Multilevel Models for Estimating the Number of Deaths in Armed Conflict (in Colombia)

Shira Mitchell

JSM 2014

Collaborators: Brent Coull, Alan Zaslavsky, Al Ozonoff, Kristian Lum, Patrick Ball, Megan Price

August 5, 2014

Colombian conflict (1964–present)

From Wikipedia, the free encyclopedia

For other Colombia-related conflicts, see List of wars involving Colombia.

This article has multiple issues. Please help **improve it** or dis
talk page.

- This article **needs additional citations for verification.** *(Ju*
- This article is **outdated.** *(June 2013)*

The **Colombian conflict** began approximately in 1964 or 1966 and is an ongoing low-intensity asymmetric war between the Colombian government, drug gangs, paramilitary groups and left-wing guerrillas such as the Revolutionary Armed Forces of Colombia, and the National Liberation Army (ELN), fighting each other to increase their influence in Colombian territory.[18][19][20][21][22][23][24][25]

Casualties and losses

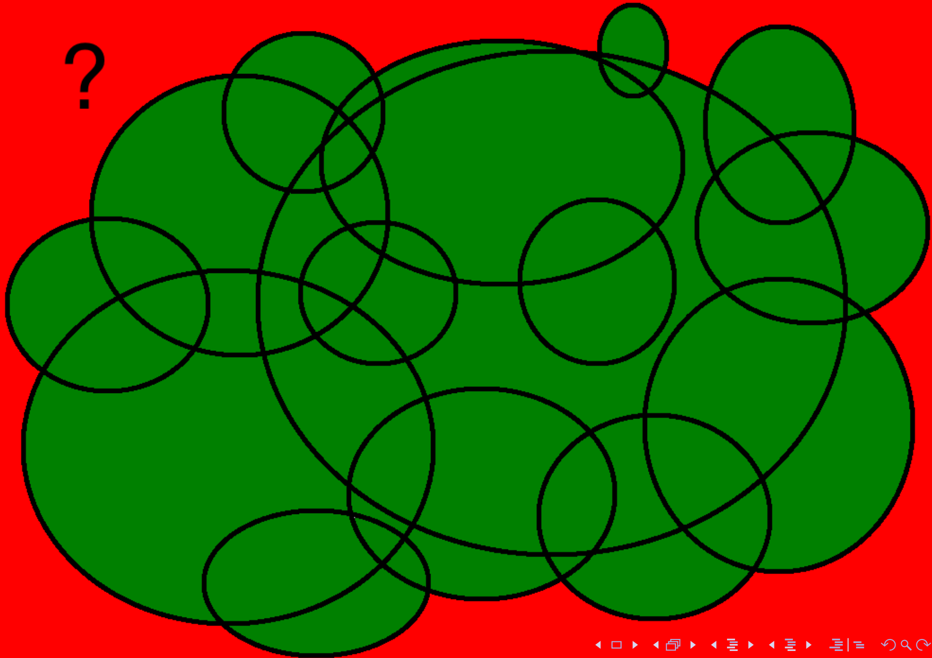
Army and Police: 4,286 killed, 13,076 injured (since 2002[6])	FARC: 12,981 demobilized (since 2002[6]) ELN: 2,789 demobilized (since 2002[6]) Since 2002, 34,512 guerrillas captured, 13,197 killed[6]

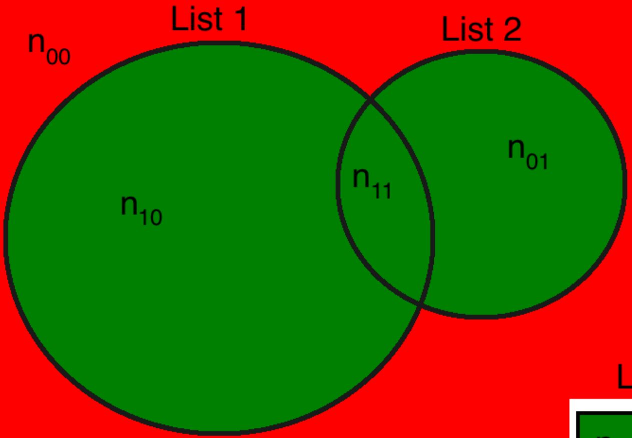
total casualties=50,000–200,000[16]

Data from the Human Rights Data Analysis Group (HRDAG)

NGOs and govt groups provide lists of killings in 1998-2007, Casanare, Colombia: department of Colombia, population 300,000, BP oil pipeline, much corruption and violence.

