussels (SAB), Records of the Brussels Municipal Police, *Amigo*, register “Vagabonds et Mendians” and register “Ivrognes, logés, prostituées”, 1880 (no series or piece number).

⁴The historical building of the *Amigo* prison in Brussels still exists today and currently houses a luxury hotel, centrally located, adjacent to the Grand-Place of Brussels. The name is a corruption of the older Dutch word *Vrunte*, which meant ‘place of detention’, misinterpreted during the Spanish occupation as ‘friend’ (Dutch *vriend*), hence the Spanish ‘*amigo*’.

N°	NOMS	PRÉNOMS	AGE	PROFESSION	LIEU DE NAISSANCE	MOIS DE DÉLIT							OBSERVATION				
						SEPTEMBRE	OCTOBRE	NOVEMBRE	DÉCEMBRE	JANVIER	FÉVRIER	MARS					
2372	Jacques	Charles	27	journalier	Uccle (Brux)												
2368	Domus	Jacques	37	"	"												
2369	Denil	Jean	31	journalier	Uccle (Brux)												
2370	Denil	Jean	31	journalier	Uccle (Brux)												
2371	Denil	Jean	31	journalier	Uccle (Brux)												
2372	Denil	Jean	31	journalier	Uccle (Brux)												
2373	Denil	Jean	31	journalier	Uccle (Brux)												
2374	Denil	Jean	31	journalier	Uccle (Brux)												
2375	Denil	Jean	31	journalier	Uccle (Brux)												
2376	Denil	Jean	31	journalier	Uccle (Brux)												
2377	Denil	Jean	31	journalier	Uccle (Brux)												
2378	Denil	Jean	31	journalier	Uccle (Brux)												
2379	Denil	Jean	31	journalier	Uccle (Brux)												
2380	Denil	Jean	31	journalier	Uccle (Brux)												
2381	Denil	Jean	31	journalier	Uccle (Brux)												
2382	Denil	Jean	31	journalier	Uccle (Brux)												
2383	Denil	Jean	31	journalier	Uccle (Brux)												
2384	Denil	Jean	31	journalier	Uccle (Brux)												
2385	Denil	Jean	31	journalier	Uccle (Brux)												
2386	Denil	Jean	31	journalier	Uccle (Brux)												
2387	Denil	Jean	31	journalier	Uccle (Brux)												
2388	Denil	Jean	31	journalier	Uccle (Brux)												
2389	Denil	Jean	31	journalier	Uccle (Brux)												
2390	Denil	Jean	31	journalier	Uccle (Brux)												

Figure 1: Photographic reproduction of two representative pages in the Amigo register of 1880, City Archives Brussels (SAB), fol. 95v–96r.

2.2. Recidivism

Recidivism is a prominent feature in our dataset, with many individuals experiencing multiple arrests. Table 1 illustrates the distribution of arrest counts. In total, the data holds evidence for 8,367 forceful arrests of 6,016 unique individuals. The mean number of nights an individual was locked up in the Amigo is therefore ~ 1.39 . However, this average masks a highly uneven distribution of arrests across perpetrators. While most individuals were registered only once, repeat offenders were common, with one extreme case involving 35 arrests for public intoxication within a single year.

Fig. 2 visualizes the temporal patterns of arrests for several individuals, revealing an interesting phenomenon of ‘burstiness’ in the data. Arrest events involving the same individual often form local temporal clusters. To investigate this pattern, we applied a simple Poisson model to all individuals arrested at least twice. This model predicts the time lag (in days) between an individual’s arrests based on their number of previous arrests. As illustrated in Fig. 3, the time lag generally decreases as the number of previous arrests increases. One way to interpret this is that the police gradually lost their patience with recidivists. Additionally, this suggests that the policing effort in arrests was not entirely neutral or random, but was guided by certain

Table 1

The distribution of the frequency with which individuals were arrested in the Amigo prison (f_1 = number of people arrested once, f_2 = number of people arrested exactly twice, etc.), incl. the total number of arrests (N) in the data and the unique individuals (S_{obs}).

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{16}	f_{22}	f_{35}	N	S_{obs}
4704	851	244	100	44	26	18	11	2	5	3	2	1	2	1	1	1	8367	6016

