

Online Appendix: Explaining Recruitment to Extremism: A Bayesian Hierarchical Case-Control Approach

Roberto Cerina¹, Christopher Barrie², Neil Ketchley³, Aaron Y. Zelin⁴

¹University of Amsterdam

²University of Edinburgh

³University of Oxford

⁴Brandeis University

Supplementary Materials

Table of Contents

A Sources	1
A.1 Independent variable details	3
B MCMC Convergence	6
C Bayesian modeling in Stan	7
C.1 Stan listings	7
D Simulation study	10
D.1 Data Generating Function	10

D.2	Candidate Models	12
D.3	Computational Constraints	12
D.4	Comparison Metrics	13
D.5	Results	13
E	Practical Advice for Researchers	18
F	Convergence diagnostics	22
G	Posterior densities of regression coefficients	27
H	Relative Deprivation Effects	32
I	Residual Area-Level Analysis	35

A Sources

It is important to consider the data-generating process for the leaked ISIS border documents. These documents contain detailed information on the home residence of each recruit, age, education, marital status, previous employment, employment status, previous combat experience, and date of entry into ISIS-controlled territory. They derive from a set of leaked documents recording the details of fighters who have crossed into ISIS-controlled territory with the intention of becoming a recruit.

Supplementary Figure [A.1](#) provides an imitation of one of the border documents. We use data for nine countries in the MENA that were included in the leak. These are: Algeria, Egypt, Jordan, Kuwait, Lebanon, Libya, Morocco, Tunisia, and Yemen. In total, we have complete records for 1,051 recruits. It remains unclear whether these constitute a representative sample of recruits. [Dodwell et al. \(2016\)](#) demonstrate, however, that 98% of these individuals can be matched against records for ISIS recruits held by the U.S. Department of Defense. Further, the Bayesian case-control approach we detail below takes into account the non-probability nature of the data-generating process through its multilevel design.

الإدارة
العامة
للحدود

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

الدولة الإسلامية في العراق والشام

الإدارة العامة للحدود

بيانات مجاهد



1	Forename and surname	Abdul Karim al-Fadl
2	Nom de guerre	Abu Hamza al-Masri
3	Mother's name	Layla
4	Blood type	A
5	Date of birth	11/01/1991
6	Marital status	<input type="checkbox"/> Married <input checked="" type="checkbox"/> Single
7	Place of Residence	Cairo, Doki
8	Education level	Bachelors in Engineering
9	Level of Sharia	<input type="checkbox"/> Low <input checked="" type="checkbox"/> Medium <input type="checkbox"/> High
10	Occupation prior to arrival	Unemployed
11	Countries transited	None
12	Point of entry and contact	Jarablus, Abu Abdi
13	Who recommended	Abu Abdi
14	Date of entry	01/09/2013
15	Previous combat experience	None
16	Fighter; Martyr; Suicide Bomber?	
17	Prefered specialization	<input type="checkbox"/> Admin <input type="checkbox"/> Security <input type="checkbox"/> Shara'i <input checked="" type="checkbox"/> Fighter
18	Current place of work	
19	Items of luggage	Suitcase
20	Level of hearing	
21	Phone number and emergency contact	Wife 123456789 Father 876543210
22	Date and place of death	
23	Notes	

الإدارة العامة للحدود

الدولة الإسلامية في العراق والشام _ سري _




Figure A.1: Example of border document (details changed)

A.1 Independent variable details

The Arab Barometer surveys were in the field at different times for each country: December, 2012-January, 2013 for Jordan; February, 2013 for Tunisia; March-April, 2013 for Egypt and Algeria; April-June, 2013 in Morocco; July 2013 in Lebanon; November-December 2013 in Yemen; and February-March 2014 in Kuwait ([ArabBarometer 2014](#)).

For both Egypt and Tunisia, we also include variables to capture subnational differences in demographic and labor-market composition, employment opportunities, as well as more context-specific variables designed to capture support for Islamist political organizations and prehistories of contentious politics. Our choice of contextual variables is based on existing research finding that lack of employment opportunities, prehistories of mobilization and repression, as well as support for political Islam, are predictive of ISIS recruitment ([Devarajan et al. 2016](#); [Rosenblatt 2018](#); [Grewal et al. 2020](#); [Barrie and Ketchley 2018](#)).

Table A.1: Individual-level variable codings across border documents and survey data

Variable	Border Documents	ABIII
coedu	1 if Education level mentions “university”	1 if q1003 >5 (or >4 for Tunisia; >6 for Yemen)
age	Date of entry - Date of birth	q1001
married	1 if Marital status is “married”	q1010
student	1 if Occupation prior to arrival is “student”	q1004 = 3 (Student)
lowstat	1 if Occupation prior to arrival is agricultural or manual/unemployed	q1004 = 5 (Unemployed) or q1010 = 4/5 (Agricultural or manual worker)

Table A.2: Egypt district-level covariates

Variable	Details	Source
Population density	number of individuals in district/district area in km ²	2006 Census
Population	number of individuals in district aged 10 or over	2006 Census
% Christian	percentage of individuals in district recorded as Christian	2006 Census
% College Edu.	percentage of individuals in district who are university educated	2006 Census
% Agriculture	percentage individuals employed in agriculture denominated by total active population	2006 Census
% Mursi	percent of total votes in district for Muhammad Mursi in the first round of the 2012 presidential election	El-Masry and Ketchley (2021)
Unemployment rate	number individuals aged without employment denominated by total active population	2006 Census
Killed at Rabaa	number of deaths of individuals from district at the 2013 Rabaa Massacre (square-rooted)	Ketchley and Biggs (2017)
Post-revolutionary protest	number of protests recorded in district in 12 months after Jan 25 Revolution (square-rooted)	Barrie and Ketchley (2019)

Table A.3: Tunisia district-level covariates

Variable	Details	Source
Population	number of individuals in district aged 10 or over	2014 Census
Population density	number of individuals in district aged 10 or over/district area in km ²	2014 Census
% College Edu.	percentage population with higher education certificate denominated by total population	2014 Census
% Agriculture	percentage individuals employed in agriculture denominated by total active population aged 15 or over	2014 Census
Unemployment rate	number individuals aged 18-59 without employment denominated by total active population aged 18-59	2014 Census
Graduate unemployment rate	number individuals with higher education certificate without employment denominated by total active population aged 18-59	2014 Census
% Ennahda 2011	percentage of total votes in district for Ennahdha in 2011 election	INS Tunisia
% Ennahdha 2014	percentage of total votes in district for Ennahdha in 2014 election	INS Tunisia
Post-revolutionary protests	number of protests recorded in district in 12 months after Jan 14 Revolution (square-rooted)	Barrie and Ketchley (2019)
Distance to Libya	distance to Libyan border from centroid of target district (square-rooted)	NA

B MCMC Convergence

Convergence diagnostics provide a first measure of the reliability of our parameter estimates for both the Bird’s Eye and Worm’s Eye models. Here, we follow Vehtari et al (2021) and implement multiple state-of-the-art tests.

Per Supplementary Figures F.1, F.5, and F.7, we examine four versions of the Gelman-Rubin statistic (\hat{R}) to verify convergence is obtained broadly, as well as when we encounter heteroskedasticity across chains, or when these are heavy-tailed. There exist various convergence-thresholds in the literature – the most stringent requires $\hat{R} < 1.01$, a medium-stringency threshold suggests $\hat{R} < 1.05$ (especially if we are estimating a large number of parameters), whilst the historical recommendation was $\hat{R} < 1.1$ (Gelman and Rubin 1992). Recent work demonstrates that this latter threshold prematurely diagnoses convergence in most cases (Vats and Knudson 2021). The parameters of all of our models are broadly convergent under the harshest 1.01 threshold for all of the measures of \hat{R} , with the exception of a very small number of spatial effects which are convergent under a slightly more lax threshold, though still well below the ‘premature convergence’ threshold $\hat{R} < 1.1$.⁸

Supplementary Figures F.2, F.6, and F.8 present five measures of Effective Sample Size (ESS), which tell us about the true number of independent draws from the joint posterior distribution after accounting for auto-correlation within chains. The measures check that the independent sample is ‘large enough’ to ensure stability of summaries of the distribution at various moments (e.g. overall, at the median, at the tails, etc.). Mirroring the performance of the \hat{R} , the posterior samples for most of our estimates parameters are well above the recommended threshold ($ESS > 400$) for ensuring stability of the central and tail estimates.

We further explore convergence at different quantiles of the posterior distribution of our least-convergent parameters – those with the lowest bulk and tail ESS (Figure F.3 presents these measures for the Bird’s Eye model). The inference is that if these relatively low-ESS parameters showcase satisfactory ESS at every quantile, we can be reassured that the whole model has converged. These plots suggest broad reliability of estimates at every section of the distribution. Finally, we explore the mixing properties of our chains for these least-convergent parameters (Figure F.4).⁹ These plots broadly suggest good mixing properties of our model, even for these relatively inefficient posterior samples.

⁸Note that in the Bird’s Eye model, this struggle is slightly exacerbated by the inclusion of governorates from Israel and Saud Arabia, for which we have no observations, and whose effects are fully interpolated via the spatial process.

⁹Here, we choose to include mixing diagnostics for the Bird’s Eye model.

C Bayesian modeling in Stan

The model that we propose is amenable to Bayesian estimation via Monte Carlo Markov Chain (MCMC) methods. Previous contributions to the case-control literature [e.g., Rota et al. 2013; Rosenfeld 2017] have used WinBUGS (Lunn et al. 2000) or JAGS (Plummer et al. 2003) as software to implement some variations on a simple Gibbs sampler. Due to the heavy computational burden imposed by the spatial prior, we propose instead to innovate by estimating this model in Stan (Carpenter et al. 2017). Stan leverages a version of Hamiltonian Monte-Carlo (HMC) called the ‘No U-Turn Sampler’ (NUTS) (Hoffman and Gelman 2014), which dramatically improves the efficiency and speed of convergence of our Markov-Chains. A challenge we face is that Stan cannot handle the sampling of latent discrete parameters (r_i in our hierarchical model above), posing a problem for the estimation of mixture models. The state-of-the-art solution is to marginalize the latent parameter out. In practice this means replacing our model for the observed labels y with the following mixture of Bernoulli distributions:

$$f(y_i | \rho_i) = \rho_i \text{Bernoulli}(y_i | \theta_1) + (1 - \rho_i) \text{Bernoulli}(y_i | \theta_0). \quad (39)$$

Beyond allowing for model parameters to be informed by y_i according to the mixed structure above, marginalization provides significant advantages for posterior exploration and MCMC efficiency as it leverages expectations rather than sampling of discrete parameters. Listing 1 in the Supplementary Materials presents the Stan code for our final model. Note that fixed-effects covariates are standardized.¹⁰ Estimates of regression coefficients on the original, unstandardised scale are computed and available in these Supplementary Materials.

C.1 Stan listings

Listing 1: Stan Data Declaration Block.

```
1 data{
2
3   int<lower = 1> n;           // total number of observations
4   int<lower = 1> p;           // number of covariates in design matrix
5   int<lower = 0> y[n];       // vector of labels
6   matrix[n, p] X;           // design matrix
7
8   int<lower = 1> small_area_id[n]; // small-area id
9   int<lower = 1> N_small_area;     // number of small areas
10
11  int<lower = 1> N_small_area_edges; // number of edges in the spatial process
12  int<lower=1, upper=N_small_area> node1_small_area[N_small_area_edges]; // node1[i] adjacent to node2[i]
13  int<lower=1, upper=N_small_area> node2_small_area[N_small_area_edges]; // node1[i] adjacent to node2[i]
14
15  real scaling_factor; // scaling factor derived from the adjacency matrix
16
17  int<lower = 1> large_area_id[n]; // large-area ids
18  int<lower = 1> N_large_area;     // number of large-areas
19
20  vector[N_large_area] log_offset; // log-scale offset
21
22  matrix[2,N_large_area] theta; // Pr(Y = 1 | r = 1, s = 1)
23
24 }
```

¹⁰We standardize both dichotomous and continuous variables as this aids convergence.

Listing 2: Stan Parameters Declaration Block.

```
1 parameters{
2
3     // cauchy prior for individual-level coefficients expressed as scale mixture of gaussian density
4     functions
5     vector[p] aux_a;          // central component
6     vector<lower = 0>[p] aux_b; // scale component
7
8     vector[N_small_area] phi; // small-area unstructured effects
9     vector[N_small_area] psi; // small-area spatial effect
10
11     real<lower = 0, upper = 1> lambda; // mixing prior on spatial component
12
13     real<lower = 0> sigma_gamma; // small-area effect scale
14
15     vector[N_large_area] eta; // large-area unstructured effect
16
17     real<lower = 0> sigma_eta; // large-area effect scale
18
19 }
```

Listing 3: Stan Transformed Parameters Block.

```

1 transformed parameters{
2
3     vector[p] beta = aux_a ./ sqrt(aux_b);          // individual-effect prior
4
5     vector[n] mu;          // expected propensity of recruitment
6
7     vector[N_small_area] gamma = (sqrt(1-lambda) * phi + sqrt(lambda / scaling_factor) * psi)*
8         sigma_gamma;
9     // convolved small -area effect
10
11     mu = log_offset[large_area_id] + eta[large_area_id]*sigma_eta + gamma[small_area_id] + X *
12         beta;
13     // linear function of the logit -scale propensity to be a recruit
14 }

```

Listing 4: Stan Model Declaration Block.

```

1 model{
2
3     aux_a ~ normal(0,1);          // prior on the centrality of the cauchy prior
4     aux_b[1] ~ gamma(0.5,100*0.5); // prior on intercept-scale
5     aux_b ~ gamma(0.5,0.5);      // prior on individual covariate scales
6
7     target += -0.5 * dot_self(psi[node1_small_area] - psi[node2_small_area]);
8     // ICAR prior
9
10    phi ~ normal(0,1);           // unstructured random effect on small -area
11    sum(psi) ~ normal(0, 0.01 * N_small_area);
12    // soft sum -to-zero , equivalent to mean(psi) ~ normal (0 ,0.01)
13
14    lambda ~ beta(0.5,0.5);      // mixing weight prior
15
16    sigma_gamma ~ normal(0,1);   // prior small-area scale
17
18    eta ~ normal(0,1);          // prior large-area effect
19
20    sigma_eta ~ normal(0,1);     // prior large-area scale
21
22    // likelihood
23    for (i in 1:n) {
24        target += log_mix(1-inv_logit(mu[i]),
25            bernoulli_lpmf(y[i] | theta[1,large_area_id[i]]),
26            bernoulli_lpmf(y[i] | theta[2,large_area_id[i]]));
27        // labels distributed as mixture of bernoulli distributions
28    }
29 }

```

D Simulation study

The model that we propose extremism researchers should adopt is significantly more complex than the standard case-control design using rare-events logistic regression and requires a substantial understanding of Bayesian methods to be fully appreciated. Moreover, the model’s estimation becomes roughly exponentially more computationally challenging as the sample size increases. To provide evidence that our approach is nevertheless preferable to a more straightforward case-control design, we report the results of a comprehensive simulation study that compares the performance of our model against the King and Zeng model (2001), as well as a simple fixed-effects logistic regression. We score these models according to their ability to accurately predict the underlying latent propensity of recruitment, $\mu_i = \text{logit}(\rho_i)$. We further investigate these models’ performance in accurately estimating the intercept, regression coefficients, and residual area effects.

We note that we did not test the ‘coverage’ properties of our models’ estimates as part of the simulation study. To test coverage, we would have needed to run the chains of each of our simulations long enough for the parameters of our models to converge in their second-moment - this was not feasible under a simulation framework where we had to run the model 200 times. As pointed out in our discussion of model fitting strategy, the well-behaved models we use to derive our results took up to 48 hours to achieve posterior samples displaying satisfactory convergence. We therefore leave it for future work to formally quantify the coverage of our models.

Simulations show that our model is robust and general. The results suggest that in a rare-event scenario, our model outperforms King and Zeng’s rare-events logistic regression thanks to its ability to account for spatial auto-correlation, while also remaining largely unbiased to discrepancies in sample and population prevalence. As prevalence increases, our model retains a degree of robustness that neither a simple fixed-effects logistic regression, nor the ‘rare events logit’, can offer – largely thanks to the contamination layer. This robustness extends not just to the ability to correctly estimate latent propensity μ^* , but actively reduces bias and RMSE in the estimation of coefficients.

D.1 Data Generating Function

In what follows we present a more detailed view of the setup and results of the simulation study. First, we create a data-generating function to draw sample-datasets generated according to the mechanism implied by either the rare-events or contaminated case-control model. We reduce the data generating process to its essence for simplicity: a single continuous covariate x_i is considered, and large-area effects are dropped. Small-area effects are simulated according to a random intrinsic conditionally auto-regressive process from one of three widely-used maps.¹¹, available from the R package `SpatialEpi` (Kim and Wakefield 2010). This enables the random sampling of ICAR effects whilst

¹¹ $\mathcal{M} = \{\text{scotland_lipcancer}, \text{newyork_lukemia}, \text{pennsylvania_lungcancer}\}$

preserving a plausible geography (i.e. neighbourhood structure and distance between areal units). Pseudo-algorithm 1 describes the steps taken to generate the simulated data.

Algorithm 1 A pseudo algorithm displaying the steps taken by the data generating function to generate a random sample of data.

Require:

sample size:	$n \in [100, 2000]$
population prevalence:	$\pi \in \left[\frac{1}{1000000}, \frac{1}{2}\right]$
expected sample prevalence:	$\hat{\pi} \in [0.01, 0.99]$
global auto-correlation:	$I \in (0, 1)$
map:	$\mathcal{M} \in \{\text{scotland, newyork, pennsylvania}\}$

(0.) derive key quantities directly from inputs:

i. expected number of case-labelled records:

$$n_1 \leftarrow n \times \hat{\pi}$$

ii. expected number of unlabelled records:

$$n_u \leftarrow n - n_1$$

iii. relative prob. of sampling a case v. control:

$$\frac{P_1}{P_0} \leftarrow \frac{(n_1 + \pi \times n_u) / \pi}{n_u}$$

iv. prob. of sampling a case-labelled record conditional on being a true control:

$$\theta_0 \leftarrow 0$$

v. prob. of sampling a case-labelled record conditional on being a true case:

$$\theta_1 \leftarrow \frac{n_1}{n_1 + \pi \times n_u}$$

(1.) sample area effects on selected map: $\boldsymbol{\gamma} \sim \text{ICAR}(\mathcal{M})$

(2.) sample initial value for intercept: $\beta_1 \sim N(0, 1)$

(3.) sample covariate value: $x_i \sim N(0, 1)$

(4.) sample covariate effect: $\beta_2 \sim N(0, 1)$

(5.) optimise intercept to meet specified sample prevalence:

$$\beta_1^* \leftarrow \arg \max_{\beta_1} f(\hat{\pi}; \beta_1)$$

(6.) calculate latent recruitment propensity: $\boldsymbol{\mu} \leftarrow \log\left(\frac{P_1}{P_0}\right) + \beta_1^* + \mathbf{x}\beta_2 + \boldsymbol{\gamma}$

(7.) calculate recruitment propensity: $\boldsymbol{\rho} \leftarrow \text{inv_logit}(\boldsymbol{\mu})$

(8.) sample recruitment status: $\mathbf{r} \sim \text{Bernoulli}(\boldsymbol{\rho})$

(9.) sample labels: $\mathbf{y} \sim \text{Bernoulli}(\boldsymbol{\theta}_r)$

We simulate `n.sims` = 200 datasets¹² using the data-generating function. The inputs to the function (highlighted under the ‘Required’ header in the pseudo-code) are

¹²In practice we simulate datasets in two stages: first we examine the model performance by sampling 100 draws from a ‘rare event’ process ($\pi \in [1/1000000, 1/10]$); then, in a second stage, we sample another 100 draws from a process with less extreme prevalence ($\pi \in [1/10, 1/2]$). This is done to evaluate performance in two different scenarios – extreme (rare-event) v. non-extreme – and ensure a large-enough sample size to capture salient dynamics in both.

sampled at random from uniform distributions conforming to the specified range for each input - in the case of the maps, a map is chosen at random amongst the three candidates.