Goal: Estimate the number of killings in Casanare in years 1998-2007.

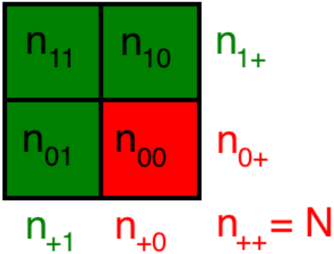




$$n_{10} + n_{11} + n_{01} = n$$

List 2

List 1			
	n_{11}	n_{10}	n_{1+}
	n_{01}	n_{00}	n_{0+}
	n_{+1}	n_{+0}	$n_{++} = N$



$$n_{k_1 k_2} \sim \mathrm{Pois}(\mu_{k_1 k_2}) \text{ independent}$$

$$\log(\mu_{k_1 k_2}) = \lambda_0 + \lambda_1 k_1 + \lambda_2 k_2$$

$$\Rightarrow \widehat{E[N]}_{MLE} = \frac{n_{1+} n_{+1}}{n_{11}}$$

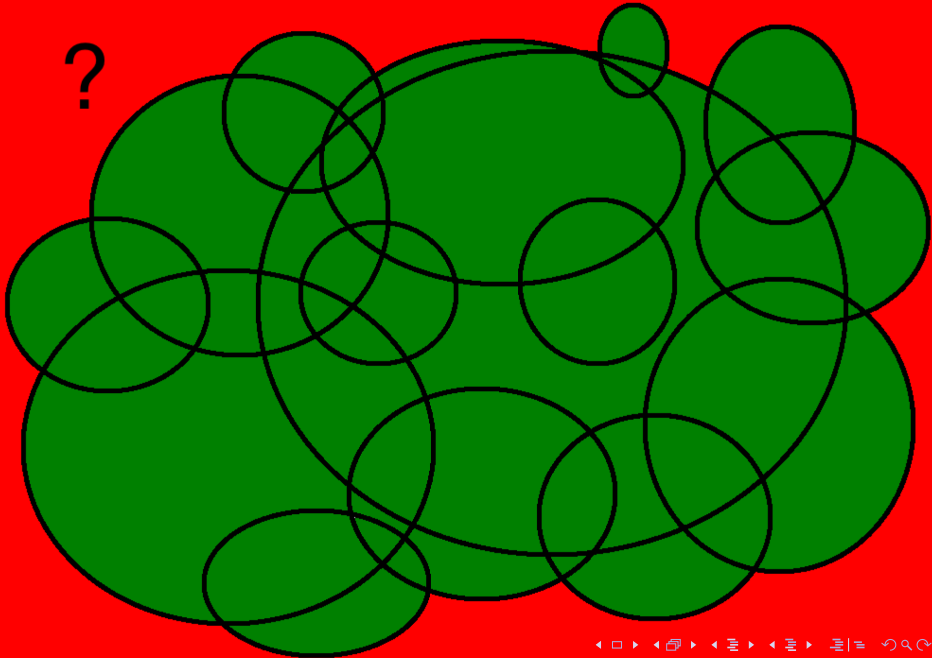
General number of lists

$$\mathbf{k} = (k_1, k_2, ..., k_J)$$

for example, if in lists 3, 4, and 6

$$= (0, 0, 1, 1, 0, 1)$$

Independence model: $\log(\mu_{\mathbf{k}}) = \lambda_0 + \lambda_1 k_1 + ... + \lambda_J k_J$



Data - from HRDAG

Matching:

Commission of Jurists		
Year	Gender	location
	⋮	
1998	male	TAMARA
	⋮	
	⋮	

National Police		
Year	Perpetrator	Gender
	⋮	
	⋮	
	⋮	
1998	FARC	male

6 lists contain among them 2619 observed killings.