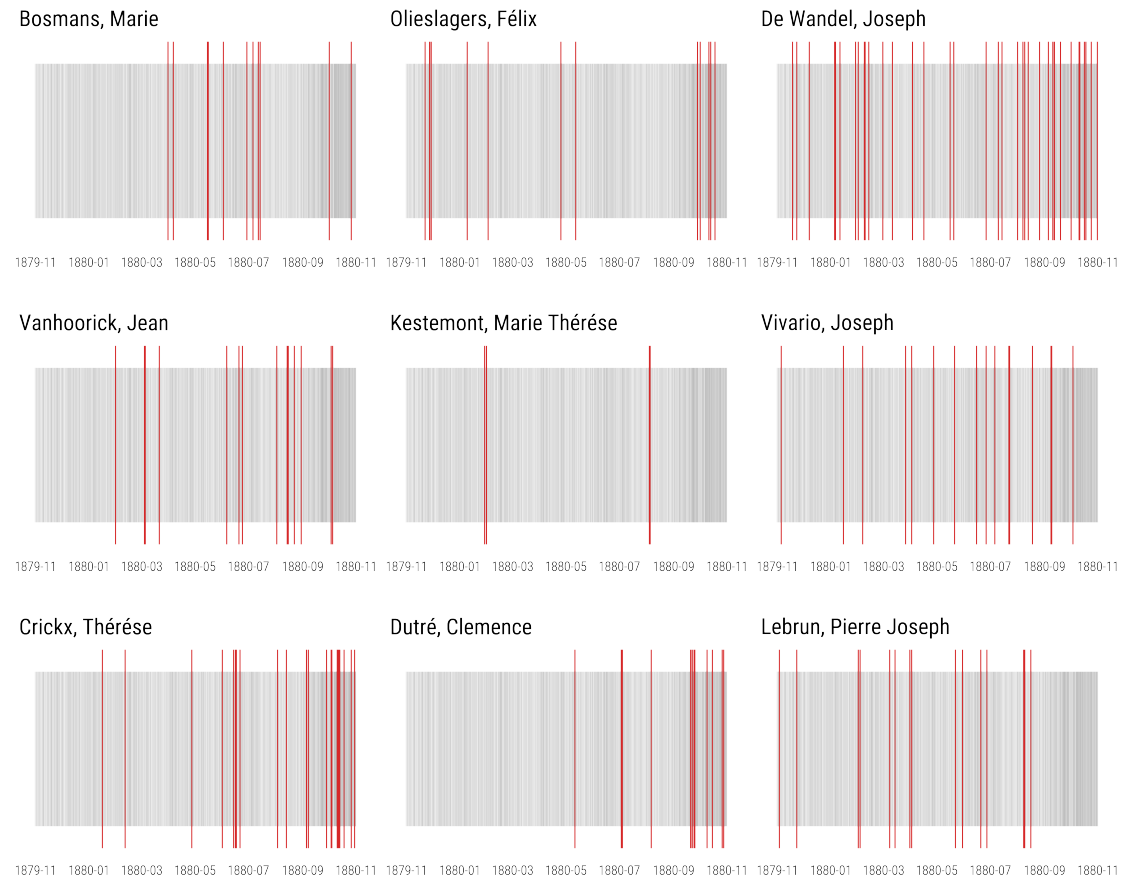


Figure 2: Visualization of the moments in time (red vertical stripes) when three representative individuals were arrested throughout the chronological range covered by the data. (Arrests of other individuals are shown as non-distinct gray vertical lines.) Note that these events are bursty and often cluster locally in time, indicating short time lapses between two consecutive arrests of the same perpetrator.

biases. The Amigo register holds information on the date of entry, but not on the release date: historians assume that the typical imprisonment was typically restricted to a short time period (i.e. a single night). The re-arrest data shows that individuals could indeed be re-registered after a short interval, which lends credibility to the assumption that incarceration was generally short-lived.

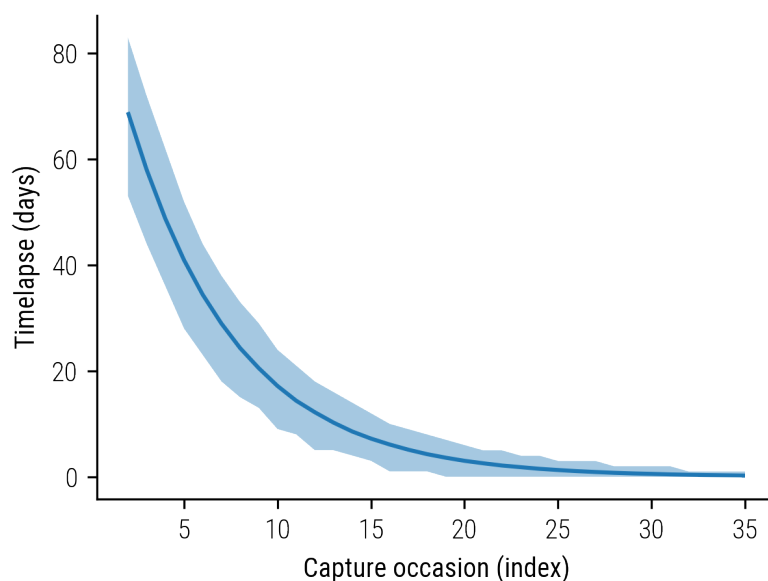


Figure 3: Predictive plot from a Poisson model (94% HDI) for recidivists (individuals with ≥ 2 arrests), showing how timelapse (i.e. the number of days after the previous arrest of an individual) varies as a function of the arrest occasion index (i.e. the number times an individual has been previously arrested).

3. Research Hypotheses

There exist valid reasons to assume that these data do not cover all of the individuals which theoretically could have been arrested by the police. One of these is the elevated number of singletons and doubletons in this distribution: if so many individuals were only recorded once or twice in the data, this renders it statistically quite likely that many perpetrators were never captured at all and don't appear in the register. The arrest data must therefore be considered incomplete and, in all likelihood, severely underestimate the true number of perpetrators in Brussels for the time period considered.

The methodology described below enables us to estimate a lower bound on these “dark numbers” in historical Brussels, i.e. *the number of perpetrators who were criminally active in this time period, but who were never formally registered by the police*. While this estimand presents a valuable scholarly objective in itself, this paper is also interested in characterizing this unobserved share of the criminal population in terms of demographics. Were men more likely to be arrested than women? Was the police effort more heavily geared towards younger individuals? Is there evidence of an active bias towards foreigners in comparison to local citizens? And, are there any intersections in these biases? Answering these questions would shed more light on the subjective drivers behind arrests and, thus, the historical vulnerability to arrest across demographic groups. On a more abstract level, explaining the drivers of under-detection in such historical datasets represents a major advance with respect to previous work in this area, which could only estimate the mere size of the non-observed share of the population, but not explain or characterize its composition.

