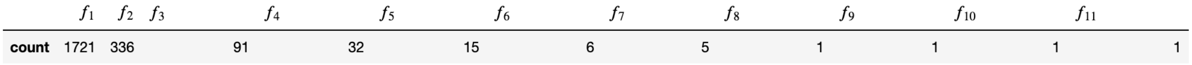
Dark Numbers. Modeling the historical vulnerability to arrest in Brussels (1879-1880) using demographic predictors

Folgert Karsdorp, Meertens Institute, Royal Netherlands Academy of Arts and Sciences
Margo De Koster, Ghent University
Mike Kestemont, University of Antwerp

Capture-recapture surveys are important bioregistration instruments in ecology, used to monitor aspects of biodiversity, such as species richness. During such campaigns, field workers use 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), doubletons (f_2), etc. Because of the imperfect observation process, however, many animal types will not be observed, 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 n (the number of observed species). Chao1, for instance, is a widely used estimator that estimates a lower bound on f_0 as follows: $\hat{f}_0 = f_1^2 / 2f_2$.

Recently, it has been demonstrated that such "unseen species models" can also be meaningfully applied to historical datasets, involving for instance the survival of medieval literature or the leaky registration of sailors working on early modern ships: because of imperfections in the persistence of such data, a similar observation bias presents itself in these sources, which can be partially corrected using these methods. So far, these models have been useful to estimate the magnitude of a loss phenomenon, but failed to explain the drivers of that loss: consequently, we often still have limited insight into the composition of the unobserved share of a population, a situation tightly related to the phenomenon of survivorship bias. Researchers in statistical sociology have broken new ground in this domain through the introduction of unseen species models that can include species-level covariates as predictors. By building on the so-called Generalized Chao estimator proposed by Bohning et al., the problem of population size estimation can be casted as a regression problem. The proposed method is characterized by a similar focus on low-frequency species (f_1 and f_2), capturing the intuition that such uncommon species carry the most information about species which were not observed at all. The method fits a binomial classifier to the singletons and doubletons in the data to model the probability that a species is sighted (exactly) twice, instead of (exactly) ones, resulting from a Bernoulli process (with a logit link function).

This method has found convincing applications already in contemporary criminology, where unobserved crime (so-called "dark numbers") are a factor of major interest. Domestic violence is a classic example in this domain, with strong evidence for under-reporting by victims. Here, we transfer this method to the field of historical studies, more specifically the register of the Amigo prison in Brussels in the period 1879-1880. The dataset includes demographic characteristics, such as age, gender and birth place, for individuals who forcefully entered the prison after being arrested. (Interestingly, this data is complemented with the registration of individuals who entered the prison for a night upon their own request (*une nuit sur demande*), e.g. because they did not have a place to stay.) Most individuals in the dataset enter the prison only rarely (see the count distribution below), but there is nevertheless much recidivism, with individuals being sighted on multiple occasions. The short time period covered is an advantage, because it allows us to treat the perpetrators' age as a species-level and not an observation-level covariate. Additionally, this will make the population less susceptible to change, because the model theoretically assumes a closed population (which ours is not entirely). Overall, individuals were not imprisoned for a lengthy time, which enables their resighting.