D.2 Candidate Models

A second step is to define the models analysed in this simulation study. The candidate models are: **m.1** – a simple fixed-effects logistic regression, where the area-effects are also estimated via fixed-effects; **m.2** – similar to **m.1**, but importantly augmented with the use on an offset (prior-correction) a-la (King and Zeng 2001); **m.3** – an essential version of our rare-events, Bayesian contaminated-controls model with a BYM2 area-effects prior. The models are detailed in Figure D.1.

(m.1): Fixed-effects logit	(m.2): King & Zeng	(m.3): Cerina et al.
$y_i \sim \text{Bernoulli}(\rho_i);$	(40) $y_i \sim \text{Bernoulli}(\rho_i);$	(45) $y_i \sim \text{Bernoulli}(\theta_{r_i});$
$\text{logit}(\rho_i) = \beta_1 + x_i\beta_2 + \sum_l z_{i,l}\gamma_l;$	(41) $\text{logit}(\rho_i) = \log \left[\left(\frac{1-\pi}{\pi} \right) \left(\frac{\bar{y}}{1-\bar{y}} \right) \right] +$	$r_i \sim \text{Bernoulli}(\rho_i);$
		(46) $\text{logit}(\rho_i) = \log \left(\frac{1}{\pi n_u} + 1 \right) +$
$\beta_1 \sim N(0, 10);$	(42) $\beta_1 \sim N(0, 10);$	$+ \beta_1 + x_i\beta_2 + \gamma_i;$
$\beta_2 \sim N(0, 1);$	(43) $\beta_2 \sim N(0, 1);$	(47) $\beta_1 \sim \text{Cauchy}(0, 10);$
$\gamma_l \sim N(0, 1).$	(44) $\gamma_l \sim N(0, 1).$	(48) $\beta_2 \sim \text{Cauchy}(0, 1);$
		(49) $\gamma_l = \sigma \left(\phi_l \sqrt{(1-\lambda)} + \psi_l \sqrt{(\lambda/s)} \right);$
		(50) $\gamma_l = \sigma \left(\phi_l \sqrt{(1-\lambda)} + \psi_l \sqrt{(\lambda/s)} \right);$
		(51) $\lambda \sim \text{Beta}(0.5, 0.5);$
		(52) $\phi_l \sim N(0, 1);$
		(53) $\psi_l \sim N \left(\frac{\sum_{l' \neq l} \psi_{l'}}{d_{l,l}}, \frac{1}{\sqrt{d_{l,l}}} \right);$
		(54) $\sigma \sim \frac{1}{2} N(0, 1).$
		(55) $\sigma \sim \frac{1}{2} N(0, 1).$
		(56) $\sigma \sim \frac{1}{2} N(0, 1).$
		(57) $\sigma \sim \frac{1}{2} N(0, 1).$
		(58) $\sigma \sim \frac{1}{2} N(0, 1).$
		(59) $\sigma \sim \frac{1}{2} N(0, 1).$
		(60) $\sigma \sim \frac{1}{2} N(0, 1).$
		(61) $\sigma \sim \frac{1}{2} N(0, 1).$

Figure D.1: Hierarchical formulation of the three competing models considered in the simulation study.

D.3 Computational Constraints

In order to fit the 600 models necessary for this simulation study, we have to ‘live dangerously’¹³, and lower our expectations over the stringent convergence properties of any given model. What we are interested in is the stability of the simulation results, and this is an aggregate set of quantities which is relatively robust to the semi-convergence of any given model. We therefore run each model in **Stan**, with the following settings: `n.cores = 4`; `n.chains = 4`; `n.thin = 4`; `n.iter = 4000`; `n.warmup = $\frac{2}{3}$ n.iter`;

¹³As others have done before when frequently fitting complex **Stan** models on large datasets – see (Lauderdale et al. 2020) for an example running multiple short chains.

all other settings are set to the `Stan` default. This gives us posterior samples which have relatively small effective sample-sizes, but are nevertheless able to give us reliable central-estimates for the parameters of interest – proof of this is that the results from the simulation study are replicable over multiple samples. Note that very rarely the chains will diverge for m.3 under high-levels of contamination. When this happens, we drop these simulations from the analysis and re-run the model.

D.4 Comparison Metrics

Finally, we define the parameters of the comparison. The simulations are intended to investigate the ability of competing models to estimate the following quantities of interest: $\boldsymbol{\mu}^* = \boldsymbol{\mu} - \log\left(\frac{P_1}{P_0}\right)$, the latent propensity to be a recruit; β_1 , the baseline propensity to be a recruit; β_2 , the effect of simulated covariate \boldsymbol{x} ; $\boldsymbol{\gamma}$, the set of area-level effects which contribute to the latent propensity. $\boldsymbol{\mu}^*$ is a good summary metric of performance on all of these dimensions, so our primary inference refers to this quantity. The models are scored only on their point-estimates, as an evaluation of uncertainty is computationally unfeasible due to the large number of MCMC iterations necessary to obtain convergent estimates of the second-moment for all these parameters.

The models are generating parameter estimates \hat{f} to approximate the true simulated parameters f ; they are scored on three dimensions: i. $\text{bias}(\hat{f}) = \frac{1}{n} \sum_i \hat{f}_i - f_i$; ii. Root-mean-square-error $\text{RMSE}(\hat{f}) = \frac{1}{n} \sum_i (\hat{f}_i - f_i)^2$; iii. Pearson correlation coefficient $r(\hat{f}) = \frac{\sum_i (\hat{f}_i - \bar{\hat{f}})(f_i - \bar{f})}{\sqrt{\sum_i (\hat{f}_i - \bar{\hat{f}})^2 \sum_i (f_i - \bar{f})^2}}$. The bias tells us the average direction of the estimation error; the RMSE tells us about the average magnitude of the error, penalising large deviations more heavily than smaller-ones; the Pearson correlation tells us about the ability of the model to correctly order (rank) the parameters.

D.5 Results

Figure D.2 presents a comparison of m.2 and m.3 in their ability to estimate latent propensity $\boldsymbol{\mu}^*$, for each scoring function (on the y-axis) across key characteristics of the data (on the x-axis). A comparison including m.1 is initially omitted here as the scale of the errors in m.1 is so large that it makes it visually impossible to distinguish between the (otherwise substantial) differences in m.2 and m.3 performance. A complete plot including m.1 is available below.

A visual analysis of Figure D.2 presents two clear dimensions in which our model advances the literature: i) m.3 is superior at moderate levels of prevalence ($\pi > 0.1$), a feat obtained thanks to the contamination layer of the model; ii) m.3 is superior under moderate-to-high levels of spatial auto-correlation ($I > 0.2$), due to the BYM2 spatial component. Related to the first advantage, we note that as the discrepancy between population and sample prevalence becomes positive ($\pi - \hat{\pi} > 0$), the upward bias

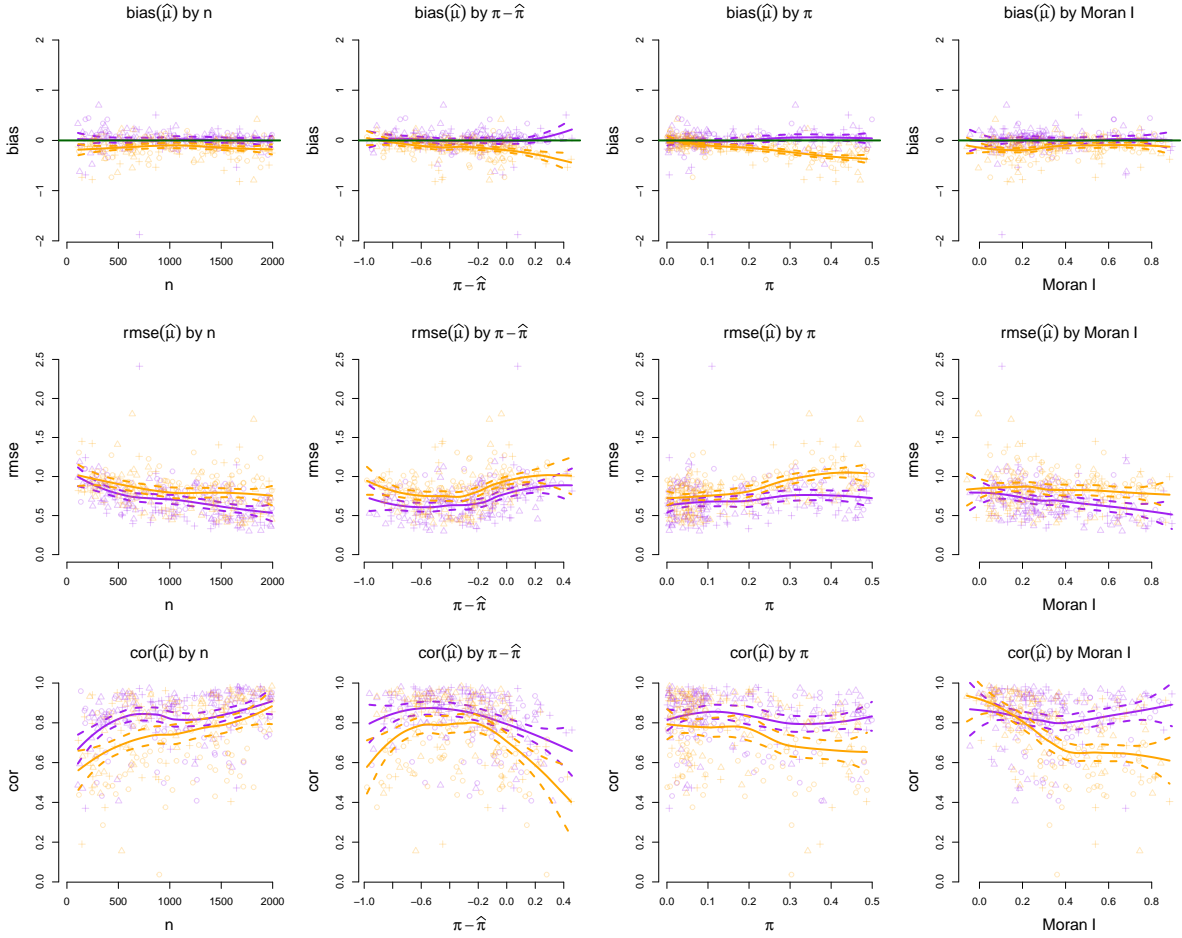


Figure D.2: Results of the simulation study, comparing the performance of our model (m.3, in purple) and the more traditional rare-events logistic regression with prior-correction for the intercept (King and Zeng 2001) (m.2, in orange) in estimating the true latent propensity $\mu^* = \mu - \log(\frac{P_1}{P_0})$.

which m.3 suffers from as a result of contamination is significantly more contained than the downward bias which characterises m.2 as a result of a non-contaminated offset, again highlighting another robustness advantage, pertaining to the relationship between sample and population prevalence. Moreover, Figures D.5 and D.6, which present the ability of m.2 and m.3 to estimate respectively the correct intercept parameter β_1 and the covariate effect β_2 , also paint a favourable picture. The ability of our model to perform under high levels of prevalence affords significant reductions in bias and RMSE, in both β_1 and β_2 , already at moderate levels of contamination. Figure D.7 compares models in their ability to estimate the correct area-level effect. Though all three models are, unsurprisingly, unbiased, m.3 is clearly more precise (lower RMSE) and better at ordering areas according to their propensity (higher Pearson correlation), in the presence of spatial auto-correlation.

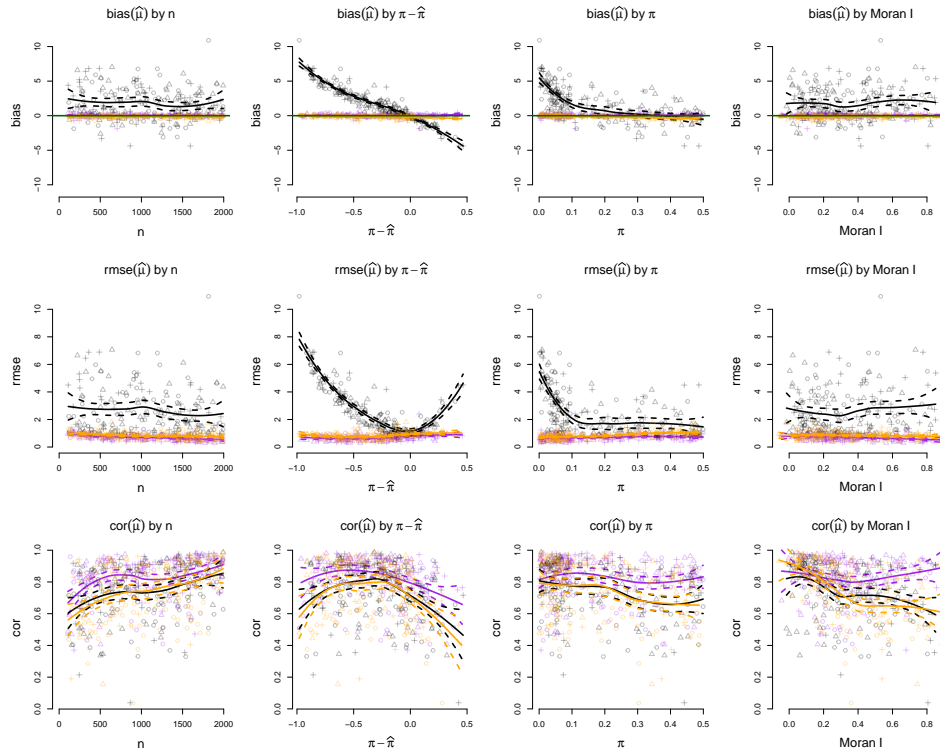


Figure D.3: Results from the simulation study, capturing the ability of the simple fixed effects model (m.1, in black), the King & Zeng model (m.2, in orange) and our proposed approach (m.3, in purple) to estimate the latent propensity of recruitment for each record in our sample μ^* .

Figure D.3 presents the scoring of models in their ability to predict latent propensity μ ; Figure D.4 displays the models' performance in estimating the baseline propensity β_1 , with Figure D.5 zooming-in to a comparison between our proposed model and the rare-events logit by King & Zeng; Figure D.4 shows model performance in estimating covariate effect β_2 ; Figure D.7 presents a comparison with respect to the estimation of area-level effects γ .

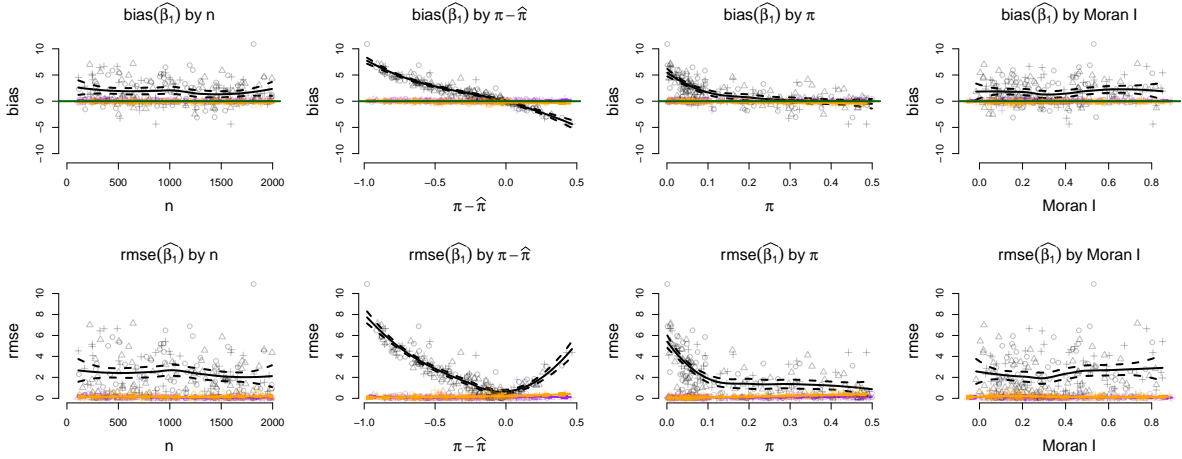


Figure D.4: Results from the simulation study, capturing the ability of the simple fixed effects model (m.1, in black), the King & Zeng model (m.2, in orange) and our proposed approach (m.3, in purple) to estimate the true intercept β_1 .

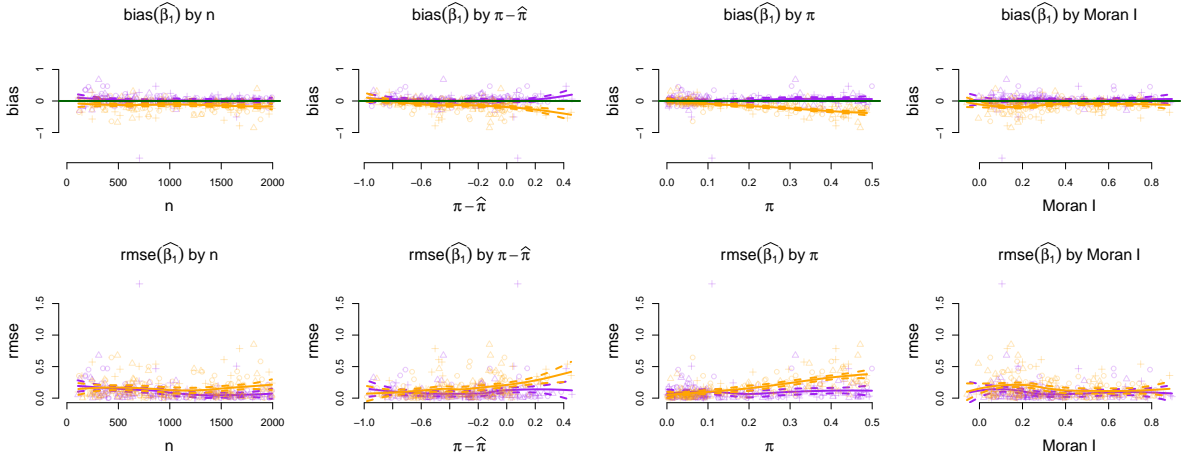


Figure D.5: Results from the simulation study, capturing the ability of the King & Zeng model (m.2, in orange) and our proposed approach (m.3, in purple) to estimate the true intercept β_1 .