Based on previous historical studies of the Amigo register [9], we formulated several hy-

potheses to guide our experimental design:

- H.1 **Impact of voluntary stays:** We hypothesize that a previous confinement on request (*nuit sur demande*) might have increased the visibility of socially fragile individuals to the local police, thus increasing their vulnerability to future arrest. For example, Joseph Janssens, a 25-year-old construction worker, was granted night shelter on March 20, 1880, but was arrested for public drunkenness just days later on March 28.
- H.2 **Age-related arrest patterns:** We expect that vulnerability to arrest generally increased with an individual's age. Prior work suggests that the police were more lenient towards younger individuals, particularly children, although minors were surprisingly often incarcerated. Among recidivists, there is a high presence of elderly citizens, who were more socially vulnerable due to a lack of suitable care institutions.
- H.3 **Gender disparities:** While there appear to be fewer criminally active women in the city, many women are among the recidivists. This suggests that the police effort may have been negatively biased towards socially vulnerable women.
- H.4 **Migratory status:** Contrary to older assumptions about discriminatory police biases against migrants, recent work found no clear evidence for this in the register's data. In fact, the data suggest that local, sedentary citizens were much more vulnerable to arrest than perpetrators born outside of Brussels.
- H.5 **Family-based biases:** We hypothesize that there existed negative biases against members of perceived "criminal families." The data shows frequent reappearance of the same family names across different individuals, sometimes with family members co-arrested. For instance, in autumn 1880, we find frequent arrests of 10 children (aged 8-10) from the same migrant family from Naples, Italy, who were active as beggars in the city.

After preprocessing and manual disambiguation of individuals appearing in the dataset, the following data was available for each individual:

- $\beta.1$ **prior:** Binary indicator ('prior' (reference level) and 'no prior') whether the individual was granted at least one "night on request" (*nuit sur demande*) prior to the date of their first arrest;
- $\beta.2$ **age:** An individual's mean age at the time of arrest, measured in years (scalar variable, centered and standardised).
- $\beta.3$ **sex:** An individual's biological sex (binary: 'female' or 'male' as reference level). We deliberately use the term 'sex' as the data contains no information on an individual's gender role.
- $\beta.4$ **origin:** Manually coded factor based on the individual's place of birth ('ABROAD' for individuals born outside of Belgium, 'BE' for those born in Belgium, and 'BXL' (reference level) for locally born individuals from Brussels).
- $\beta.5$ **family:** Binary indicator ('no family' (reference level) and 'family') whether an individual with the same family name occurs in the dataset. This variable aims to capture bias against known "criminal families" but should be interpreted cautiously, because we cannot rule out the possibility that unrelated individuals shared the same last name.

4. Estimating dark numbers with unseen species models

In this paper, we apply a so-called unseen species model to the Amigo data, borrowed from the biostatistical literature in ecology, to help solve the problem of estimating the number of unobserved perpetrators. Capture-recapture surveys are important bioregistration instruments in the field of ecology, used to monitor aspects of biodiversity, such as species richness (the number of unique species living in a certain area) [12]. During such campaigns, field workers use a variety of trapping devices (e.g. cameras) to register animals, mark them and release them again, so that they can be re-sighted at a later time. This process results in what is known as “abundance data”: counts that record how often animal types have been observed, such as singletons (f_1 or the number of species sighted exactly once), doubletons (f_2 or the number of species sighted exactly twice), etc. Because of the imperfect observation process, however, many animal types will not be observed during such campaigns, leading to an underestimation of the true ecological diversity (“unseen species”). The resulting count data must therefore be treated as censored, because it is zero-truncated: the number of relevant species which exist in the area but which were never observed (f_0) are missing. Statistical methods are therefore used to estimate f_0 as \hat{f}_0 and correct for the observation bias, by adding \hat{f}_0 to S (the number of observed species) to obtain an estimate of the true population size \hat{S} . Chao1 [7], for instance, is a widely used estimator that estimates a lower bound on f_0 as follows: $\hat{f}_0 \geq f_1^2/2f_2$.

4.1. Unseen heterogeneity

For theoretical reasons, it is important to stress that Chao1 only estimates a lower bound on the true f_0 , i.e. it estimates the minimum number of unobserved criminals; f_0 , in reality, could in fact could have been larger than \hat{f}_0 . (Readers should take time to convince themselves of the fact that, conversely, the detection ratio $S/(\hat{f}_0 + S)$ is an upper bound for that reason.) The fact that Chao1 only offers a lower bound is related to the fact that it does not take into account any heterogeneity in the data [4]: it is derived from the standard Poisson distribution, which assumes homogeneity across the data points. And yet, it is clear that some individuals in our data might have had higher detection rates, just like some species in ecology might be easier to observe in nature (because of their large size, bright color, loud vocalization, etc.) [8]. It has been shown that ignoring the heterogeneity in a dataset, if any were present, will inevitably lead to (lower bound) estimates for the true population size that are excessively conservative [5]. Conversely, if we can account for potential differences in detection probability, this conservative bias in Chao1 can be reduced, and the population estimate is adjusted upwards, consequently becoming less of a lower bound. To test whether there is indeed any such heterogeneity in a dataset, Böhning and colleagues [2, 13] recommend the use of a so-called ratio plot for the function:

$$\hat{r}_x = \frac{(x+1)f_{x+1}}{f_x}, \quad (1)$$

where f_x refers to the number of items that occur exactly x times [2]. If the data points are homogeneous and do not violate the Poisson assumptions, the resulting points should present as a constant, i.e. a horizontal line. If that is not the case, there is reason to assume that there is unobserved heterogeneity. In Fig. 4, we present a ratio plot for the arrest data up to $x = 6$