Data - from HRDAG

year	IMLM (govt)	PN0 (govt)	VP (govt)	CCJ (NGO)	CIN (NGO)	CCE (NGO)
1998	1	0	0	14	13	3
1999	2	0	0	6	8	2
2000	213	0	5	22	23	0
2001	262	0	2	21	12	0
2002	268	1	0	33	9	0
2003	348	274	2	12	11	0
2004	412	324	295	14	11	1
2005	210	155	138	8	13	16
2006	104	71	26	3	2	15
2007	54	0	33	27	36	35

We're far from the independence model

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

Heterogeneity of a person's recordability

$$P_j(\theta) = \mathrm{P}(\text{person with recordability } \theta \text{ is recorded on list j})$$

$$\log\left(\frac{P_j(\theta_{govt})}{1 - P_j(\theta_{govt})}\right) = \theta_{govt} + \lambda_j \text{ for } j \in govts$$

$$\log\left(\frac{P_j(\theta_{NGO})}{1 - P_j(\theta_{NGO})}\right) = \theta_{NGO} + \lambda_j \text{ for } j \in NGOs$$

Heterogeneity of a person's recordability

Let $(\theta_{NGO}, \theta_{govt}) \sim p(\theta_{NGO}, \theta_{govt})$.
Then

$$\log(\mu_{\mathbf{k}}) = \lambda_0 + \lambda_1 k_1 + \ldots + \lambda_6 k_6 + \gamma(k_+^{NGO}, k_+^{govt})$$

$$\log(\mu_k) = \lambda_0 + \lambda_1 k_1 + ... + \lambda_6 k_6 +$$

$$\sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govts} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govts} \omega_{mix} k_j k_{j'}$$

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

$$\log\left({\mu_k}^{(t)}\right) = \lambda_{0t} + \lambda_{1,t} k_1 + \ldots + \lambda_{6,t} k_6 +$$
$$\sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govts} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govts} \omega_{mix} k_j k_{j'}$$

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

$$\log\left(\mu_k{}^{(t)}\right) = \lambda_{0t} + \lambda_{1,t} k_1 + \ldots + \lambda_{6,t} k_6 + \sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govts} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govts} \omega_{mix} k_j k_{j'}$$

$$\lambda_{j,t} \sim N(\mu_j, \tau^2) \text{ for } j = 1, \ldots, 6$$

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

AR1 Model

$$\begin{bmatrix} \lambda_{j,1} \\ \vdots \\ \lambda_{j,T} \end{bmatrix} \mid \mu_j, \rho, \tau^2 \sim N\left(\begin{bmatrix} \mu_j \\ \vdots \\ \vdots \\ \mu_j \end{bmatrix}, \begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & \ddots & \dots & \dots \\ \vdots & & & \\ \rho^{T-1} & & & 1 \end{bmatrix} \tau^2 \right).$$

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

Mixture Model

$$\lambda_{j,t} \mid \gamma_{j,t} \sim (1-\gamma_{j,t})N(\mu_{inactive}, \sigma^2_{inactive}) + \gamma_{j,t}N(\mu_j, \tau^2) \text{ for } j = 1, ..., 6$$

$$\gamma_{j,t} \sim \text{Bern}(p) \text{ independent}$$

$$p \sim \text{Unif}(0, 1)$$

- Heterogeneity of a person's recordability
- Groups collecting data interact
- Want yearly estimates, but very little data exist in some years
- Groups operating in different but overlapping time periods