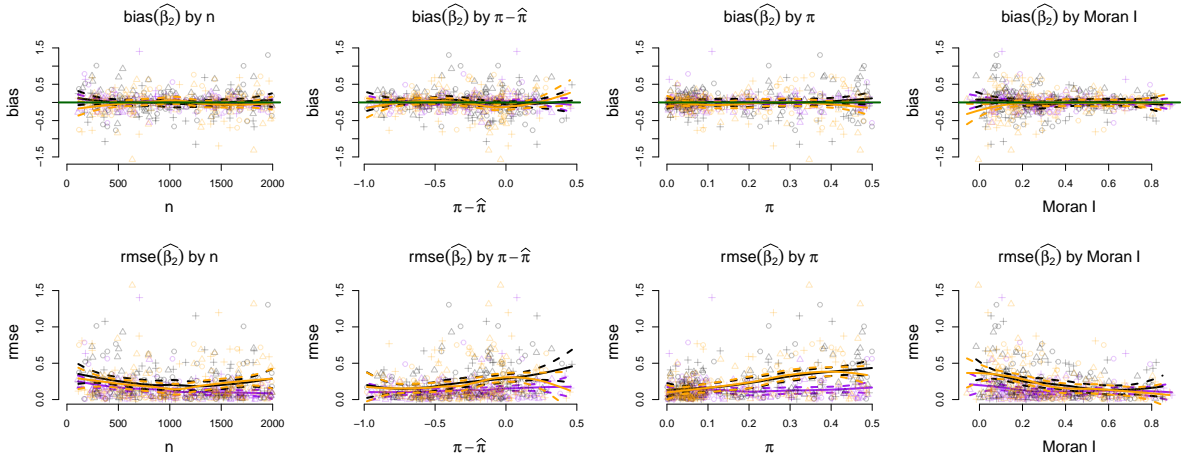


Figure D.6: Results from the simulation study, capturing the ability of the simple fixed effects model (m.1, in black), the King & Zeng model (m.2, in orange) and our proposed approach (m.3, in purple) to estimate the true covariate effect β_2 .

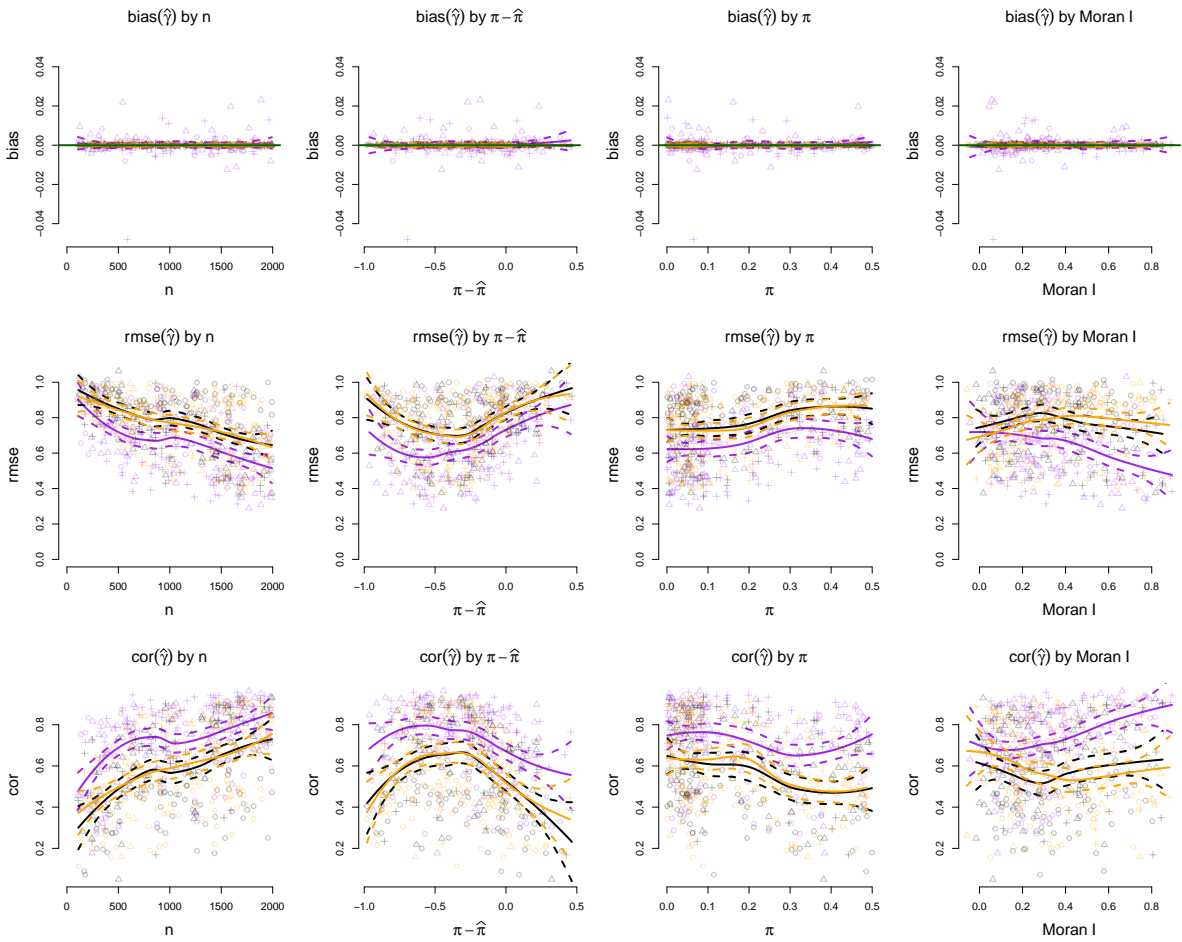


Figure D.7: Results from the simulation study, capturing the ability of the simple fixed effects model (m.1, in black), the King & Zeng model (m.2, in orange) and our proposed approach (m.3, in purple) to estimate the true area-level effects γ .

E Practical Advice for Researchers

In this section we provide guidance to applied researchers who seek to analyse data on recruitment to extremist organisations - or any other sort of data which can plausibly be affected by selection on the dependent variable, contamination, spatial auto-correlation, etc. - using our modeling framework. We focus on specific features of a typical application, and leverage lessons from our simulation study as well as our experience from the ISIS recruitment example detailed in this paper. We also address features of the modeling framework which have not been tested, presenting our current understanding of their potential impact, and outlining ways in which future research could mitigate their influence. We refer to [Rosenfeld \(2018\)](#) for further treatment of the underlying assumptions of a contaminated case-control model.

- i. **Sample size:** our proposed strategy is unbiased for estimating recruitment propensity μ starting at $n > 100$ (the minimum sample size tested in our simulation study). Compared to existing alternatives, our model affords greater returns on additional samples in terms of reduction of average error (RMSE). The advantage over other alternatives seems particularly evident for $n > 1000$ - where alternatives' RMSE tends to plateau, whilst our model's RMSE continues to decrease in seemingly linear fashion. A breakdown by estimated parameter tells us that our model affords some meaningful RMSE gains for every estimated parameter: for regression coefficients (both β_1 and β_2) we see a stable reduction of RMSE persisting after $n > 1000$, whilst alternative models plateau; the largest efficiency gains however are observed on the area-level effects γ .

Our advice with respect to sample size is to use our model over any existing alternatives at any sample size. The recommended sample size to obtain the best results is $n > 2000$. We suspect decreasing returns will kick-in at some stage, but we have not tested where that point might be, and leave it for future research.

- ii. **Population prevalence:** π plays a two-fold role in our model: it is used to calculate an optimal offset to account for selection, and it is used to account for contamination in the unlabeled controls. Our model is robust to any level of $\pi \in (0, 0.5)$. Under a contaminated data generating process, this is in stark contrast with the best available alternative - the King & Zeng model - which suffers from negative-bias as the contamination rate increases. The gains appear primarily as a result of estimating an unbiased intercept β_1 , but also as a result of having a precise estimate of β_2 at any level of π - whilst the King & Zeng model suffers from roughly linearly-increasing RMSE at increased contamination rates.

Our advice with respect to the true population prevalence is to use our model at any level of π . There is no recommended level of prevalence at which our model's ability to estimate either propensity μ or regression coefficients β_1 and β_2 underperforms - our model is simply robust to contamination, given knowledge of π is available.

- iii. **Discrepancy between sample & population prevalence:** $\pi - \hat{\pi}$ captures the difference between the true prevalence and the sample-prevalence. This arises as a direct result of the stacking procedure, where whatever available cases are artificially appended to a sample of unlabeled controls. Note that this is a measure of the degree to which our sample ends up being a non-probability sample.

Here, compared to the best available alternative, we have a trade-off: both models will tend to have biased intercepts at very high-levels of $\pi - \hat{\pi}$. The King & Zeng model will tend towards a negative bias, as the offset used here does not take into account contamination, and the relatively large numbers of unlabeled controls will be ‘hiding’ a very large number of cases, meaning the offset is inaccurately calibrated. This model essentially believes that there are less cases in the data than in reality, and therefore estimates a smaller intercept than it should. On the other hand, our model will tend to do the opposite: by accounting for contamination, it will tend to believe there are relatively more cases in the sample than there actually are in the population. This bias however tends to be smaller, and appear at higher levels of discrepancy, compared to that of the King & Zeng model.

Our advice with respect to the size of the discrepancy between sample and population prevalence is as follows: researchers should prefer to account for contamination rather than not, due to the relatively smaller bias and RMSE on the intercept. They should however be aware that in a regime of large under-sampling of the cases, the bias will be in the positive direction under our contamination model, but in the negative direction if contamination is not accounted for. Researchers should attempt to create stacked-samples that have sample prevalence roughly equal to population prevalence; over-sampling of cases is not an issue, but under-sampling is, if the discrepancy is large. As a conservative guide, we advise researchers to ensure their stacked sample does not under-sample cases by more than 10 percentage points, relative to population prevalence, when using our model.

- iv. **Spatial auto-correlation:** spatial auto-correlation in the area-level effects γ tends not to affect the bias of the model estimates, but it does impact efficiency (in terms of RMSE) and the ability of models to properly rank individuals according to their underlying recruitment propensity μ . Large discrepancies in RMSE and correlation of γ tend to kick-in around a Moran- I of 0.2; the comparison with other models increasingly favours our approach as I increases, generating an advantage as large as 0.2 correlation points at high-levels of spatial auto-correlation. Note that under low-levels of auto-correlation, our model performs at least equally well as any other alternatives.

Our advice with respect to spatial auto-correlation is to explicitly account for it in the model. There are no drawbacks to doing so in terms of the metrics we test. There might be some issues relative to the ‘coverage’ ability of the model,

which we could not test due to computational burden, in the sense that estimates of γ under our model will be greatly ‘shrunk’ and therefore tend to have smaller uncertainty than fixed effects. But regardless of the potential coverage issues, the gains in terms of ability to order and discriminate between profiles are sufficiently large we feel comfortable advising to use the ICAR model to account for spatial auto-correlation.

- v. **Prior knowledge of population prevalence:** we encode in our model’s contamination layer an expectation that the unlabeled cases will be recruited at a rate roughly equal to the population prevalence of recruitment. The validity of this assumption depends on the application at hand. For us, the known population propensity made for a good prior because our hypothetically-contaminated observations came from a random sample of the population, so the known prevalence, and the sampling design of the Arab Barometer, were at the same level. However this assumption becomes increasingly inappropriate as the sampling frame of the contaminated units veers further away from a random sample of the population of interest. The degree of error induced by misspecification is not tested in our simulation study - it is assumed that the correct prevalence is always known.

Speculating on the potential effects of misspecification we can consider the nature of the bias we would be introducing: artificially increasing the contamination rate relative to the true population rate will positively bias the intercept of the model, by ‘flipping’ an unreasonable amount of unlabeled observations to ‘cases’. This is similar to what we see in the effect of sample and population prevalence discrepancy $\pi - \hat{\pi}$.

In a fully Bayesian model this parameter can be estimated from the data ([Rota et al. \(2013\)](#); [Rosenfeld \(2018\)](#)) though we have found, anecdotally in our experimenting for this paper, that under this fully-Bayesian approach the estimated posterior of π tends to be necessarily biased by the selection effects into the artificially-stacked samples. This is somewhat in contradiction with [Rosenfeld \(2018\)](#), and we merely point this out to encourage further exploration of this question.

Our advice with respect to how to best use prior knowledge on the population prevalence is to ‘use it with care’. If it is known with certainty, we advise to use it, and introduce it as an ‘observed value’ for the prevalence in the Bayesian model. If no knowledge on the true population prevalence is available, we do encourage researchers to perform a fully-Bayesian analysis and estimate the prevalence as part of the model parameters; however we would further advise, where possible, to use strongly informative priors to counteract the bias introduced by the artificial sampling design.

- vi. **Exogenous Selection:** a dimension on which our model is untested is the degree to which non-random samples of cases could bias the analysis. Examples of this would be if the sample of recruits in our data was obtained via snowball sampling,

or any other sampling design which is vulnerable to systematic distortions brought about by exogenous effects.

This paper takes the view that it is generally hard-to-impossible to obtain representative samples of recruits, and therefore builds-in a series of robustness measures - such as regularising priors, random effects, and mixture models, to limit the negative effects of non-representativity. More regularising approaches are worth exploring and introducing into the modeling framework - for instance the use of regularised horseshoe priors ([Piironen and Vehtari \(2017\)](#)) could contribute to excluding irrelevant covariates from the analysis, as well as regularising coefficients, hence encouraging the avoidance of over-fitting to potentially biased data. It is further worth noting that this model is amenable to post-stratification ([Hanretty et al. \(2018\)](#); [Park et al. \(2004\)](#)), which would allow for more representative estimation of average recruitment propensities at the small-area level. Enhancing this analysis with a post-stratification layer would enable the use of even more extreme unrepresentative samples of recruits, such as individuals observed to be extremists on social media or in other unconventional samples ([Wang et al. \(2015\)](#); [Cerina and Duch \(2020\)](#)). But this approach would not solve the bias in regression coefficients, and is only relevant if the small-area is indeed the desired level of analysis.

Our advice with respect to the potential impact of exogeneous selection effects is to build into the model reasonable protections against over-fitting to biased data. In our case, this is possible to some degree through the use of various regularising priors. In general the conventional wisdom stands: if it is at all possible to obtain a representative sample of cases, researchers should do whatever they can to obtain it. However, reality dictates that this is very rarely possible, especially in the context of extremist movements. Our regularised modeling approach therefore becomes the preferred solution.

F Convergence diagnostics

To ensure absolute convergence of all model parameters we run our model with extremely conservative settings: `n.iter` > 10,000, `n.warmup` > 9,000,¹⁴ `n.chains` > 4; `n.thin` = `n.cores` = `n.chains`; `max_treedepth` = 25, `adapt_delta` = 0.99. Note that the Worm’s Eye models take around 12 hours to run for Egypt, 24 hours for Tunisia, whilst the Bird’s Eye model takes 48 hours. As a final note, it’s worth highlighting that convergence of point estimates for the individual-level covariates happens under far more laxed estimates, and exploratory versions of this model can be fit under 1 hour in all cases.

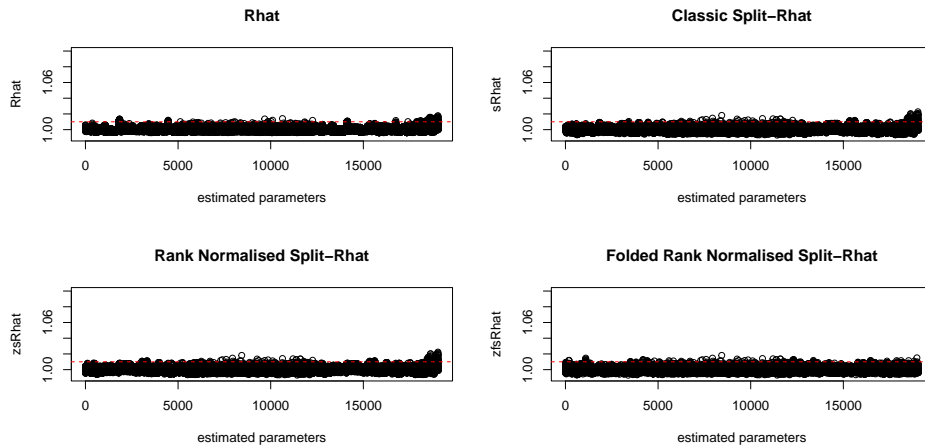


Figure F.1: Gelman-Rubin Statistics for the *Bird’s Eye* model.

¹⁴For the ‘Bird’s Eye model, we set `n.iter` = 10,000 and `n.warmup` = 9,000, and ran the model over 8 chains spread over 8 cores, thinning by a factor of 8 – whilst for the Worm’s Eye models we can afford a larger number of iterations – `n.iter`= 25,000 and `n.warmup` = 22,500, running 4 chains spread over 4 cores, and thinning by a factor of 4.

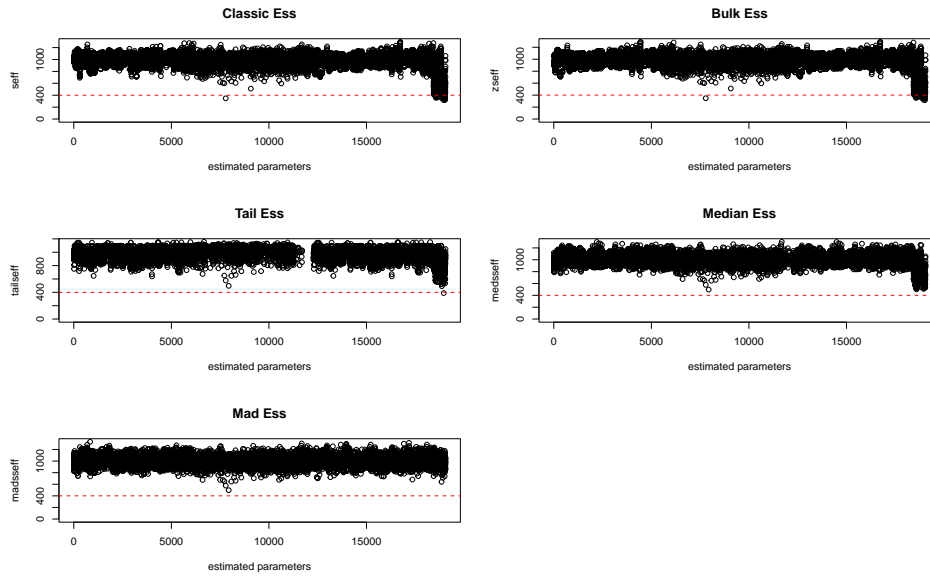


Figure F.2: Effective sample-size (ESS) for the parameters of the *Bird's Eye* model.