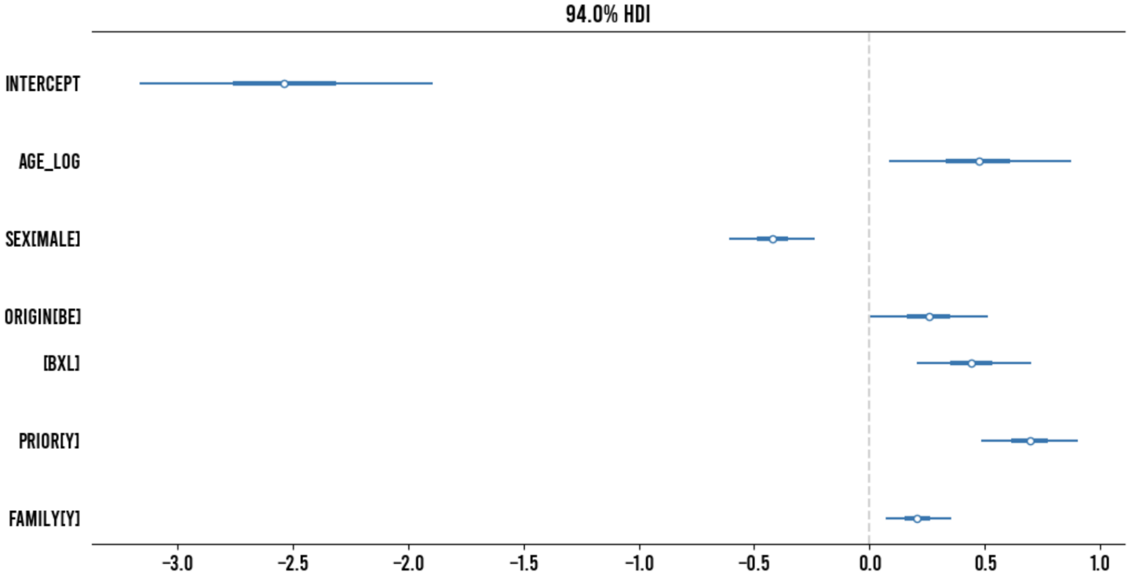


	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
count	1721	336	91	32	15	6	5	1	1	1	1

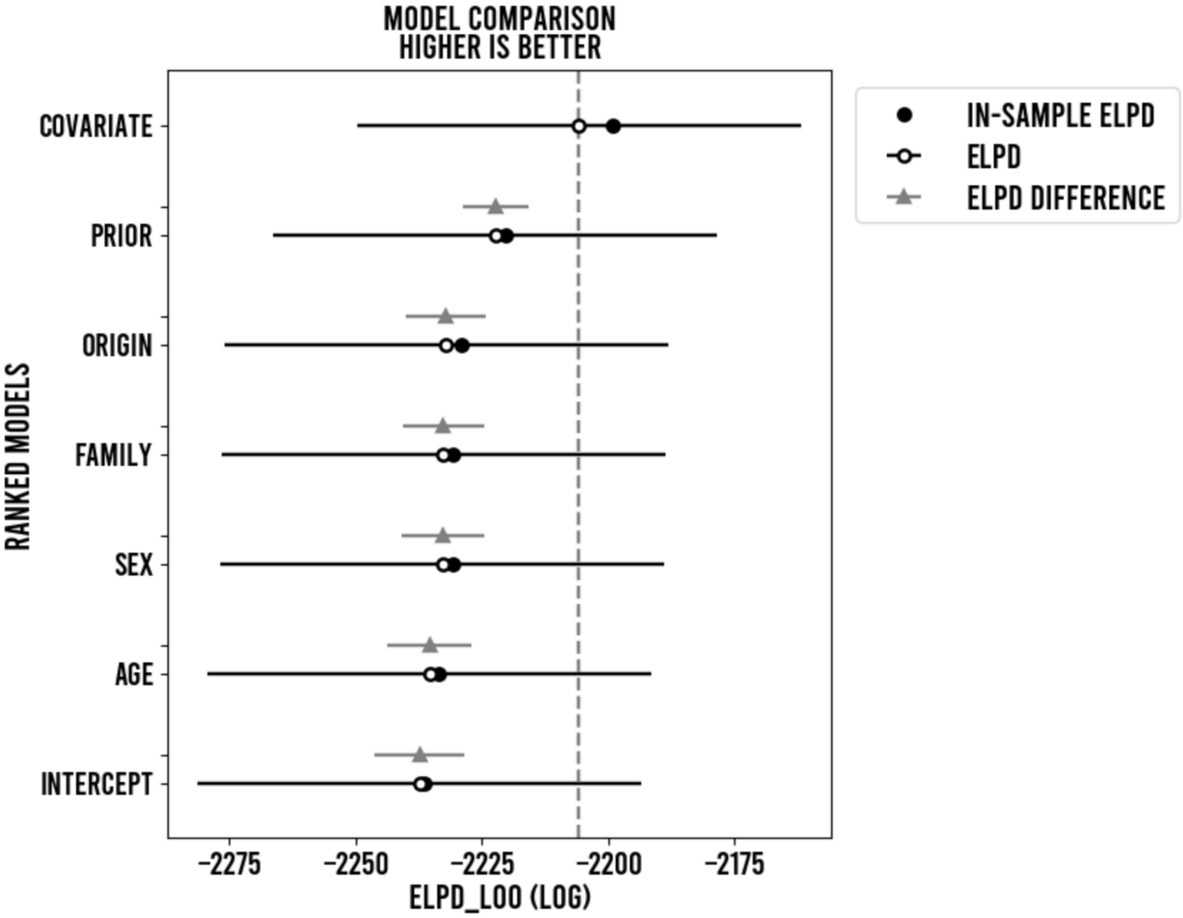
Prior research has already established that policing effort regarding arrests could be biased and demographic groups suffered differential arrest rates. We aim to apply the Generalized Chao estimator described above with a twofold aim: (1) estimate the number of non-apprehended individuals (i.e. the "dark number" of perpetrators who were not arrested); (2) model the vulnerability to arrest across different categories of perpetrators. In our model, we resort to the following predictors:

- *log(age)*: an individual's mean age at the time of arrest (scalar);
- *sex*: an individual's sex (binary: 'female' (reference level) or 'male');
- *origin*: manually coded factor on the basis the individual's place of birth ('ABROAD' (reference level), 'BE', 'BXL');
- *name*: binary indicator ('N' (reference level) and 'Y') whether an individual with the same family name occurs in the dataset; this variable aims to capture bias against known "criminal families";
- *prior*: binary indicator ('N' (reference level) and 'Y') whether the individual was granted at least one "night on request" prior to their first arrest; this variable is included because individuals who were previously known with the police might be more vulnerable to arrest.

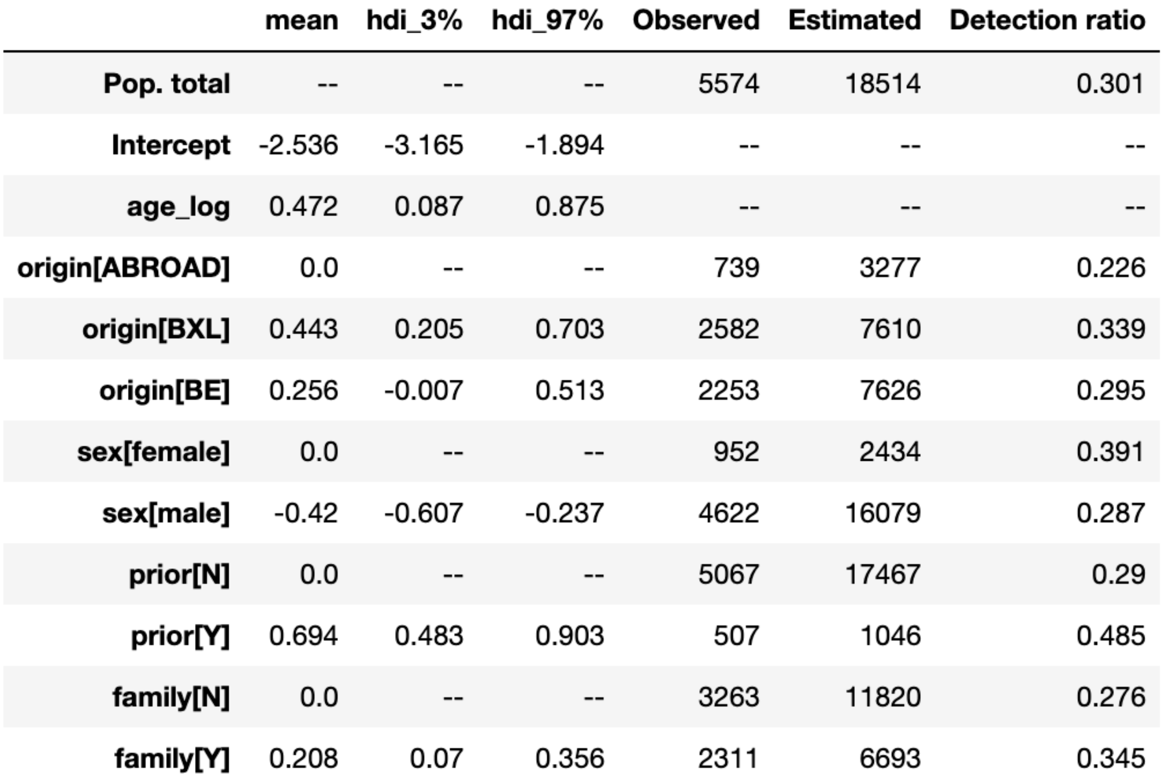
We apply the model in a Bayesian framework, using the BAMBI library in Python as an interface to Stan. We report an intercept-only model; a model for each predictor in isolation; and an additive model that includes all covariates. The figure below shows a forest plot of the Highest Density Interval (HDI; on the horizontal axis) for the predictor coefficients in the latter model: the HDI's generally do not intersect with zero, indicating that they all contribute meaningfully to the model. Females are generally more likely to end up as f_2's than men; age contributes positively to the vulnerability to arrest too. Local perpetrators were more easily arrested than other non-local Belgian nationals or foreigners. A prior *nuit sur demande* also makes the resighting of an individual more likely; likewise, we find weak evidence for a bias against individuals with a "known last name", potentially indicating more police effort against members from families perceived as "criminals" and "troublemakers".