Missing Data

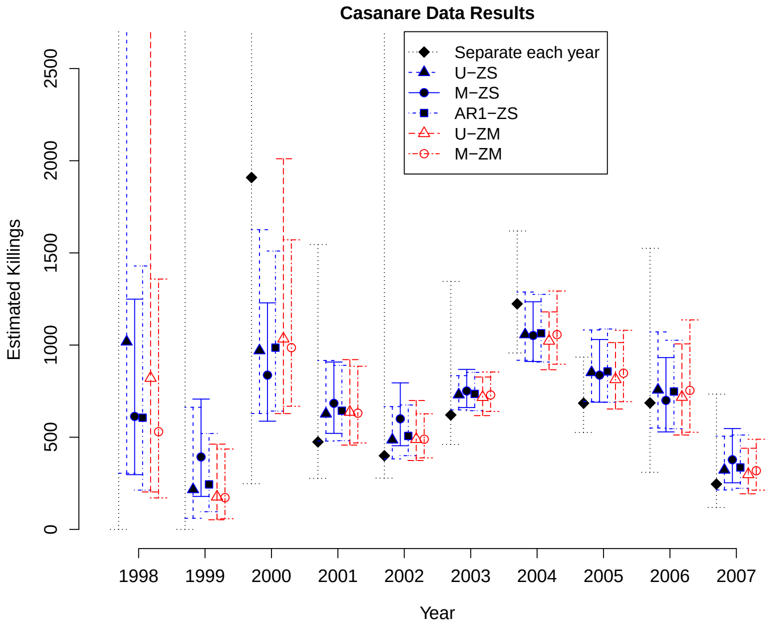
Consider inactive lists as missing data [Zwane et al., 2004].

In year t, lists 3 and 4 are inactive.

Treat $n^{(t)}_{01000}$ as margin $n^{(t)}_{01++0}$, and cells $n^{(t)}_{01000}$, $n^{(t)}_{01010}$, $n^{(t)}_{01100}$, $n^{(t)}_{01110}$ as missing data.

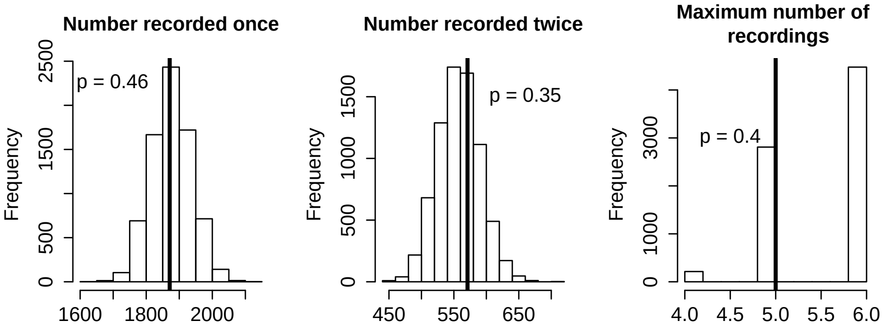
Zeros from missing data (ZM) vs *Zeros from sampling (ZS)*

	Zeros from missing data (ZM)	Zeros from sampling (ZS)
unpooled main effects (U)	U-ZM	U-ZS
Multilevel model (M)	M-ZM	M-ZS AR1-ZS



Casanare Data Results
Separate each year
U−ZS
M−ZS
AR1−ZS
U−ZM
M−ZM
Estimated Killings
0
500
1000
1500
2000
2500
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Year

Posterior Predictive Checks: M-ZS



Simulations

Simulate from posterior predictive distribution of M-ZM, M-ZS, and AR1-ZS fit to Casanare data.

Fit all the models.

- Coverage is similar for all models.
- Multilevel models have narrower intervals, and lower bias.

Recommendations

- In many applications, lists concentrate effort in different years, locations, or demographics.
- If these groups are overlapping $\Rightarrow$ fit joint models, to be able to model more list interactions, and to borrow information across strata.