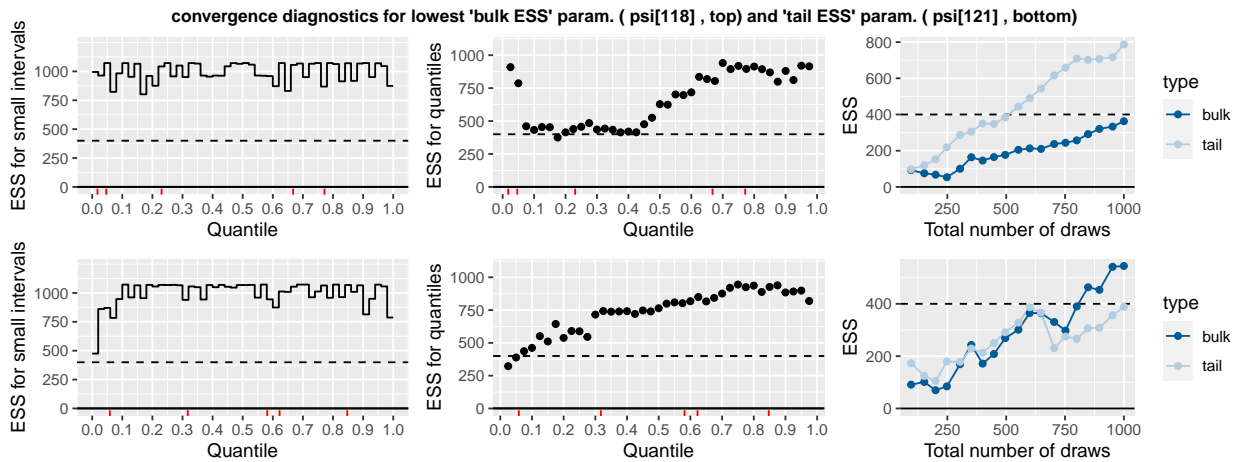


Figure F.3: Convergence dynamics for the parameters with the lowest bulk (top) and tail (bottom) ESS, for the *Bird's Eye* model. The quantile plots show satisfactory ESS for every section of the posterior distribution, whilst the positive and close-to-linear gradient in the ‘total number of draws’ plot suggests ESS would improve further by drawing more samples – a sign that the posterior is well-explored.

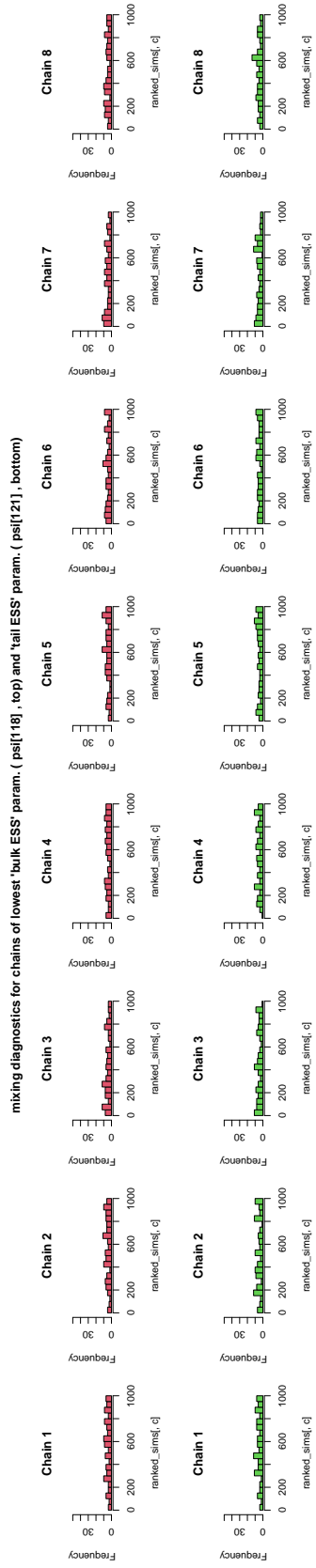


Figure F.4: Histogram of the ranked posterior draws for the parameters with the lowest bulk (top) and tail (bottom) ESS, for the *Bird's Eye* model. This plot is evidence of reasonably good mixing also for our 'least convergent' parameters, as the ranked draws from each chain could be reasonably drawn from a uniform distributions.

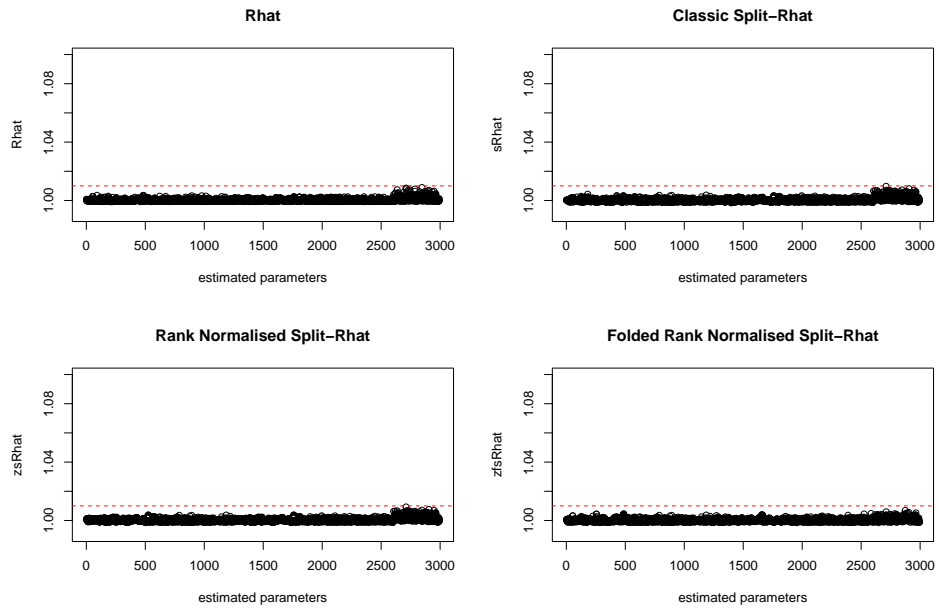


Figure F.5: Gelman-Rubin Statistics for the Egypt *Worm's Eye* model.

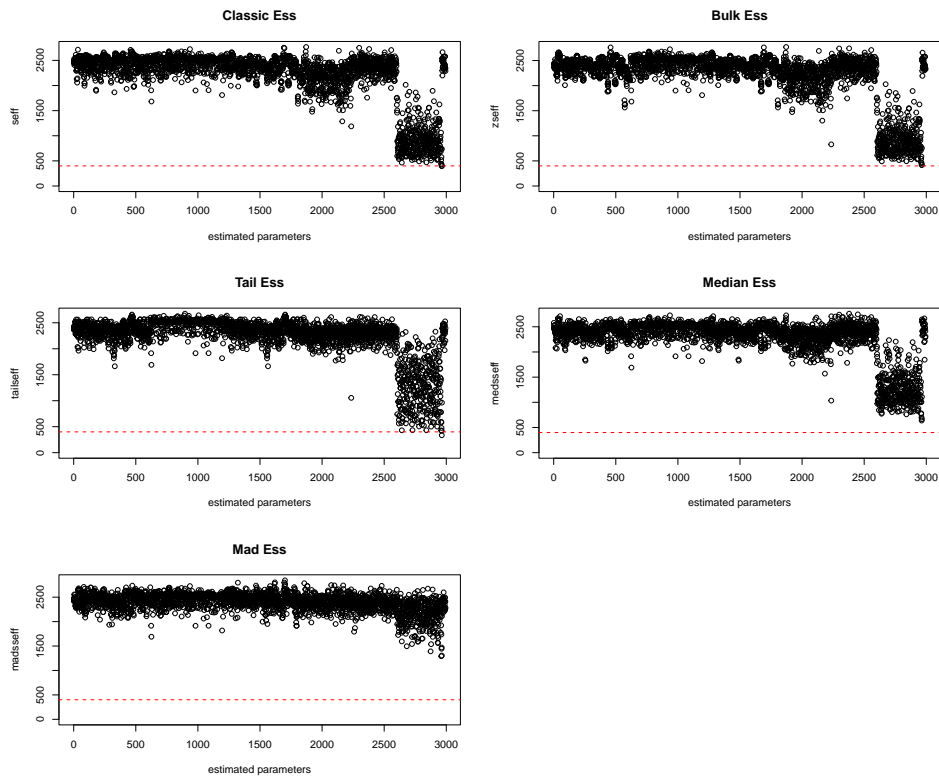


Figure F.6: Effective sample-size (ESS) for the parameters of the Egypt *Worm's Eye* model.

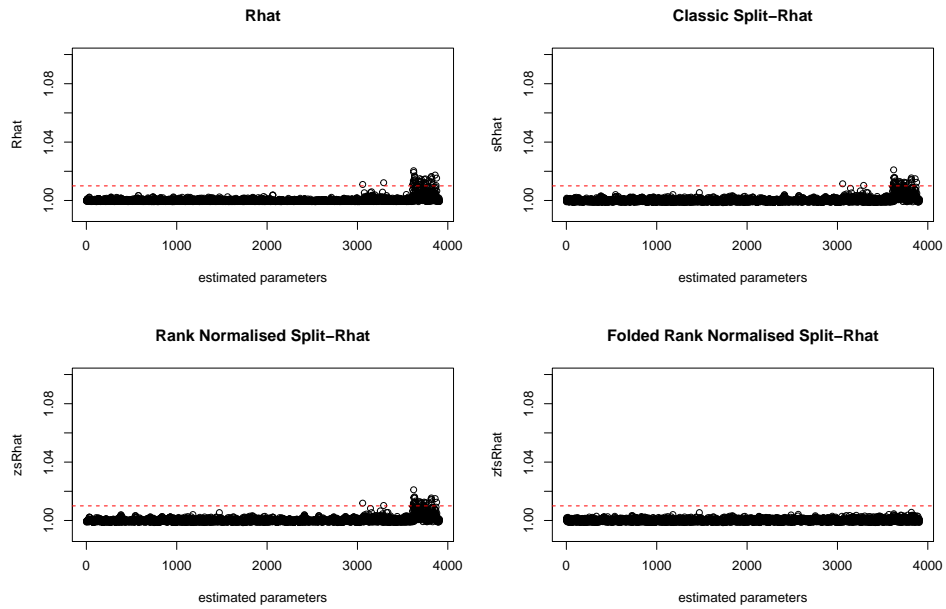


Figure F.7: Gelman-Rubin Statistics for the Tunisia *Worm's Eye* model.

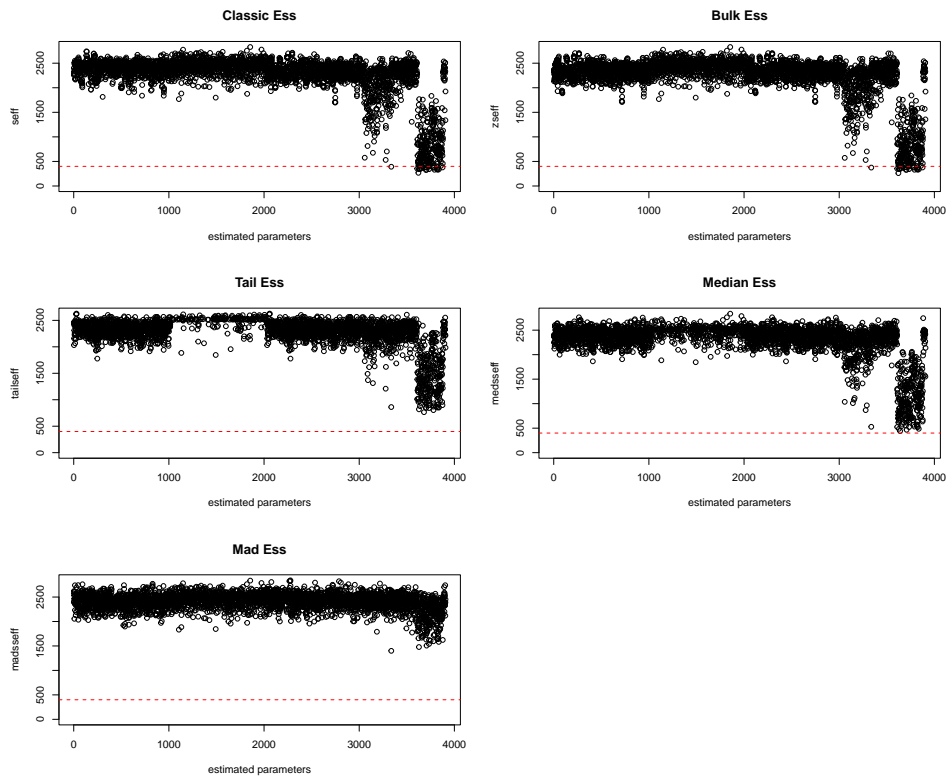


Figure F.8: Effective sample-size (ESS) for the parameters of the Tunisia *Worm's Eye* model.

G Posterior densities of regression coefficients

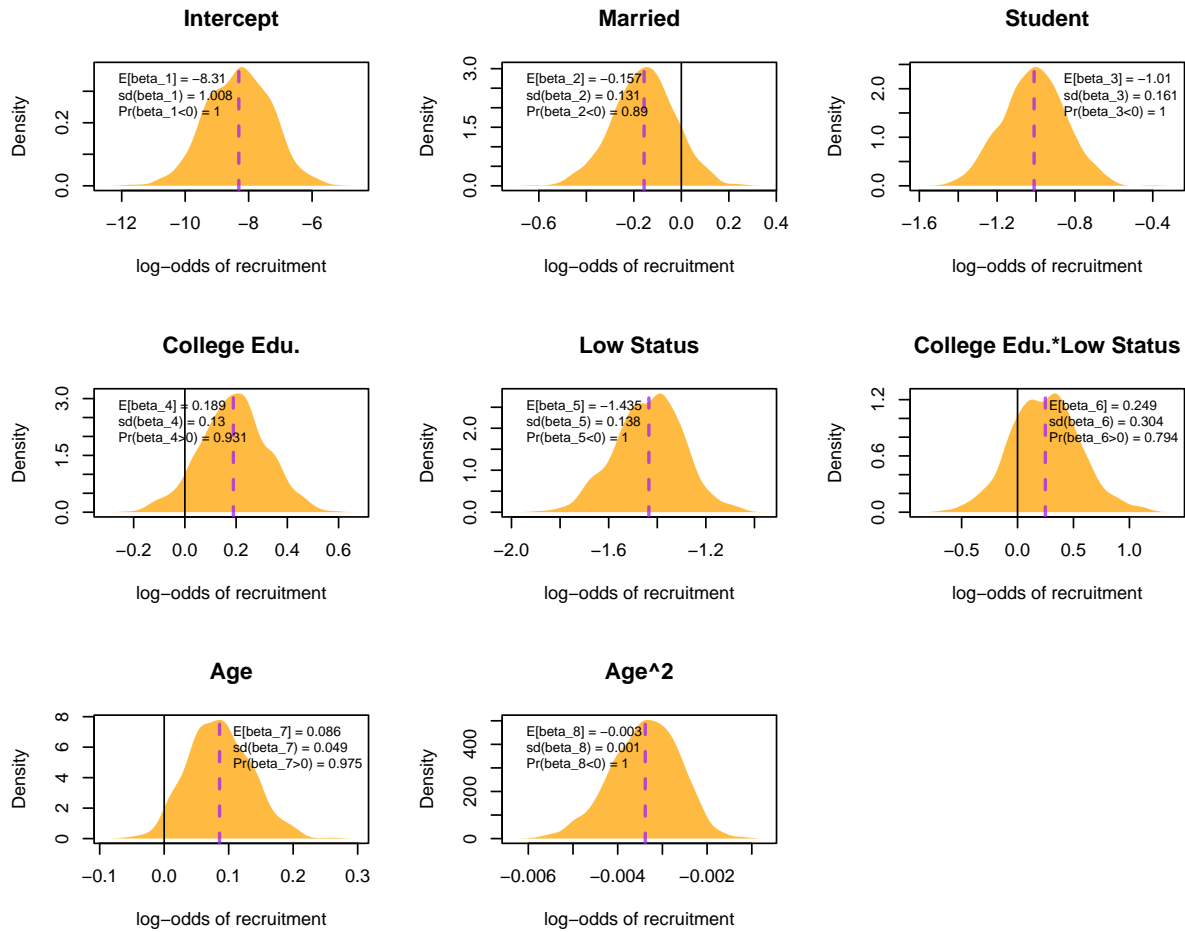


Figure G.1: Posterior density of individual-level fixed-effect coefficients for the 'Bird's Eye' model. These effects are presented on the original, non-standardized scale.