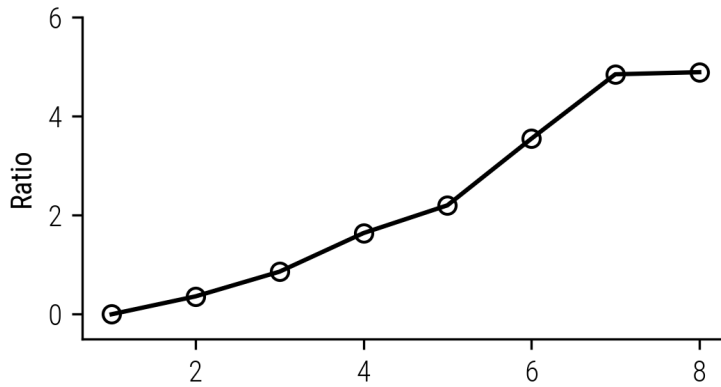


Figure 4: A ratio plot for the count statistics from the arrest data.

from the Amigo prison. It is clear that we are not dealing with a straight horizontal line. Thus, we have reason to believe that there is heterogeneity in the data that is not accounted for by the standard Chao1.

4.2. Generalized Chao

In this section, we discuss the Generalized Chao: a generalization of the Chao1 method that can take into account statistical covariates and thus model heterogeneity across individual perpetrators in a dataset. This estimator, originally proposed by Böhning et al., aims to model heterogeneity in the data by modeling the detection probabilities as a regression problem. The method is characterized by a similar focus on low-frequency species (f_1 and f_2) as Chao1, capturing the intuition that such uncommon species carry the most information about species which were not observed at all. These data consisting of f_1 and f_2 counts can be said to arise from a Poisson distribution, $y_i \sim \text{Poisson}(\lambda)$, but a truncated one with $y_i \in 1, 2$.

Böhning and colleagues show that by maximizing the binomial likelihood, we can obtain an estimate for λ_i using $\hat{\lambda}_i = 2 \frac{\hat{p}_i}{(1-\hat{p}_i)}$ [5]. Here, \hat{p}_i refers to the estimated probability of species i to occur once or twice. To estimate \hat{p} , then, we fit a Generalized Linear Model assuming a binomial distribution:

$$y_i \sim \text{Binomial}(1, p_i)$$

$$\text{logit}(p_i) = \alpha + \beta_x x_i + \dots$$

Here, α represents the intercept and β_x represents the coefficient for a predictor x_i which is available for the i -th species in the data; the outcome variable is binary with the negative class representing species that occur once and the positive class those that occur twice. Given \hat{p}_i , the lower bound on the true population size can be estimated as follows:

$$\hat{S} = S + \sum_{i=1}^{f_1+f_2} \frac{1}{\hat{\lambda}_i + \hat{\lambda}_i^2/2} \quad (2)$$

We employ Bayesian logistic regression models as implemented in the Python package Bambi[6] as an interface to PyMC [1]. The ‘No U-Turn Sampler’ (NUTS) was used for sampling [14], also known as the adaptive Hamiltonian Monte Carlo (HMC) algorithm. We use weakly informative priors for all model terms. The convergence of all models was verified through examination of their trace plots and the \hat{R} statistic [11]. As all \hat{R} values were well below the 1.1 threshold, indicating convergence, we opted not to report individual values. To assess and compare the performance of the regression models on unseen data, we employ approximate Leave-One-Out Cross Validation (LOO-CV) [22].

4.3. Estimating the probability of repeated occurrence

We fit a series of models for diverse (additive) combinations of the available covariates. No interactions or random effects are considered. We include linear and quadratic terms for the age predictor. We compare all models against an intercept-only model which, as it ignores any heterogeneity in the data, should produce estimates equal to the Chao1 model, and can thus be considered a baseline in these experiments. We compare the predictive performance of the models using LOO-CV [22], and visualize the expected log pointwise predictive density (ELPD) and error in the left panel of Fig. 5. The model ranking suggests that the inclusion of predictors generally improves the predictive performance. Akaike model weights are reported in the middle panel: these in turn suggest that the more complex, additive models offer added value over simpler models, but caution is warranted as these weights are determined as point estimates and do not account for uncertainty [20]. In fact, given the considerable overlap in the errors of the ELPD estimates, no single model clearly outperforms the others and should be unequivocally preferred.