We recommend Multilevel Models

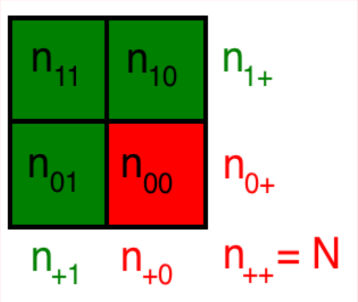
- In years with little data, we might not trust unpooled estimates - high variance, likely to get extreme estimates.
- Exchangeability and normality can be assessed via posterior predictive checks, relaxed by expanding the model.
- If we want monthly estimates at municipality-level, less and less data per stratum.
  - Colombia (2003-2011)
  - Syria (2011-2013)

References I

F Dominici. Combining contingency tables with missing dimensions. *Biometrics*, 56(2):546–553, 2000.

A Gelman, A Jakulin, M G Pittau, and Y Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

E N Zwane, K van der Pal-de Bruin, and P G M van der Heijden. The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, 23:2267–2281, 2004.



$$n_{11} | n_{1+}, n_{+1}, N \sim \text{HGeom}(n_{1+}, N - n_{1+}, n_{+1})$$

$$\hat{N}_{MLE} = \left\lfloor \frac{n_{1+} n_{+1}}{n_{11}} \right\rfloor$$

EM-like algorithm

E step:

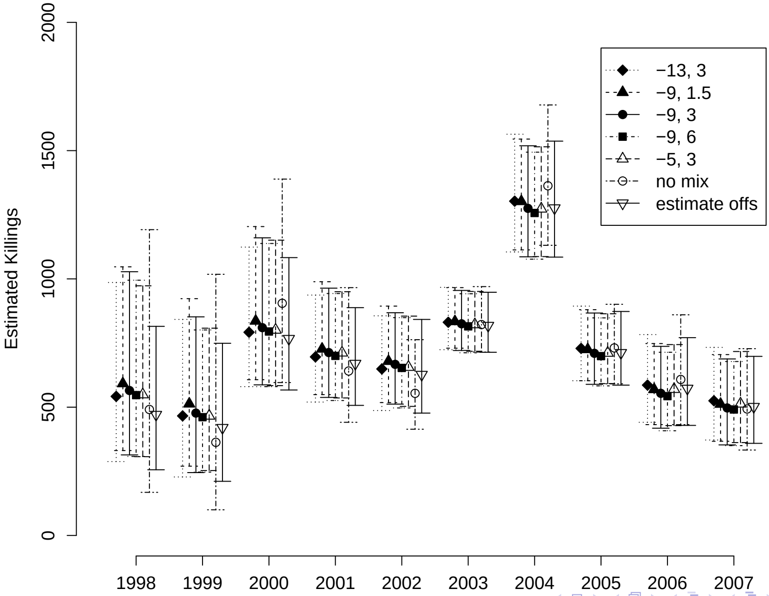
$$\hat{n}^{(t)}_{01010} = \frac{\sum_{s=1}^{T} \mu^{(s)}_{01010}}{\sum_{s=1}^{T} \left(\mu^{(s)}_{01000} + \mu^{(s)}_{01010} + \mu^{(s)}_{01100} + \mu^{(s)}_{01110} \right)} n^{(t)}_{01++0}.$$

M step:

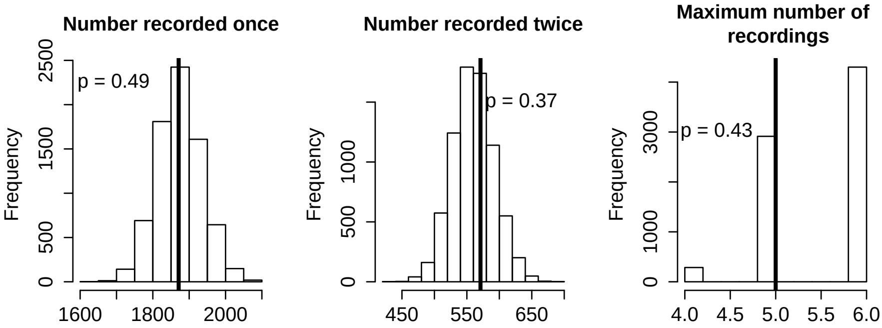
Fit log-linear model to completed data $\{n^{(t)}_{\mathbf{k}}\}_{\mathbf{k} \neq 00000, 00010, 00100, 00110}$.