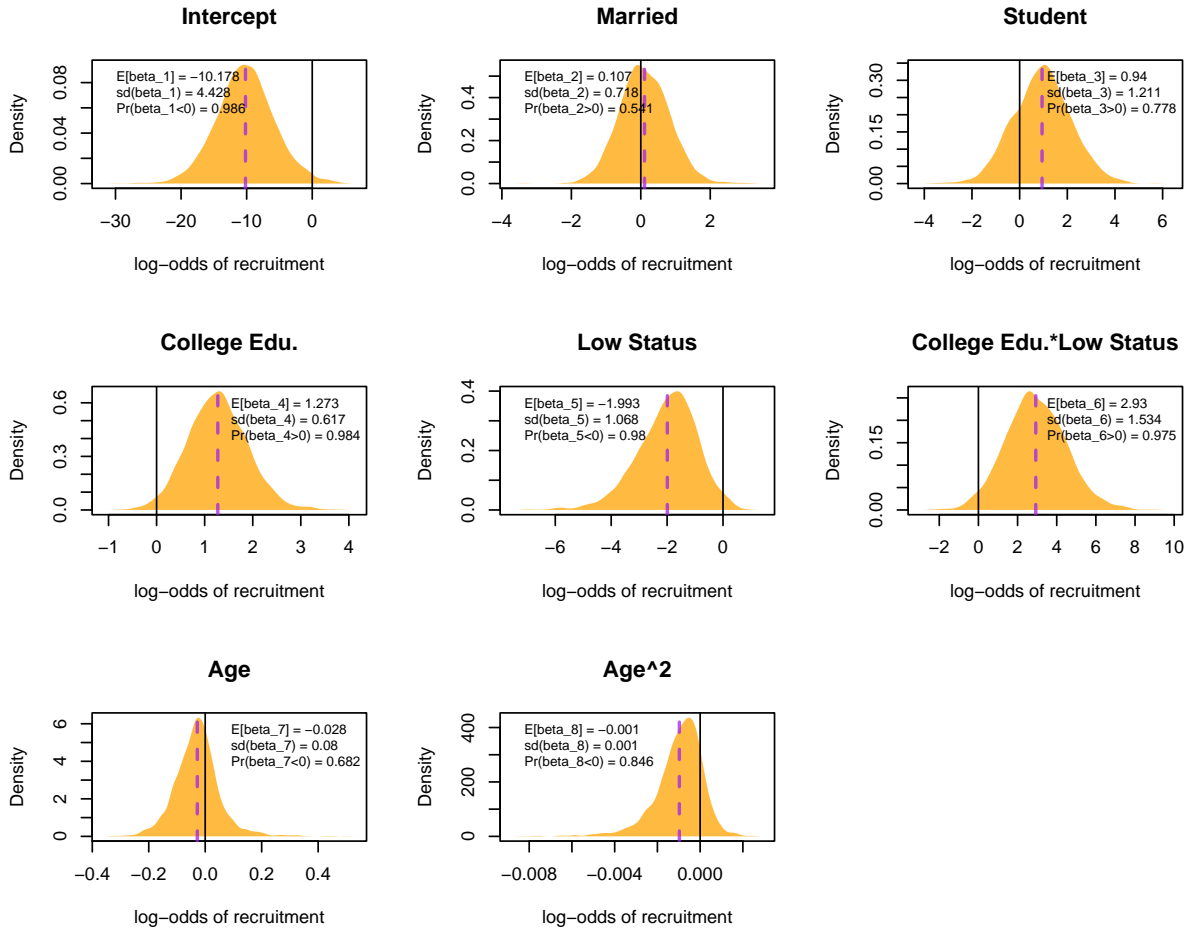


Figure G.2: Posterior density of individual-level fixed-effect coefficients for the Egypt 'Worm's Eye' model. These effects are presented on the original, non-standardized scale.

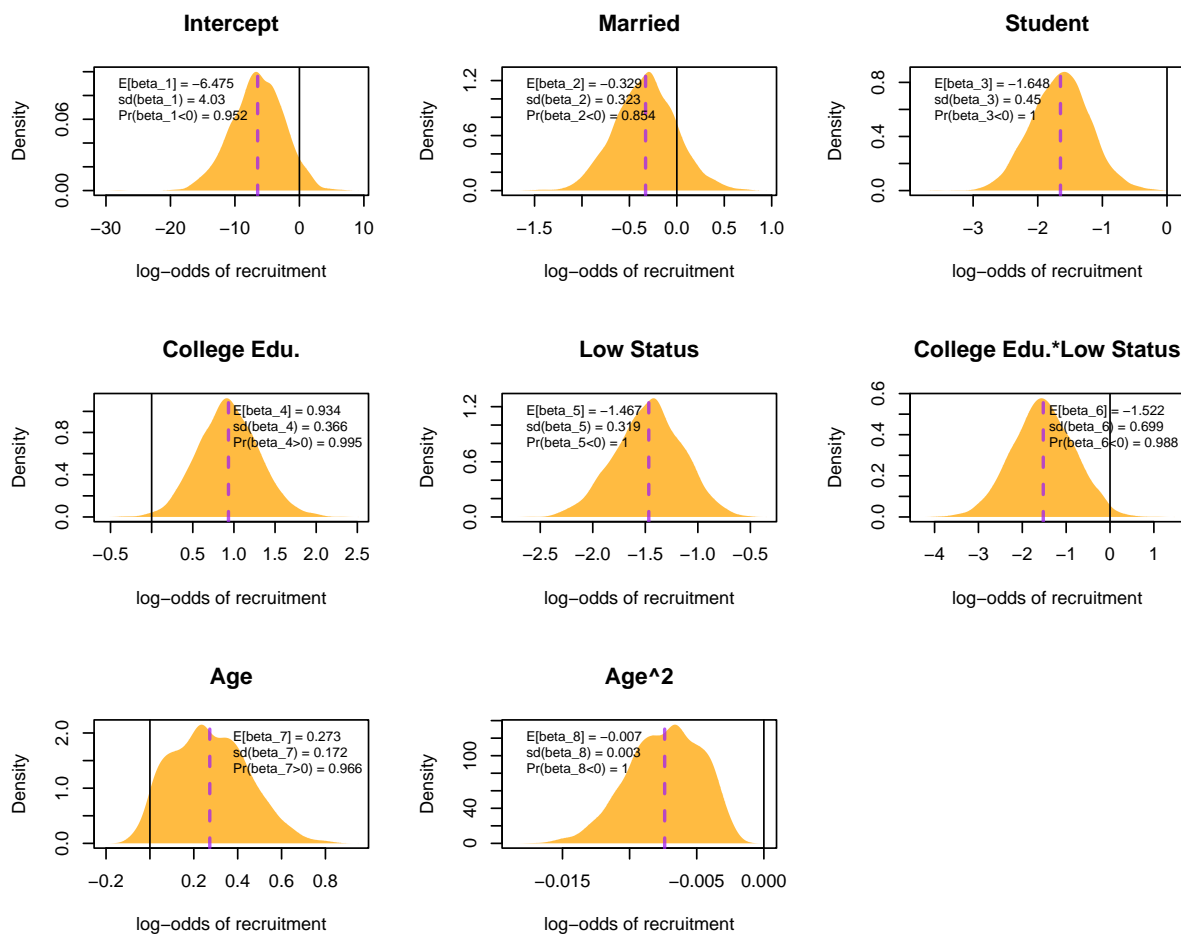
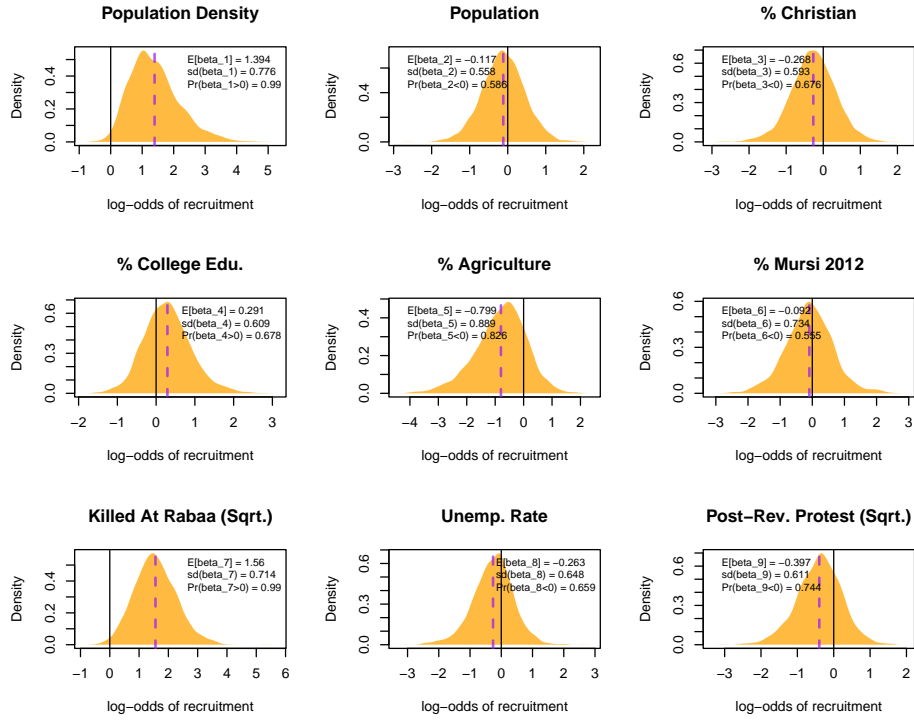
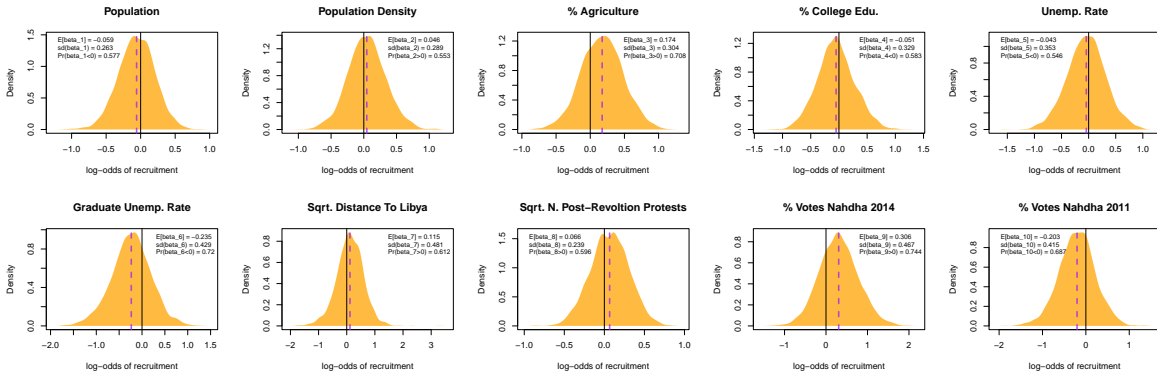


Figure G.3: Posterior density of individual-level fixed-effect coefficients for the Tunisia ‘Worm’s Eye’ model. These effects are presented on the original, non-standardized scale.

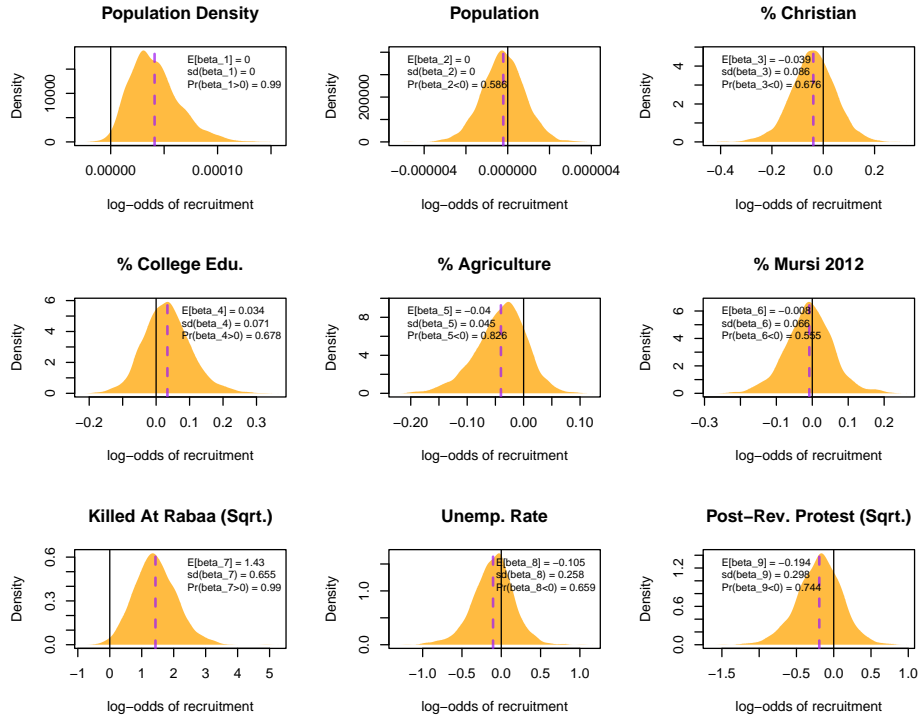


(a) Egypt

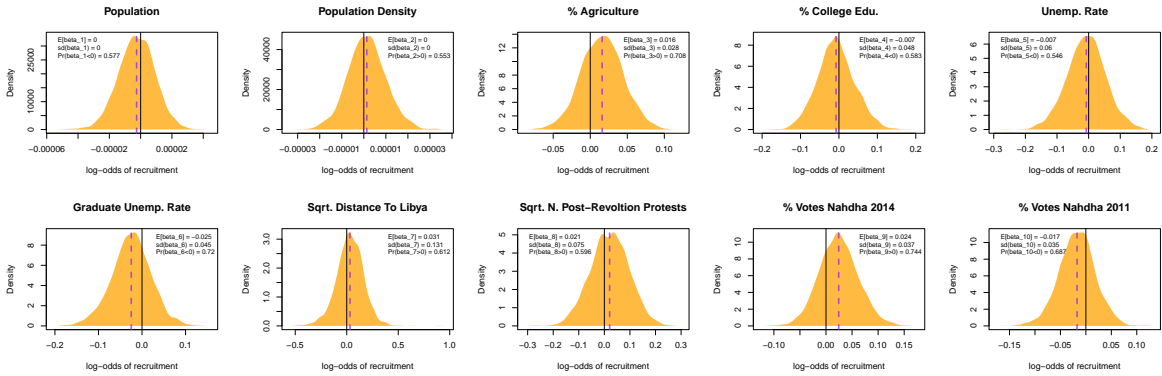


(b) Tunisia

Figure G.4: Posterior density of district-level fixed-effect coefficients for the *Worm's Eye* models.



(a) Egypt



(b) Tunisia

Figure G.5: Posterior density of district-level fixed-effect coefficients for the *Worm's Eye* models. These effects are presented on the original, non-standardized scale.

H Relative Deprivation Effects

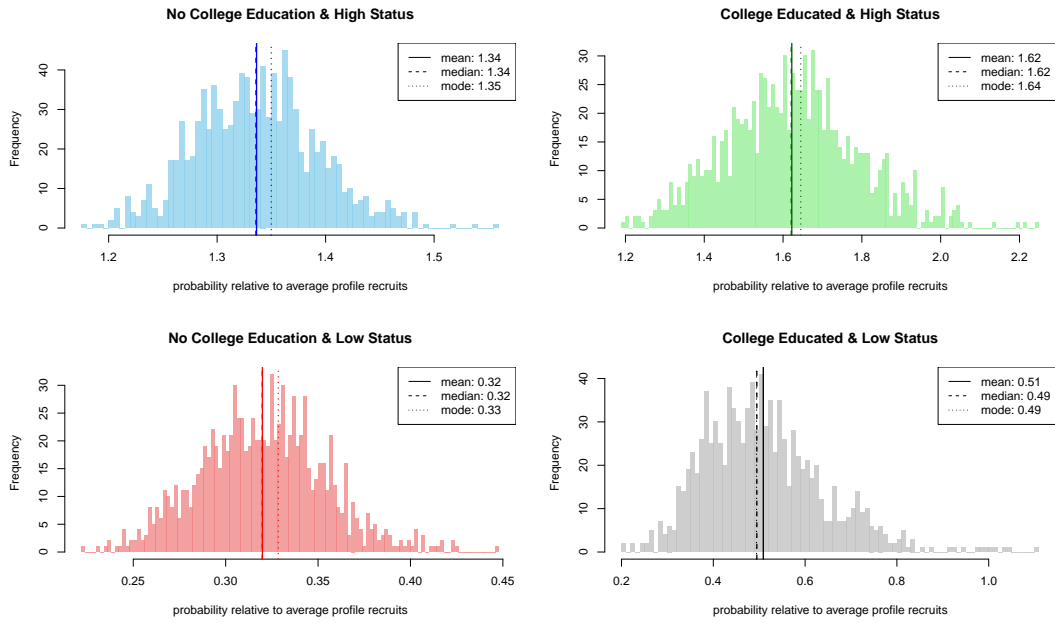


Figure H.1: Predicted propensity of recruitment for relative-deprivation profiles according to the 'Bird's Eye' model. The effects are presented as odds relative to the 'average' recruitment profile.

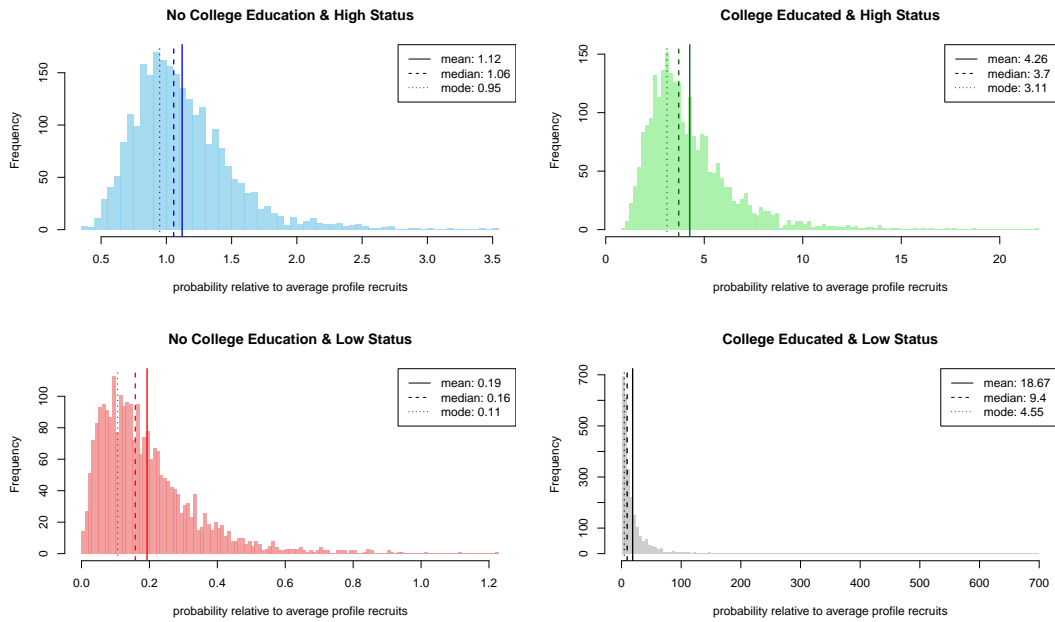


Figure H.2: Predicted propensity of recruitment for relative-deprivation profiles according to the Egypt 'Worm's Eye' model. The effects are presented as odds relative to the 'average' recruitment profile.