In Fig. 6, we show a forest plot for the coefficients included in the highest ranked model, which is also the most complex in terms of parameters. The predictor coefficients’ HDI values do not intersect with zero suggesting that they each contribute meaningfully to the model. Given the strong class imbalance, however (cf. the intercept estimate), their effect sizes are however small, which is both reflected in the non-distinctness of the model comparison above and the marginal effect plots in Fig. 7. The coefficients show how this particular model assesses the likelihood of an individual being a doubleton (as opposed to a singleton), i.e. $p(f_2)$. First of all, the arrest rates vary with gender: we see that female individuals were more likely to be a doubleton than men. There appears to be a non-linear relationship with “age”, which peaks around the age of 50. Doubletons are also slightly common in the local population from Brussels, who suffered higher arrest rates than people born elsewhere in Belgium – the vulnerability to arrest of native Belgians appears comparable to individuals born outside of the country. People who previously stayed the night in the Amigo at their own request also saw a higher doubleton incidence than others. Finally, this model finds weak evidence for a bias against individuals with a known last name, potentially indicating more police effort against members from families perceived as “troublemakers”.

Given the lack of decisive arguments to conclusively prefer one model over the rest, we opt for the common practice of (weighted) model averaging. This approach allows us to account for model uncertainty by combining the predictions of multiple models, rather than relying solely on a single ‘best’ model. Specifically, we employ the stacking of predictive distributions

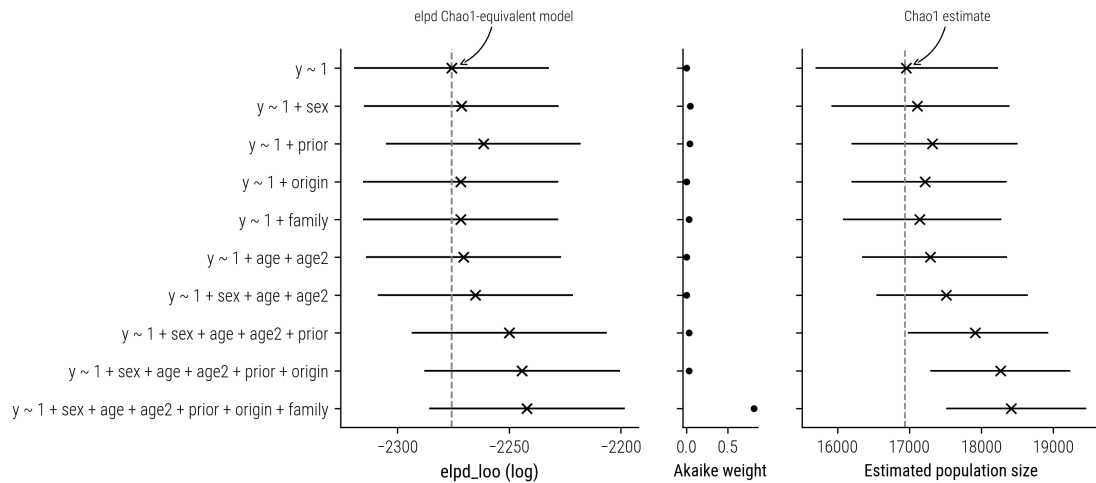


Figure 5: Left panel: Visualization of the model comparison for the LOO information criterion (ELPD = expected log pointwise predictive density), including the difference with the baseline intercept model ($y \sim 1$). The intervals for the out-of-sample predictive fit overlap across nearly all models. Middle panel: Akaike weights for each model. Right panel: Adjusted population size estimate using the Generalized Chao. Only for the best performing additive models, we see that the lower end of the credible intervals does not overlap with the point estimate for Chao1.

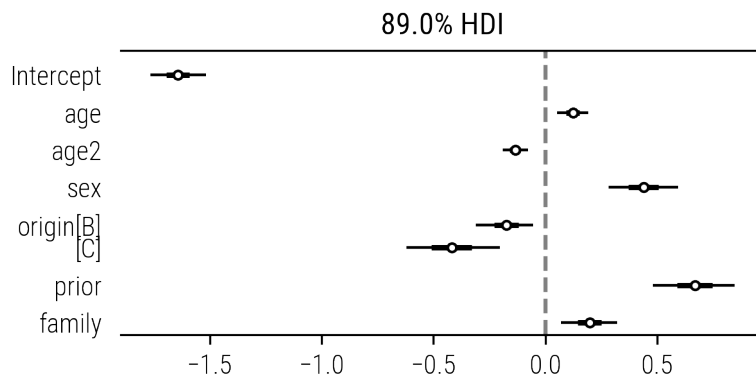


Figure 6: Forest plot visualization for the model coefficients of the best performing individual model, excluding the intercept (with 89% HDIs). None of the shown ranges intersect with zero, although many of the effects are relatively small. (Origin[B] refers to 'BE' in relation to the reference level 'BXL' and Origin[C] to 'ABROAD'.)

method [25], as implemented in PyMC. This technique generates a meta or ensemble model by computing a weighted average of the posterior predictions from all considered models. The weights assigned to each model are derived from the model weights in our previous model comparison analysis (cf. the middle panel of Fig. 5).