Bayesian version [Dominici, 2000].

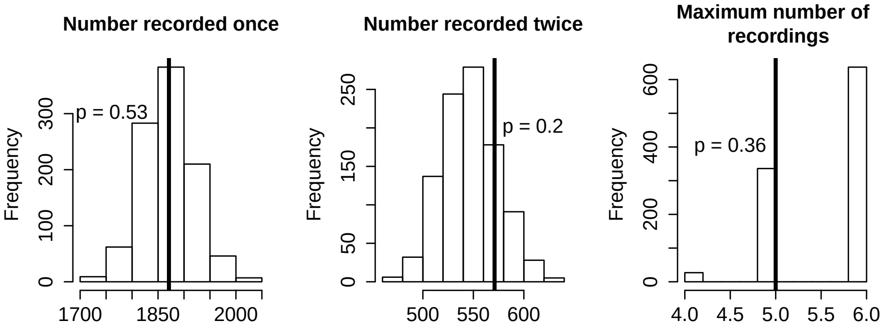
Sensitivity Analysis: Choice of $\mu_{inactive}$, $\tau^2_{inactive}$



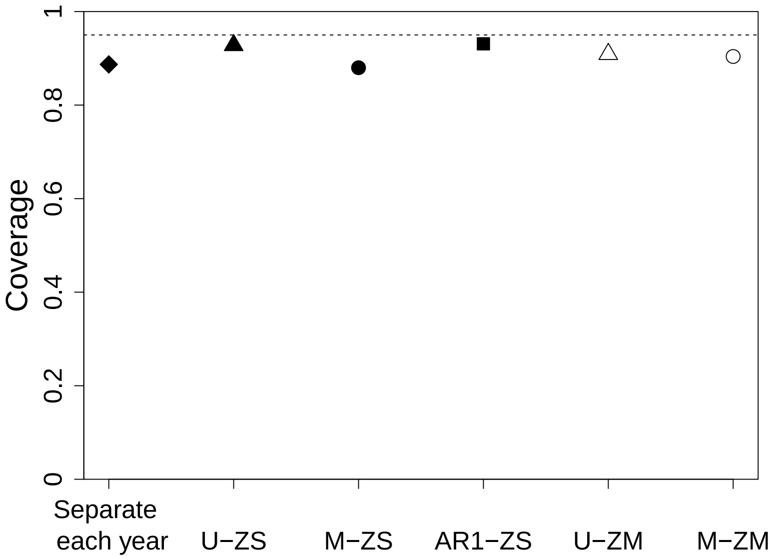
Posterior Predictive Checks: AR1-ZS