We also compared the individual models using the leave-one-out method that is recommended in Bayesian modeling: while the covariate model receives most of the "weight" (i.e. probability given the data), the models all have large standard errors, leading to considerable overlap, making it difficult to distinguish between them.



The resulting model probabilities can be used to estimate the number of observed individuals in each group, but also the absolute number of unobserved perpetrators – see the estimated numbers in the overview table below. The detection rates can differ strongly across categories.



	mean	hdi_3%	hdi_97%	Observed	Estimated	Detection ratio
Pop. total	--	--	--	5574	18514	0.301
Intercept	-2.536	-3.165	-1.894	--	--	--
age_log	0.472	0.087	0.875	--	--	--
origin[ABROAD]	0.0	--	--	739	3277	0.226
origin[BXL]	0.443	0.205	0.703	2582	7610	0.339
origin[BE]	0.256	-0.007	0.513	2253	7626	0.295
sex[female]	0.0	--	--	952	2434	0.391
sex[male]	-0.42	-0.607	-0.237	4622	16079	0.287
prior[N]	0.0	--	--	5067	17467	0.29
prior[Y]	0.694	0.483	0.903	507	1046	0.485
family[N]	0.0	--	--	3263	11820	0.276
family[Y]	0.208	0.07	0.356	2311	6693	0.345

Whereas the population of perpetrating women, for instance, was much smaller than men, we see that (older, local) females were much more vulnerable to arrest; the same is true for Belgian nationals, especially Brussels locals, to whom much policing effort was geared. No meaningful bias can be discerned against foreigners. Our results tie in closely with prior research and demonstrate the usefulness of this method for studying biased observation processes in history, for instance regarding dark numbers in the light of policing effort. The surplus value of the Generalized Chao method may be limited in the case of a dataset with only a few categorical predictors and well-balanced levels: in such cases, one could in principle simply filter out the relevant subsets and calculate $\hat{f}_0$ for the groups in isolation using the conventional Chao1. This rapidly becomes less feasible however for many predictors with many ill-balanced levels, because the resulting subsets will become prohibitively small for a reliable estimate. A point of attention is that the difference between singletons and doubletons often proves hard to model in historical data. Future research should consider the use of simulations to establish the effect of such under-modeled heterogeneity on the resulting estimates. Finally, the paper will highlight some unresolved aspects of the model's assumptions, for instance, whether we can assume a reasonably closed population, which is problematic for a metropole like Brussels.

References

- Böhning, D. et al., ‘A Generalization of Chao’s Estimator for Covariate Information’, *Biometrics* (2013).
- Böhning, D. & Van der Heijden, P., ‘A Covariate Adjustment for Zero-Truncated Approaches to Estimating the Size of Hidden and Elusive Populations’, *The Annals of Applied Statistics* (2009).
- Chao, A., ‘Nonparametric Estimation of the Number of Classes in a Population’, *Scandinavian Journal of Statistics* (1984).
- De Koster, M. & Erkul, A., ‘Removing local nuisances, arresting masterless strangers, and granting ‘nights on request’. The policing of vagrancy in late-nineteenth-century Antwerp and Brussels’, *TSEG - The Low Countries Journal of Social and Economic History* (2023).
- Karsdorp, F., ‘Population Size Estimation as a Regression Problem’, https://www.karsdorp.io/posts/20220405110456-population_size_regression_estimators/ (2022).
- Kestemont M., Karsdorp F. et al., ‘Forgotten books: the application of unseen species models to the survival of culture’, *Science* (2022).
- Wevers, M.; Karsdorp, F. & Van Lottum, J., ‘What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model’, in: *CHR 2022: Computational Humanities Research Conference* (2022).