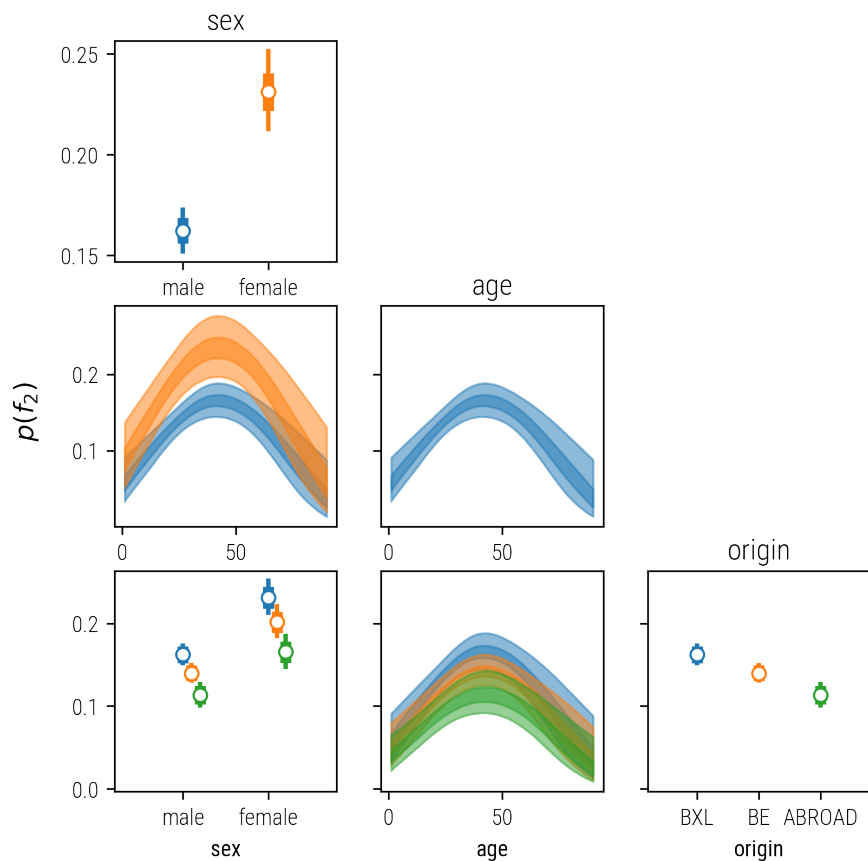


Figure 7: Marginal effect plots contrasting the predictors 'sex', 'age', and 'origin'. Shown are the 50% and 89% credible intervals. The top row displays the marginal effect of 'sex', with females showing a higher effect than males. The middle row contrasts 'age' effects, revealing a non-linear relationship peaking around age 50, with separate curves for males and females. The bottom row illustrates the interaction between 'sex', 'age', and 'origin', where 'BXL' (Brussels), 'BE' (Belgium), and 'ABROAD' categories are compared.

4.4. Adjusted population estimates

The main results are presented in Table 2, which is primarily insightful for the categorical predictors. Here, the adjusted population size estimates are shown with 89% upper and lower credible intervals, both for the highest ranked model and the ensemble model, together with the observed counts and the original Chao1 estimate for each predictor level. Detection rates are presented for each predictor. Asterisks in the table indicate that the lower end of the credible interval for a level does not include the corresponding Chao1 estimate: this is of particular interest, because these are instances where the novel estimators deviate meaningfully from the baseline estimator, identifying demographic categories where the detection rate is potentially underestimated by Chao1. Crucially, we see that the Generalized Chao estimates consistently yield an upwards adjustment for the total population in comparison to Chao1: as the estimated

Table 2

Tabular overview of the Generalized Chao estimates (best performing model (\hat{S} , and model averaging \hat{S}_{MA}) across the different demographic levels in the data, including per-level observed counts and the original Chao1. Asterisks in the table indicate that the lower end of the credible intervals for a level does not include the corresponding Chao1 estimate.

		\hat{S}_l	\hat{S}_u	S	\hat{S}	\hat{S}_{MA}	Chao1	ratio
Pop. total		17409	19505	5610	18414*	18238	16937	0.305
origin	BXL	7160	8202	2582	7660*	7690*	7148	0.337
	BE	6940	8154	2275	7530	7457	7029	0.302
	ABROAD	2759	3783	753	3223	3093	2982	0.234
sex	male	15042	16993	4649	15969*	15779	14818	0.291
	female	2193	2730	961	2444	2460	2272	0.393
prior	no	16364	18458	5100	17363*	17155	16299	0.294
	yes	942	1174	510	1050	1084	1007	0.486
family	no	10708	12407	3237	11536*	11312	10688	0.281
	yes	6403	7401	2373	6877	6927	6409	0.345

population total went up from 16,937 to 18,414, our estimated detection rate dwindled down from 33%, to 31%. While the adjustments can be limited in some instances, this suggests that the Generalized Chao has successfully managed to capture at least some of the heterogeneity across the demographic group in the data. As expected, the model averaging yields slightly lower estimates than the single-model approach and presents a more cautious approach that might be preferable in the case of ranking ties between models, like here.

When we compare the detection rates across the levels of specific factors, interesting trends can be discerned. These insights are further supported by the marginal effects plots in Fig. 7. The estimated number of female perpetrators is much lower than for male perpetrators, but we see that women's vulnerability to arrest was considerably higher. In the case of migratory status, we see that local citizens form the largest group, but this time we see that the arrest efforts clearly targeted this group more intensely than people born outside of the city or country. Again, no bias against migrants is manifest in the data. A factor that considerably increased detect rates was the stay of a prior night in the Amigo. Somewhat weaker is the effect of belonging to a family of perceived troublemakers: again, a smaller demographic group, but with much more policing effort geared against them. Often, we see that the effects are strongest for the smallest, most underrepresented subgroups in the criminalized population, such as women, which are heavily outnumbered by men. This might help explain why even the best performing model in the evaluation in the end does not clearly outperform the baseline model.