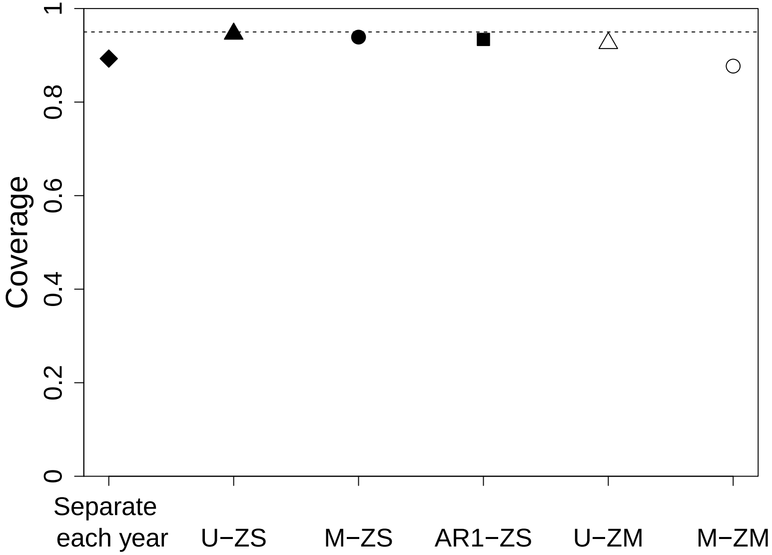
Posterior Predictive Checks: M-ZM



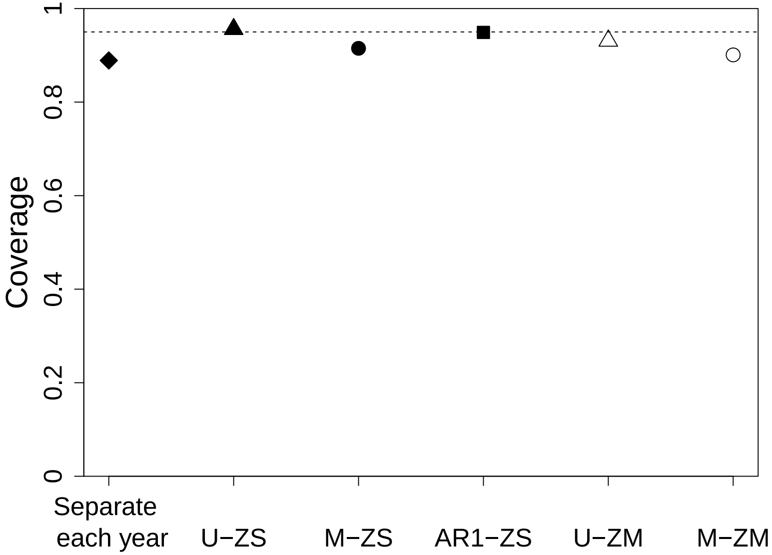
Generate data from M-ZM: Coverage



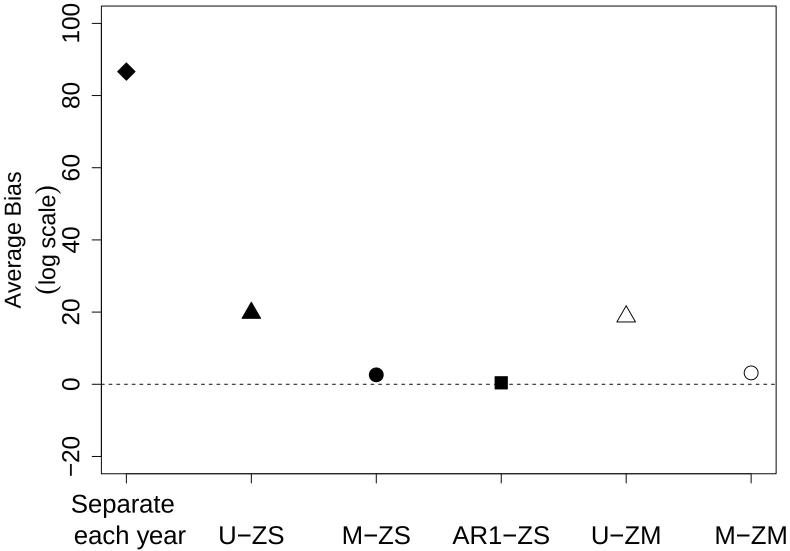
Generate data from M-ZS: Coverage