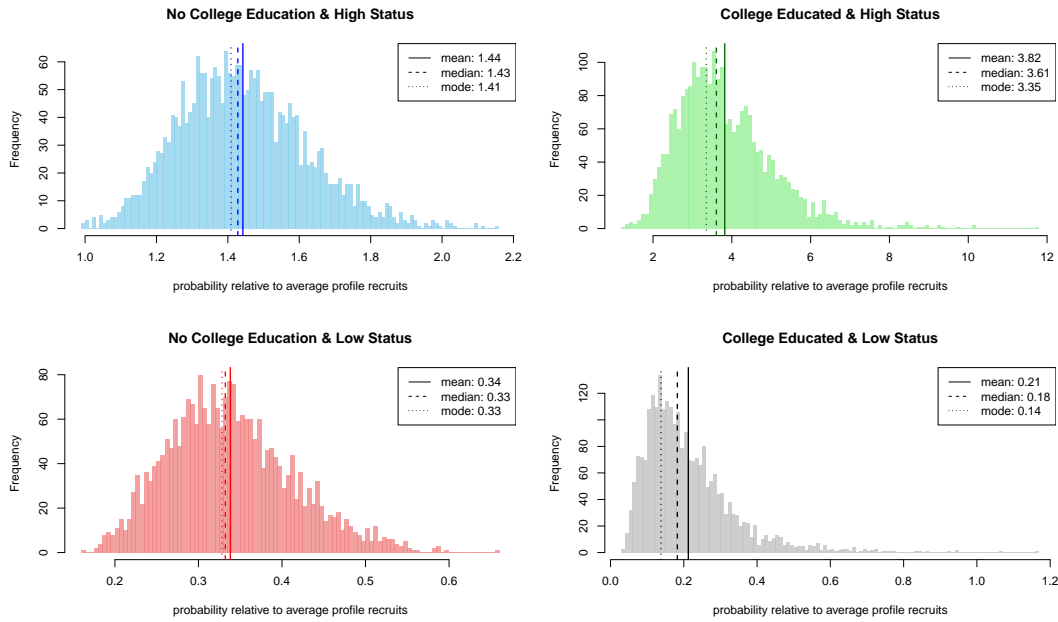


Figure H.3: Predicted propensity of recruitment for relative-deprivation profiles according to the Tunisia 'Worm's Eye' model. The effects are presented as odds relative to the 'average' recruitment profile.

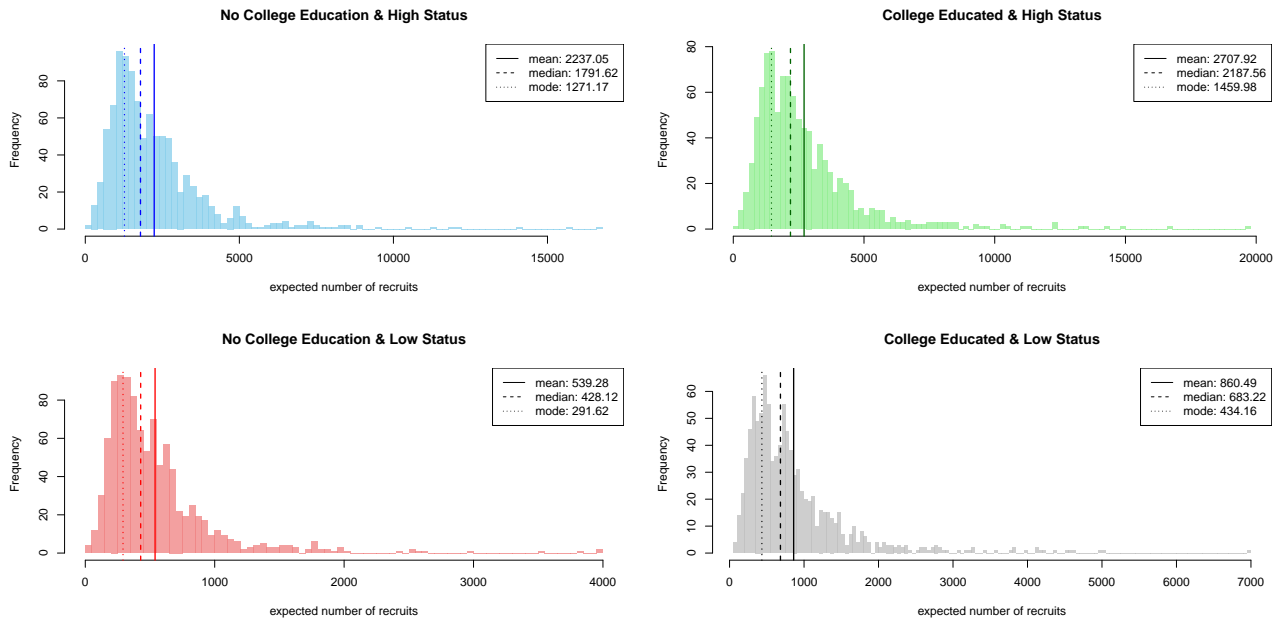


Figure H.4: Predicted propensity of recruitment for relative-deprivation profiles according to the 'Bird's Eye' model. The effects are presented as predicted counts under the assumption that everyone in the population is an 'average profile', and only changing the profile's relative deprivation status.

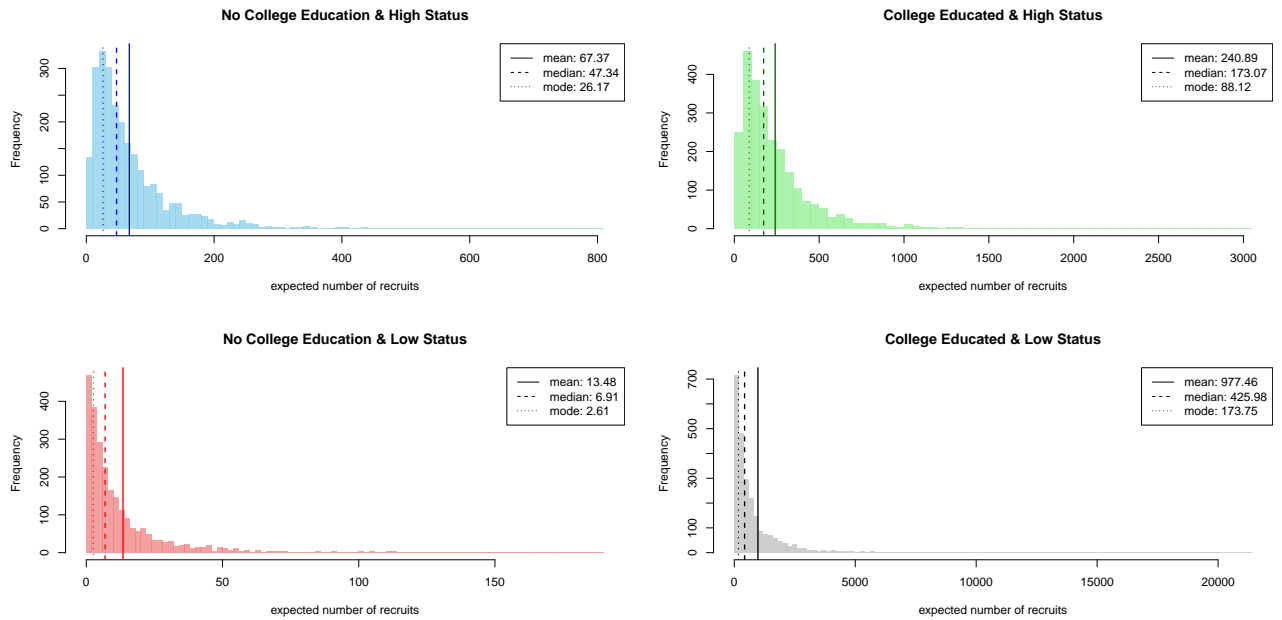


Figure H.5: Predicted propensity of recruitment for relative-deprivation profiles according to the Egypt ‘Worm’s Eye’ model. The effects are presented as predicted counts under the assumption that everyone in the population is an ‘average profile’, and only changing the profile’s relative deprivation status.

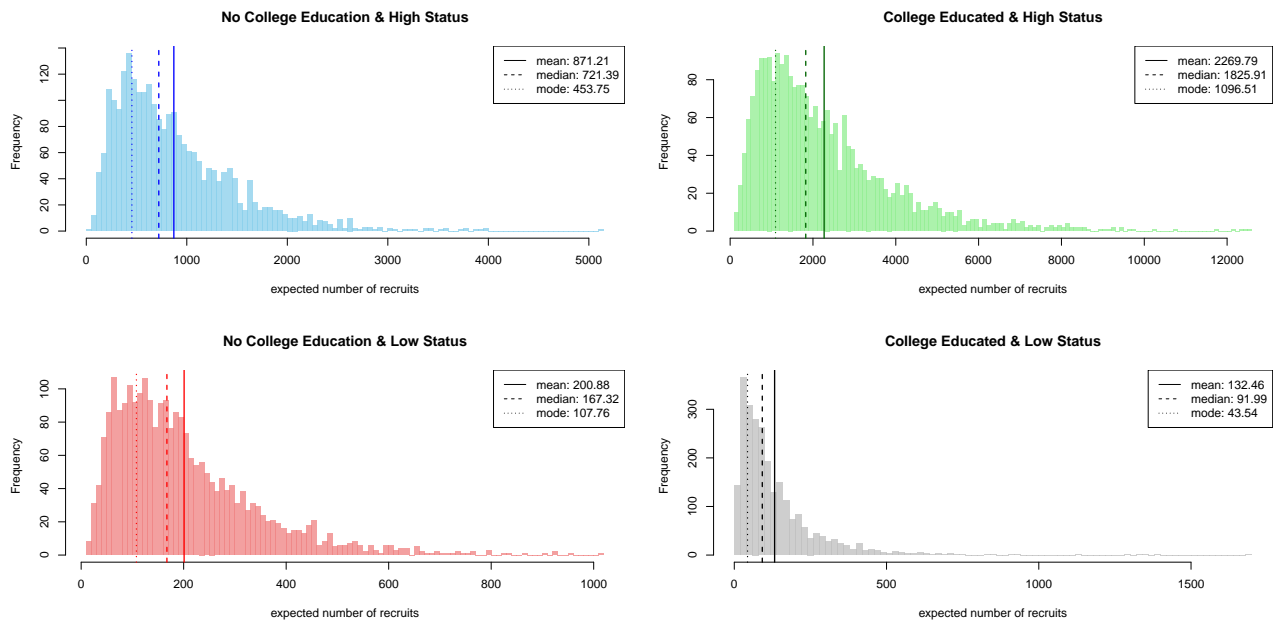


Figure H.6: Predicted propensity of recruitment for relative-deprivation profiles according to the Tunisia ‘Worm’s Eye’ model. The effects are presented as predicted counts under the assumption that everyone in the population is an ‘average profile’, and only changing the profile’s relative deprivation status.

I Residual Area-Level Analysis

In Figure 5a in the main manuscript, edges connect nodes identified by the centroids of governorates for each country. Minor adjustments were performed to ensure the absence of islands or sub-graphs, which would have made the analysis needlessly complicated. Note also that Israel and Saudi Arabia are included for the purpose of obtaining this fully-connected graph, but no observations were available for either of these countries in terms of recruits or Arab Barometer observations, and hence the direction of the estimates for their governorates is entirely driven by the spatial component. Supplementary Figure I.2 displays the observed number of recruits per area, along with the residual for each governorate.

We are satisfied that the spatial pattern implied by the adjacency matrix derived from the fully connected graph is completely extracted from the residuals, as shown by the relatively uniform color pallet of the rightmost map in Figure I.2, and most importantly the posterior distribution of the residuals' Moran's I in Figure I.1, which is normally distributed around the expected null value.

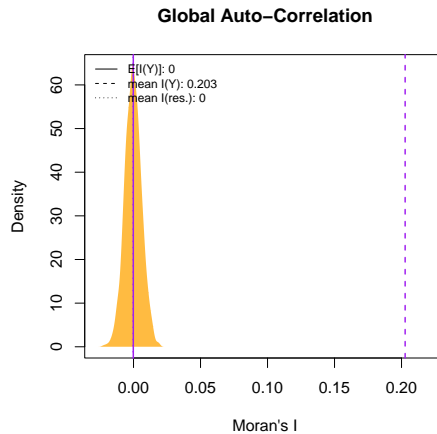


Figure I.1: Posterior distribution of Moran's I, a coefficient of global spatial auto-correlation. The adjacency matrix implied by Figure 5a is used as the weight matrix. $I(Y)$ indicates the coefficient value prior to spatial modeling; $I(res)$ shows the complete nullification of auto-correlation as a result of the ICAR prior. The expected value under the null distribution, $E[I(Y_{null})]$, is calculated as $\frac{-1}{(n_1+n_u)-1}$.

Figure I.2 displays the observed number of recruits per area, along with the residual for each governorate.

In I.2b the residual is calculated as follows: take $z = z_1, \dots, z_L$ to be the subset of individuals $i \in z_l$, who belong to small-area l ; take $s = 1, \dots, S$ to be the index of posterior sample draws; then $res_l = \frac{1}{S} \sum_s \left[\frac{1}{\sum_i \mathbb{1}(i \in z_l)} \sum_{i \in z_l} (y_i - \hat{y}_{i,s}) \right]$. A first concern is the presence of spatial autocorrelation in the recruitment data, which could bias individual-level coefficients. The spatial distribution in Figure I.2a seem to suggest the possibility of spillover effects around high-density coastal areas. This is confirmed by

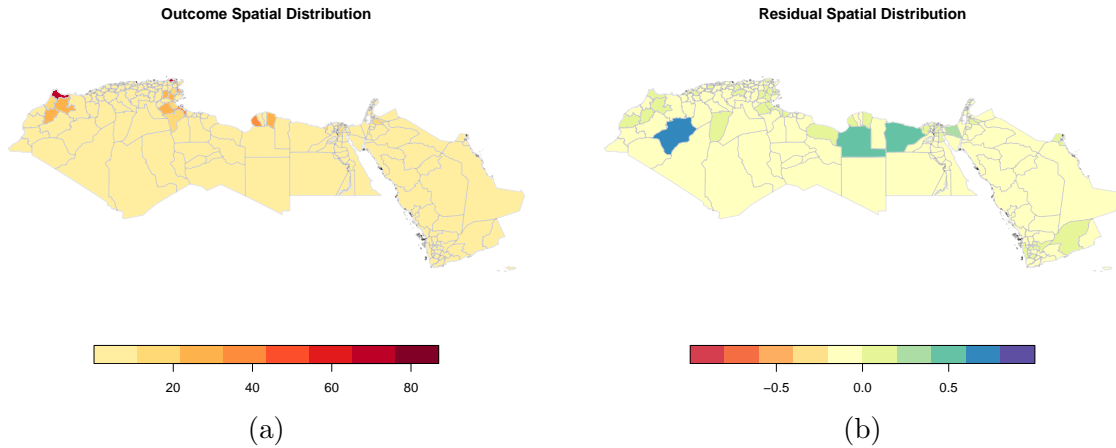


Figure I.2: Spatial distribution of observations (a) and residuals (b) at the Governorate level.

the Moran's I ($I(Y)$), which shows statistically significant spatial auto-correlation.¹⁵

We display below the spatial distribution of the point estimates for Governorate and Country-level random effects in Figure I.3. The corresponding prediction intervals for country and governorate effects are shown in Figure I.4 for country-effects and Figure I.5 for Governorate effects. It is worth noting that part of the reason for heightened recruitment propensity around the eastern Governorates could be the increasing proximity to Syria and the ISIS caliphate itself, as well as higher proportions of refugees from destabilized regions of Syria, and in general more potential for pro-ISIS unobservable network-dynamics. We see a strong unexplained effect in Tunisia, highlighting unobserved but systematic variance in favour of recruitment, while Algeria, Egypt and Yemen show significant unexplained negative effects on recruitment over and above their spatial and unstructured Governorate-level variance.

¹⁵As $I(Y)$ is an observed, not modeled, quantity, it carries no uncertainty around it; it is reasonable to assume that the distribution of the $I(Y)$ would be the same as that of the $I(res)$ in terms of its shape and variance, and only differ as a result of the mean parameter. This is what is commonly assumed under standard hypothesis testing. Hence, it is easy to see that by applying the extremely narrow simulated variance around the $I(Y)$ dotted line, there would be a 0 probability of that distribution crossing the $E[I(Y)]$ line, and hence we can say the $I(Y)$ is highly significant. Calculating the significance of $I(Y)$ in frequentist terms, using the `ape` package, reveals a p-value of 0. We plot and describe our calculation for Moran's I in Figure I.1

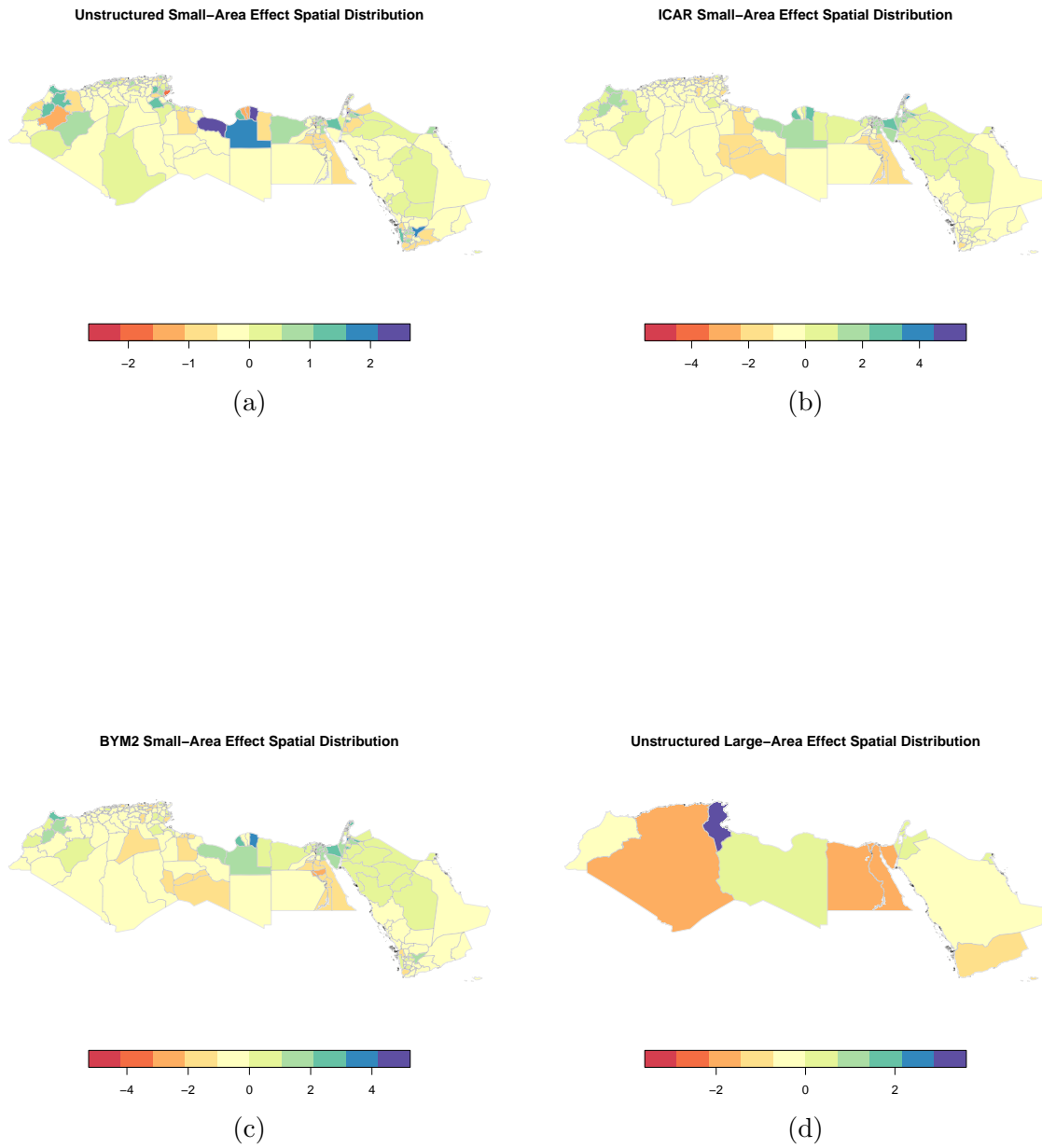


Figure I.3: Spatial distribution of: (a) the unstructured Governorate-level effect - ϕ ; (b) the spatial Governorate level effect - ψ ; (c) the total Governorate effect - $\gamma = \sigma(\phi\sqrt{1-\lambda}) + \psi\sqrt{\lambda/s}$; (d) the unstructured Country effect - η .