5. Discussion and model criticism

Our application of the Generalized Chao method to the Amigo prison register data has yielded valuable insights into the 'dark number' of criminality in late 19th-century Brussels. By in-

corporating demographic covariates into our model, we've been able to estimate not only the total number of unobserved perpetrators but also to characterize patterns of arrest vulnerability across different social groups. Our finding suggests that the true criminal population was likely much larger than can be observed in the register data, and that the Generalized Chao, albeit slightly, adjust the population estimate relative to the Chao1 model.

However, it's crucial to acknowledge the limitations of our approach. Our statistical models primarily focus on relatively stable individual characteristics and do not incorporate event-specific information such as the date, location, or reason for arrest. The inclusion of age as a predictor, while informative, is only justified by the narrow one-year time frame of our dataset. In ecological terms, our model considers species-level covariates but lacks observation-level predictors. This constraint represents a significant limitation of the method and highlights an area for future methodological development.

Despite these limitations, our rigorous model comparison revealed valuable insights. Distinguishing between singletons and doubletons proved to be a complex task, with considerable model uncertainty. While our models incorporating demographic covariates didn't markedly outperform the baseline model in terms of out-of-sample predictive ability, this finding itself is valuable. It underscores the complexity of historical criminality and arrest patterns, suggesting that simple demographic factors alone may not fully explain arrest vulnerability.

While our model comparison revealed limitations in predictive power, it's crucial to consider additional factors that may influence our results. One such factor is the potential violation of core assumptions underlying the Chao1 estimator and its generalizations, particularly the assumption of a closed population during the observation period. In many applications, including ecology, this assumption is often violated due to births and deaths within the population. However, in our historical urban context, migration presents a more significant concern. Late 19th-century Brussels experienced extreme population turnover, with historians estimating annual rates as high as 10% [9]. This dynamic population flux challenges the closed population assumption inherent in our models. Furthermore, the specific nature of our data introduces additional complexities. Some arrested vagrants were temporarily sent to vagrancy colonies outside Brussels. While many of these individuals eventually returned to the city, they were temporarily unobservable [9]. This periodic absence of individuals from the observable population further complicates our modeling efforts.

These factors likely contributed to one-inflation in our data, an excess of singletons (individuals observed only once) compared to doubletons (individuals observed twice). This phenomenon requires careful consideration, as standard models, including non-parametric lower bound estimators, tend to overestimate the true population size in such cases [3]. Several mechanisms can contribute to one-inflation in a population:

1. **Behavioral change:** Identified individuals may alter their behavior after initial detection. In our context, the experience of incarceration might have had a strong dissuasive effect, reducing the likelihood of repeat offenses and contributing to one-inflation;
2. **Mismatching:** Individuals actually detected twice might be incorrectly recorded as single detections, artificially inflating the singleton category. Given the nature of our historical data, such mismatches are plausible. For instance, an intoxicated arrestee with impaired speech might have been difficult for police to accurately identify and record.

Population turnover: As discussed earlier, the high rate of population flux in late 19th-century Brussels could contribute to one-inflation by introducing many individuals who were only briefly present in the city.

3. **Temporary absences:** The practice of sending some arrested vagrants to colonies outside Brussels, as previously mentioned, could also contribute to one-inflation by interrupting the observability of these individuals.

The presence of one-inflation in our data may partially explain the limited improvement in predictive power we observed when incorporating demographic covariates into our models. It also underscores the importance of developing and applying methods that can account for such data characteristics in historical criminology research.

Looking beyond the specific challenges of our study, this work has significant implications for the broader field of computational humanities, particularly from the perspective of critical data studies. A persistent challenge in historical scholarship is the under-representation of marginalized groups in archival records. This systemic bias not only skews our understanding of historical societies but also perpetuates the silencing of under-represented voices. Our work with the Generalized Chao estimator offers a concrete, quantitative approach to addressing this issue. By estimating the ‘dark numbers’ and characterizing demographic patterns in arrest vulnerability, we provide a method for: (a) quantifying the extent of under-representation in historical records, (b) identifying specific demographic groups that may be disproportionately under-represented, and (c) adjusting historical narratives to account for these ‘invisible’ populations.

While our study focused on historical criminology, the methodology has potential applications across various domains of historical research. It could be adapted to estimate under-representation in census data, literary corpora, or other historical datasets where certain groups may be systematically excluded or under-reported. Clearly, statistical methods alone cannot fully rectify historical biases. They must be used in conjunction with critical historical analysis, interdisciplinary collaboration, and a commitment to amplifying marginalized voices. Future research should explore how unseen species models can be integrated with other quantitative and qualitative methods in the humanities to provide a more comprehensive and equitable view of historical societies.

References

- [1] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. Luhmann, O. Martin, M. Osthege, R. Vieira, T. Wiecki, and R. Zinkov. “PyMC: a modern, and comprehensive probabilistic programming framework in Python”. In: *PeerJ Computer Science* 9 (2023), e1516. DOI: 10.7717/peerj-cs.1516. URL: <https://doi.org/10.7717/peerj-cs.1516>.
- [2] D. Boehning, M. F. Baksh, R. Lerdsuwansri, and J. Gallagher. “Use of the ratio plot in capture-recapture estimation”. In: *Journal of Computational and Graphical Statistics* 22.1 (2013), pp. 135–155. DOI: 10.1080/10618600.2011.647174. URL: <https://centaur.reading.ac.uk/24755/>.