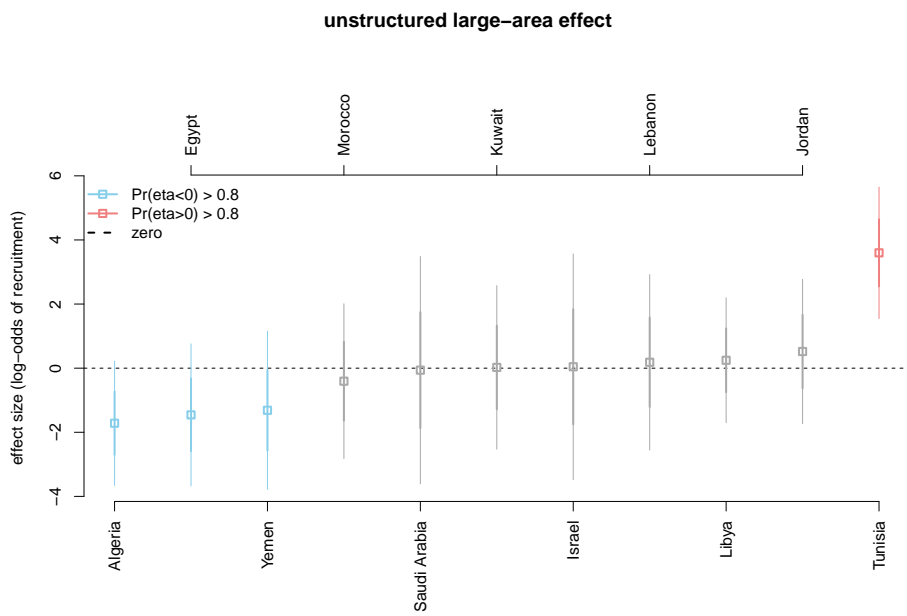


Figure I.4: Unstructured large-area effect η for the ‘Bird’s Eye’ model.

Figure I.6 presents the Egypt and Tunisia fully-connected graphs used to derive the district-level adjacency matrices fed to the ICAR model. Again, a small number of adjustments were made to connect islands and ensure full-connectivity.

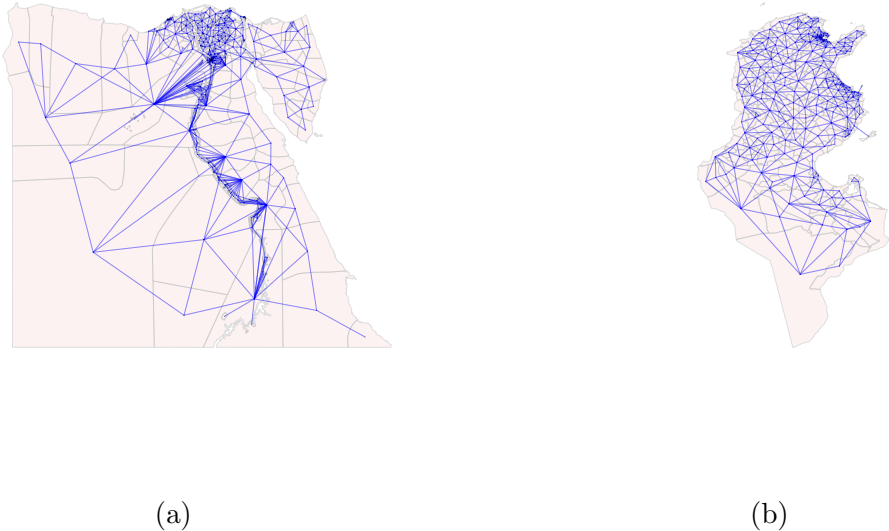


Figure I.6: Fully-connected graphs of (a) Egypt and (b) Tunisia at the District level.

The residual plots in Figure I.8, along with the Moran’s I presented in Figure I.7, convincingly show we have extracted all spatial variance from the observations: the resulting Moran’s Is are distributed around the null-value.

Figures I.9 and I.10 present the spatial distribution of point-estimates for the District and Governorate effects of Egypt and Tunisia respectively. The spatial distribution for Egypt indicates a substantially heightened propensity of recruitment in northeastern regions. No similar pattern is evident in Tunisia, though the mid-eastern costal areas do display systematically lower spatial recruitment effects than the rest of the country. Both countries estimate a number of highly significant district-level effects, which account for large portions of the variance in recruitment of both countries, with highly significant effects ranging from -5 to $+5$ log-odds points . In Tunisia, we also find evidence of a negative Sfax effect. Clearly, in order to be a recruit you must be subjected to unobserved area-level heterogeneity; individual-level covariates alone cannot counteract the underlying rarity of the event, as highlighted by the intercepts.

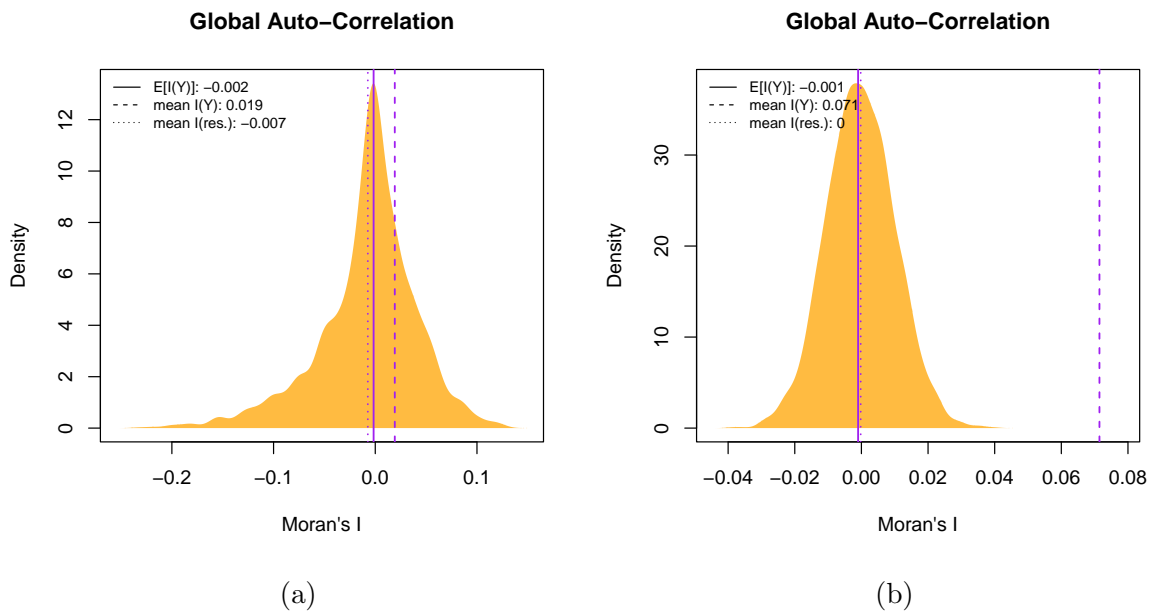


Figure I.7: Posterior distribution of Moran's I for Egypt (a) and Tunisia (b). The adjacency matrices implied by Figure I.6 are used as the weight matrices.

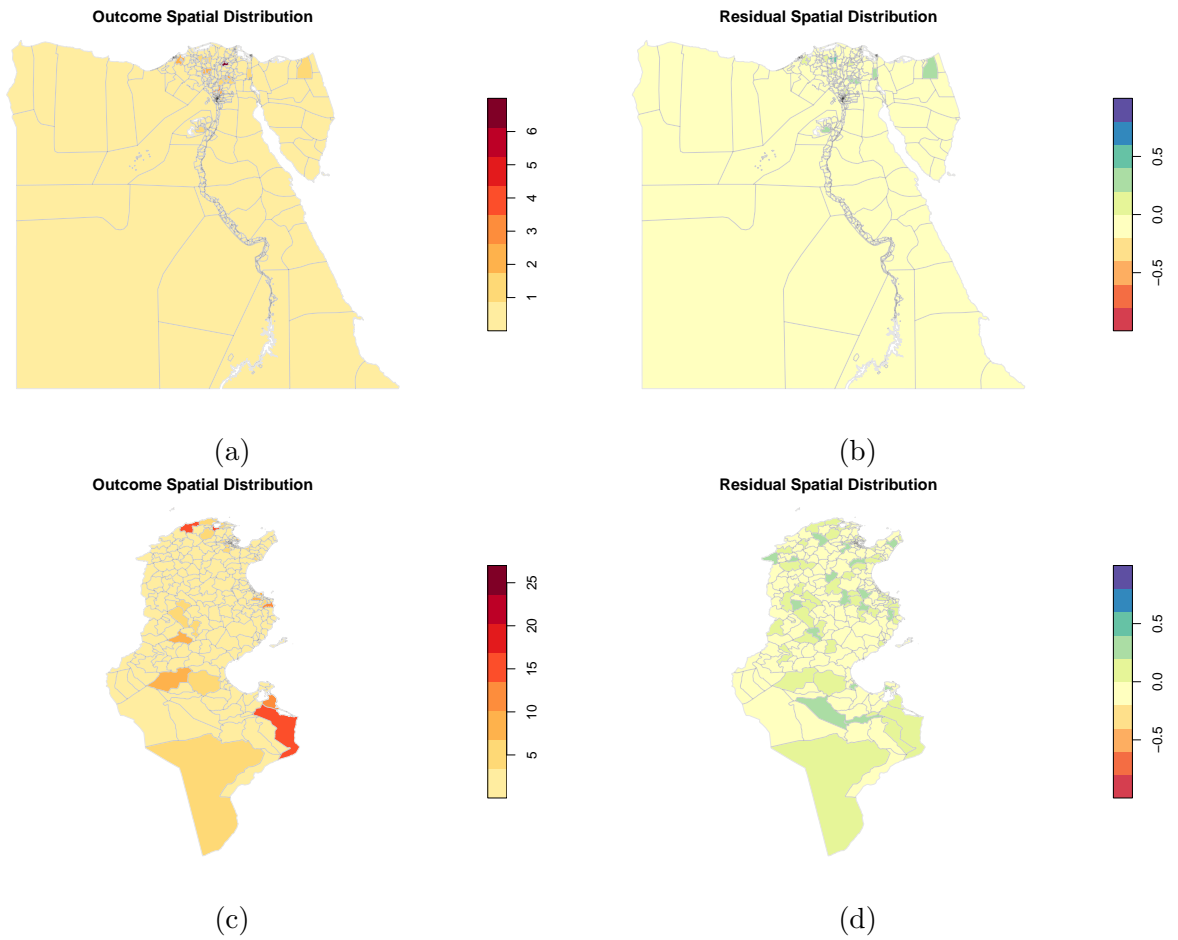
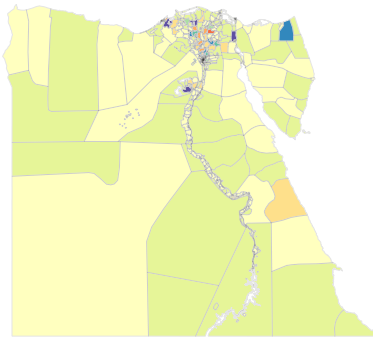


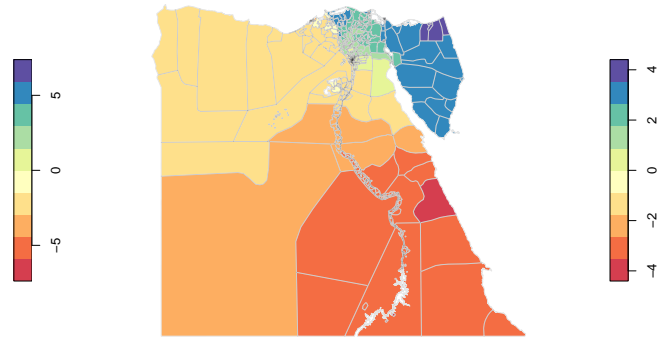
Figure I.8: Spatial distribution of Egyptian observations (a) and residuals (b); Tunisian observations (c) and residuals (d) at the District level. (a) and (c) present the spatial distribution of recruits.

Unstructured Small–Area Effect Spatial Distribution



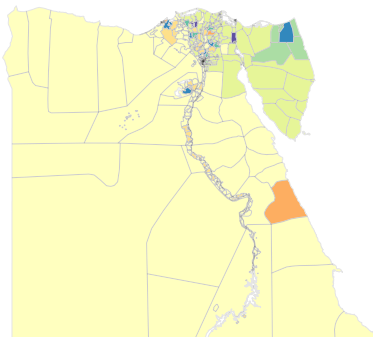
(a)

ICAR Small–Area Effect Spatial Distribution



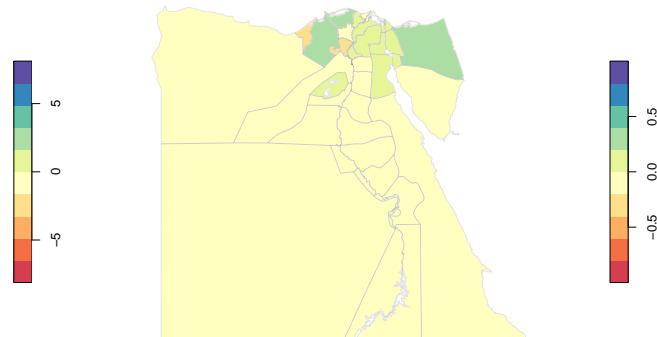
(b)

BYM2 Small–Area Effect Spatial Distribution



(c)

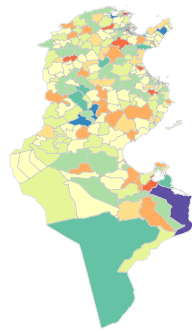
Unstructured Large–Area Effect Spatial Distribution



(d)

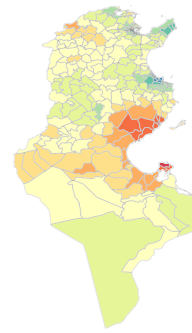
Figure I.9: Egypt’s Spatial distribution of: (a) the unstructured Governorate-level effect - ϕ ; (b) the spatial Governorate level effect - ψ ; (c) the total Governorate effect - $\gamma = \sigma(\phi\sqrt{1-\lambda}) + \psi\sqrt{\lambda/s}$; (d) the unstructured Country effect - η .

Unstructured Small–Area Effect Spatial Distribution



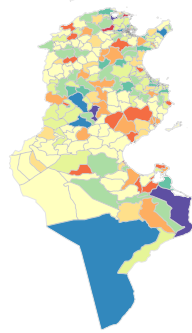
(a)

ICAR Small–Area Effect Spatial Distribution



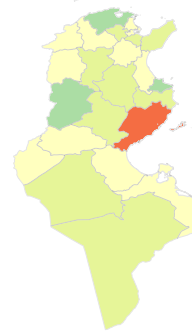
(b)

BYM2 Small–Area Effect Spatial Distribution



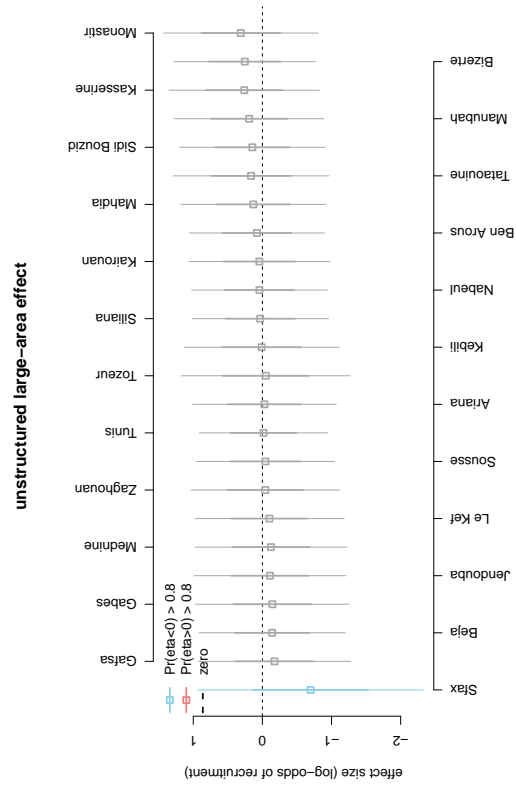
(c)

Unstructured Large–Area Effect Spatial Distribution

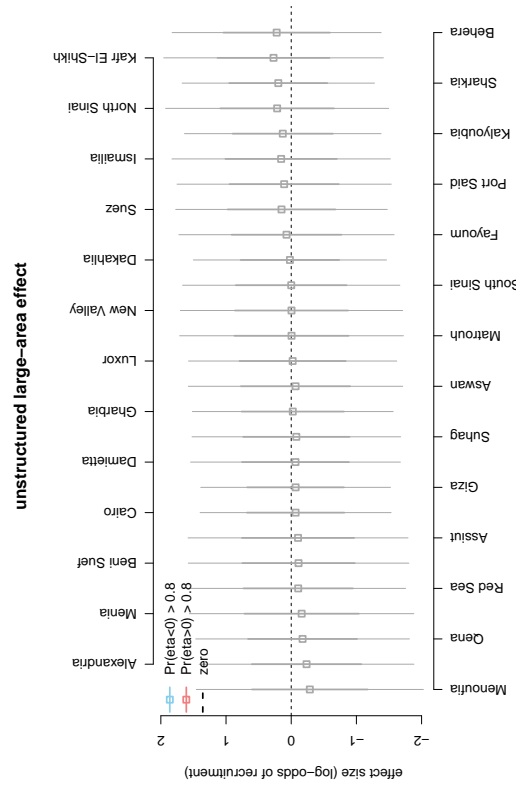


(d)

Figure I.10: Tunisia’s Spatial distribution of: (a) the unstructured Governorate-level effect - ϕ ; (b) the spatial Governorate level effect - ψ ; (c) the total Governorate effect - $\gamma = \sigma(\phi\sqrt{1 - \lambda}) + \psi\sqrt{\lambda/s}$; (d) the unstructured Country effect - η .



(b) Tunisia



(a) Egypt

Figure I.11: Governorate effect (η) ordered by proportion of posterior simulations above zero.

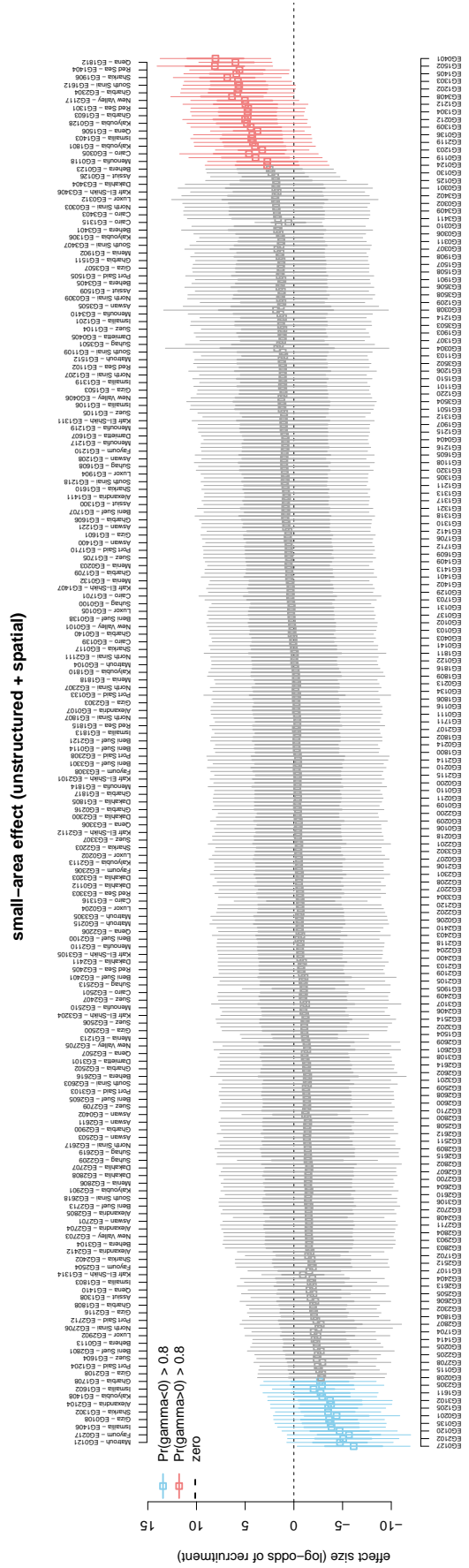


Figure I.12: Total (structured + spatial) residual District effects in Egypt, ordered by proportion of posterior simulations above zero.

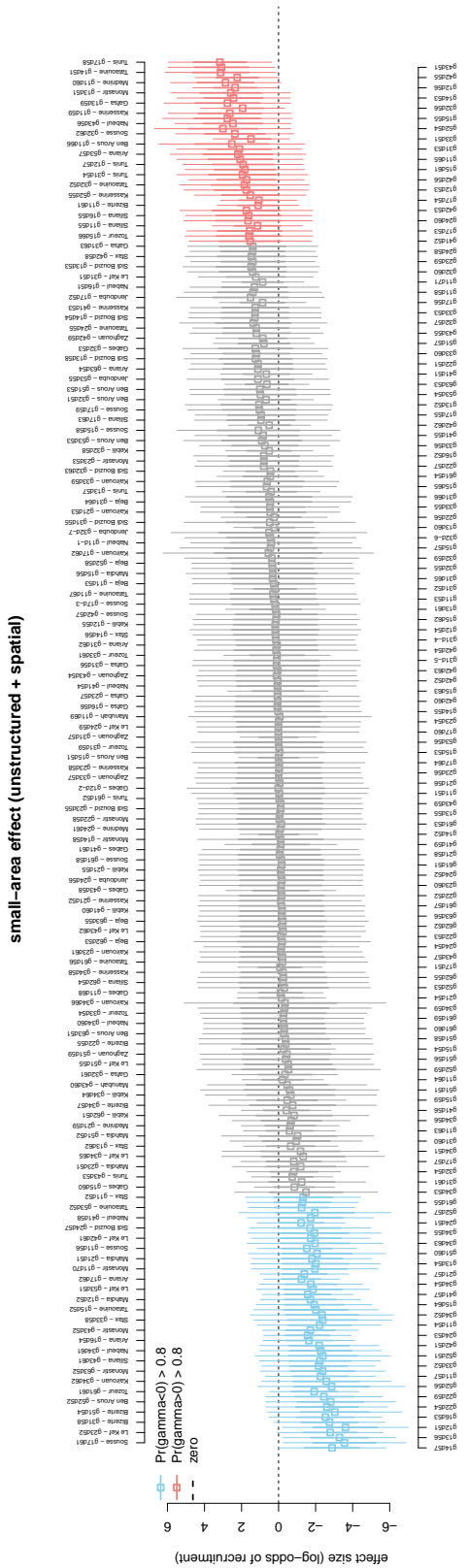


Figure I.13: Total (structured + spatial) residual District effects in Tunisia, ordered by proportion of posterior simulations above zero.

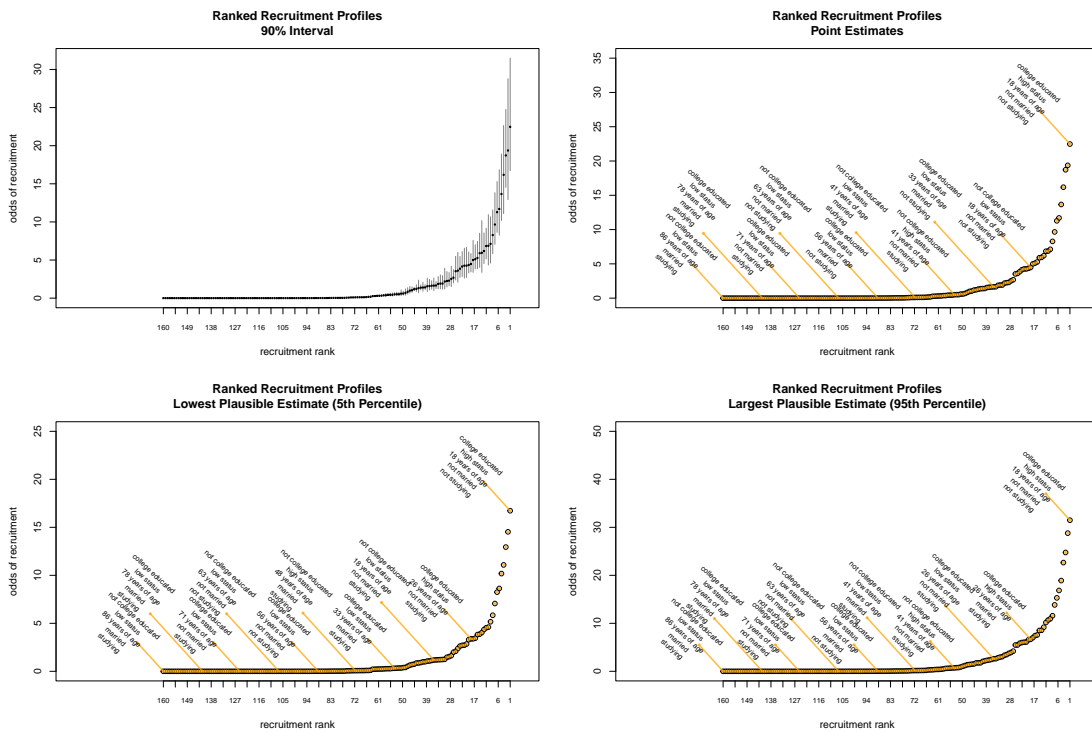


Figure I.14: Bird's Eye Distribution of predicted probabilities across a variety of hypothetical profiles. The distribution is presented on the odds relative to the average profile. To aid with interpretation, minimal and maximal estimates are presented separately. These plots help showcasing the sharp non-linearity across profile's recruitment propensities.

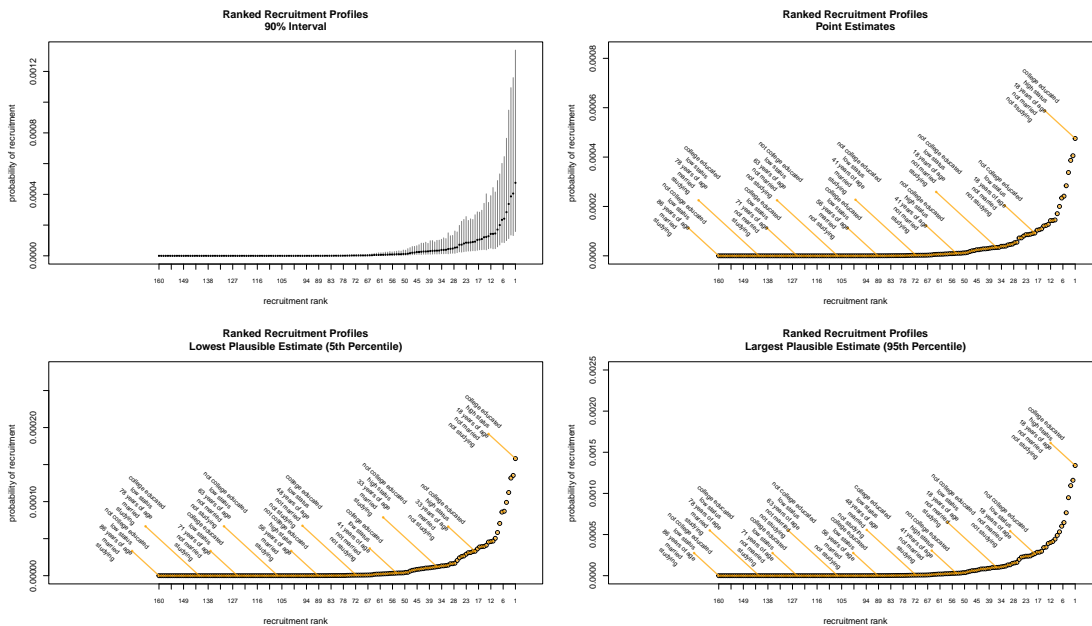


Figure I.15: Bird's Eye Distribution of predicted probabilities across a variety of hypothetical profiles. The distribution is presented on the probability scale. To aid with interpretation, minimal and maximal estimates are presented separately. These plots help showcasing the sharp non-linearity across profile's recruitment propensities.

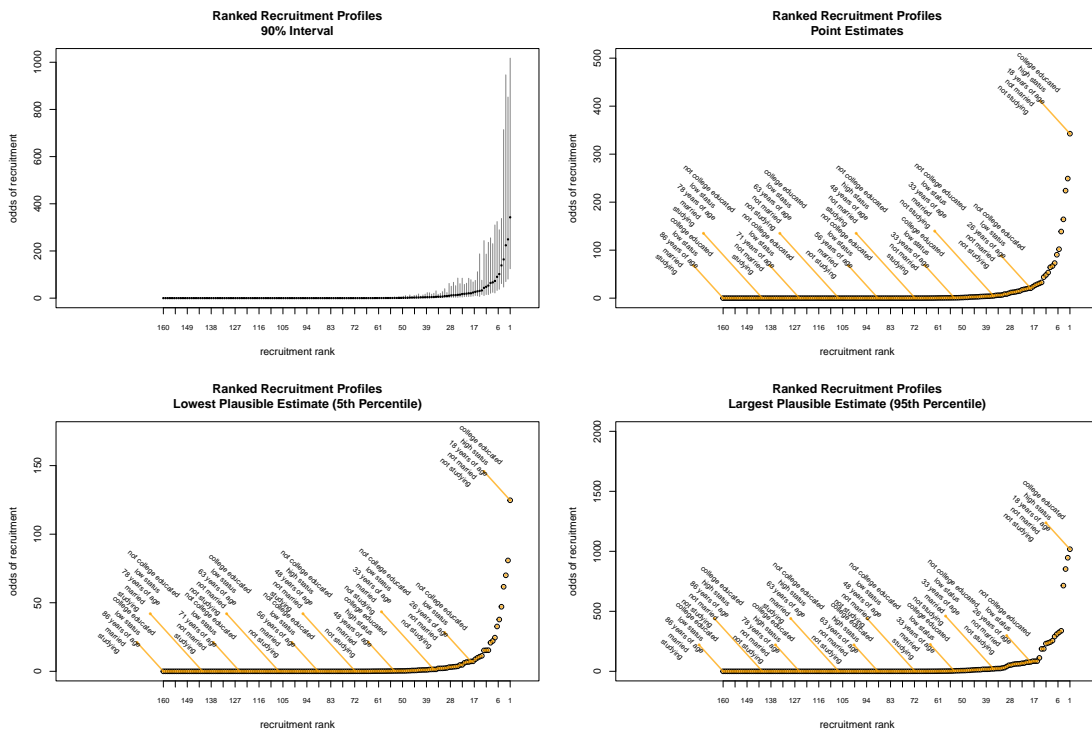


Figure I.18: Worm's Eye (Tunisia) distribution of the predicted probabilities across a variety of hypothetical profiles. The distribution is presented on the odds relative to the average profile. To aid with interpretation, minimal and maximal estimates are presented separately. These plots help showcasing the sharp non-linearity across profile's recruitment propensities.

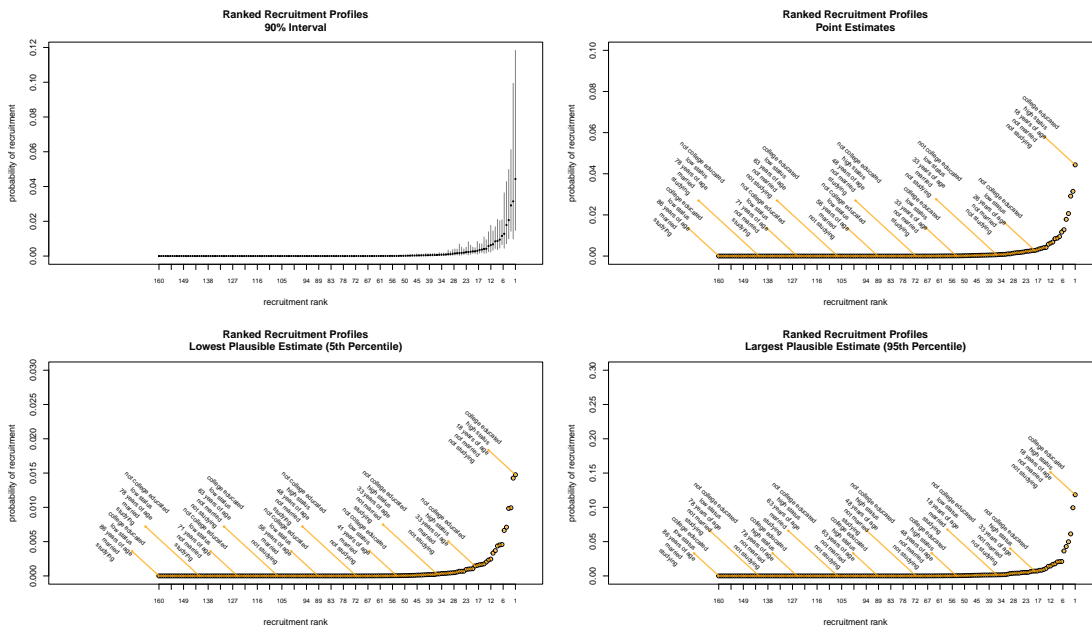


Figure I.19: Worm's Eye (Tunisia) distribution of the predicted probabilities across a variety of hypothetical profiles. The distribution is presented on the probability scale. To aid with interpretation, minimal and maximal estimates are presented separately. These plots help showcasing the sharp non-linearity across profile's recruitment propensities.

References

- Carpenter, B., et al. 2017. “Stan: A Probabilistic Programming Language.” *Journal of Statistical Software* 76 (1): 1–32.
- Cerina, R., and R. Duch. 2020. “Measuring Public Opinion via Digital Footprints.” *International Journal of Forecasting* 36 (3): 987–1002.
- Dodwell, B., D. Milton, and D. Ressler. 2016. “The Caliphate’s Global Workforce: An Inside Look at the Islamic State’s Foreign Fighter Paper Trail.” Technical report, Combating Terrorism Center.
- Gelman, A., and D. B. Rubin. 1992. “Inference from Iterative Simulation Using Multiple Sequences.” *Statistical Science* 7 (4): 457–472.
- Grewal, S., A. Yildirim, and S. Williamson. 2020. “Counting the Black Flags: Measuring Support for ISIS through Endorsement Experiments.” Working Paper.
- Hanretty, C., B. E. Lauderdale, and N. Vivyan. 2018. “Comparing Strategies for Estimating Constituency Opinion from National Survey Samples.” *Political Science Research and Methods* 6 (3): 571–591.
- Hoffman, M. D., and A. Gelman. 2014. “The no-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *Journal of Machine Learning Research* 15 (1): 1593–1623.
- Kim, A. Y., and J. Wakefield. 2010. *R Data and Methods for Spatial Epidemiology: The SpatialEpi Package*. Seattle: Department of Statistics, University of Washington.
- Lauderdale, B. E., D. Bailey, J. Blumenau, and D. Rivers. 2020. “Model-Based

Pre-Election Polling for National and Sub-National Outcomes in the US and UK.” *International Journal of Forecasting* 36 (2): 399–413.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. “Winbugs—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility.” *Statistics and Computing* 10 (4): 325–337.

Park, D. K., A. Gelman, and J. Bafumi. 2004. “Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls.” *Political Analysis* 12 (4): 375–385.

Piironen, J., and A. Vehtari. 2017. “On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior.” In *Artificial Intelligence and Statistics*, 905–913. Fort Lauderdale: PMLR.

Plummer, M. 2003. “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.” In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, 1–10. Vienna.

Rosenblatt, N. 2018. “All Jihad Is Local: What ISIS’ Files Tell Us about Its Fighters.” Technical report I, New America.

Vats, D., and C. Knudson. 2021. “Revisiting the Gelman–Rubin Diagnostic.” *Statistical Science* 36 (4): 518–529.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. 2021. “Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC (with Discussion).” *Bayesian Analysis* 16 (2): 667–718.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31 (3):

980–991.