- [3] D. Böhning and H. Friedl. “One-Inflation and Zero-Truncation Count Data Modelling Revisited With a View on Horvitz–Thompson Estimation of Population Size”. In: *International Statistical Review* (2024). DOI: <https://doi.org/10.1111/insr.12570>.
- [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), pp. 595–610.
- [5] D. Böhning, A. Vidal-Diez, R. Lerdswansri, C. Viwatwongkasem, and M. Arnold. “A Generalization of Chao’s Estimator for Covariate Information”. In: *Biometrics* 69.4 (2013), pp. 1033–1042.
- [6] T. Capretto, C. Piho, R. Kumar, J. Westfall, T. Yarkoni, and O. A. Martin. “Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python”. In: *Journal of Statistical Software* 103.15 (2022), pp. 1–29. DOI: [10.18637/jss.v103.i15](https://doi.org/10.18637/jss.v103.i15).
- [7] A. Chao. “Nonparametric estimation of the number of classes in a population”. In: *Scandinavian Journal of Statistics* 11 (1984), pp. 265–270.
- [8] A. Chao and C. H. Chiu. “Estimation of species richness and shared species richness”. In: *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*. Ed. by N. Balakrishnan. Wiley, 2012, pp. 76–111.
- [9] M. De Koster and A. Erkul. “Removing Local Nuisances, Arresting Masterless Strangers, and Granting ‘Nights on Request’: The Policing of Vagrancy in Late-Nineteenth-Century Antwerp and Brussels”. In: *TSEG - The Low Countries Journal of Social and Economic History* 20.1 (2023). DOI: [10.52024/tseg.13642](https://doi.org/10.52024/tseg.13642).
- [10] A. Esch. “Überlieferungs-chance Und Überlieferungs-zufall Als Methodisches Problem Des Historikers”. In: *Historische Zeitschrift* 240.3 (1985), pp. 529–570. DOI: [doi:10.1524/hzhz.1985.240.3.529](https://doi.org/10.1524/hzhz.1985.240.3.529).
- [11] A. Gelman and D. B. Rubin. “Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: [10.1214/ss/1177011136](https://doi.org/10.1214/ss/1177011136).
- [12] N. J. Gotelli and R. K. Colwell. “Estimating species richness”. In: *Biological Diversity: Frontiers in Measurement and Assessment*. Ed. by A. E. Magurran and B. J. McGill. Oxford University Press, 2011, pp. 39–54.
- [13] P. G. M. van der Heijden, M. Cruyff, and D. Böhning. “Capture Recapture to Estimate Criminal Populations”. In: *Encyclopedia of Criminology and Criminal Justice*. New York, NY: Springer New York, 2014, pp. 267–276. DOI: [10.1007/978-1-4614-5690-2_662](https://doi.org/10.1007/978-1-4614-5690-2_662).
- [14] M. D. Hoffman and A. Gelman. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1593–1623.
- [15] M. Kestemont, F. Karsdorp, E. de Bruijn, M. Driscoll, K. A. Kapitan, P. O. 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](https://doi.org/10.1126/science.abl7655).

- [16] N. Löhr. “Off the Record: On Studying Lost Arabic Books and their Networks”. In: *Medieval Worlds* 18 (2023). DOI: 10.1553/medievalworlds_no18_2023s219.
- [17] M. Mangel and F. J. Samaniego. “Abraham Wald’s Work on Aircraft Survivability”. In: *Journal of the American Statistical Association* 79.386 (1984), pp. 259–267.
- [18] A. Martynenko. “Unread, yet preserved: A case study on survival of the 19th-century printed poetry”. In: *Literatura: teoría, historia, crítica* 25.2 (2023).
- [19] M. Maxfield, B. Weiler, and C. Widom. “Comparing Self-Reports and Official Records of Arrests”. In: *Journal of Quantitative Criminology* 16 (2000), pp. 87–110. DOI: 10.1023/a:1007577512038.
- [20] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall: CRC Press, 2020.
- [21] T. L. Penney. “Dark Figure of Crime (Problems of Estimation)”. In: *The Encyclopedia of Criminology and Criminal Justice*. John Wiley & Sons, Ltd, 2014, pp. 1–6. DOI: 10.1002/9781118517383.wbeccj248.
- [22] A. Vehtari, A. Gelman, and J. Gabry. “Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC”. In: *Statistics and Computing* 27.5 (2017), pp. 1413–1432. DOI: 10.1007/s11222-016-9696-4.
- [23] H. Wainer. *Visual revelations: graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Lawrence Erlbaum Associates, 2000.
- [24] M. Wevers, F. Karsdorp, and J. van Lottum. “What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model”. In: *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*. Ed. by F. Karsdorp and K. L. Nielbo. Vol. 3290. CEUR Workshop Proceedings. CEUR-WS.org, 2022, pp. 189–197. URL: <https://ceur-ws.org/Vol-3290/short%5C%5Fpaper1793.pdf>.
- [25] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. “Using Stacking to Average Bayesian Predictive Distributions (with Discussion)”. In: *Bayesian Analysis* 13.3 (2018). DOI: 10.1214/17-ba1091.

A. Online Resources

The full code and data for replicating our analysis will be made available from the following GitHub repository upon the publication of the paper: <https://doi.org/10.5281/zenodo.13969373>.