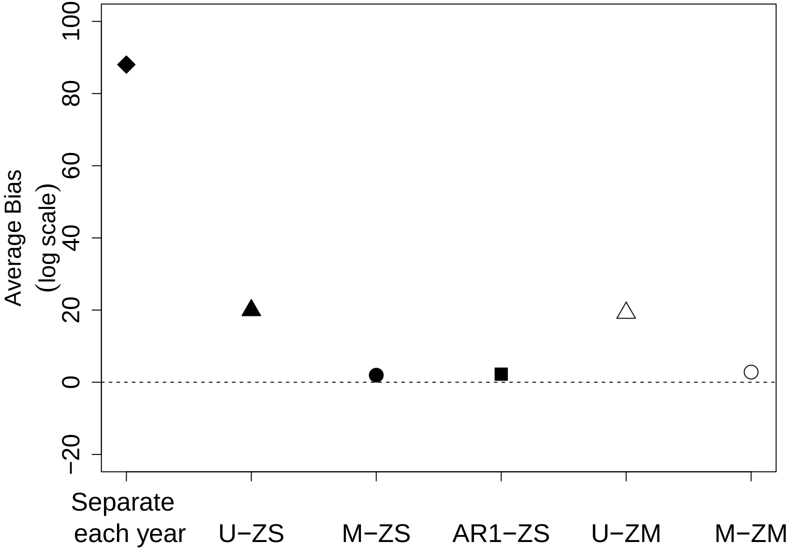
Generate data from AR1-ZS: Coverage



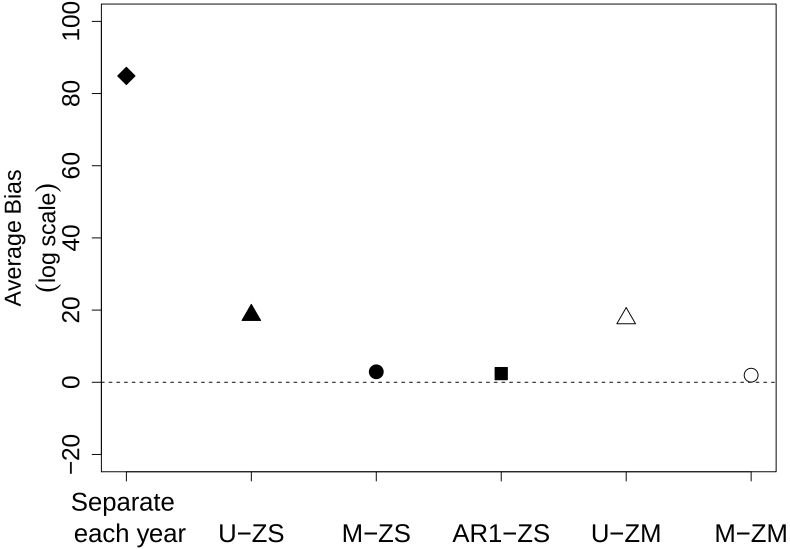
Generate data from M-ZM: Bias



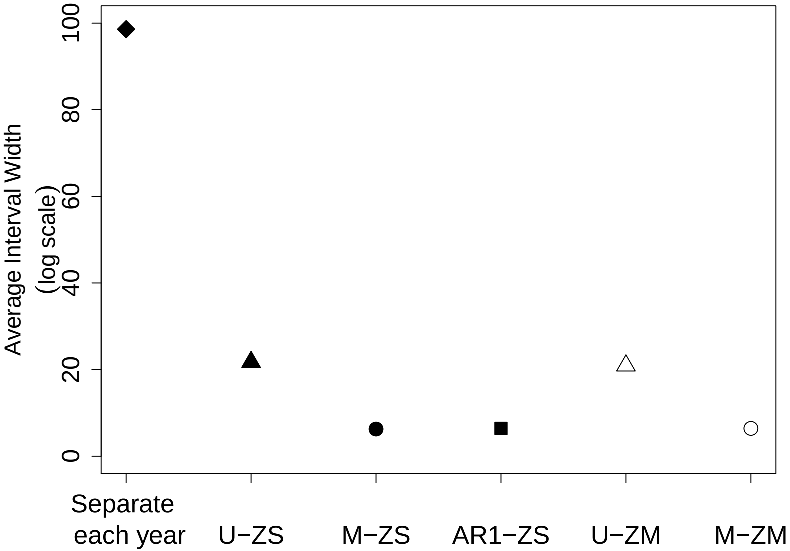
Generate data from M-ZS: Bias



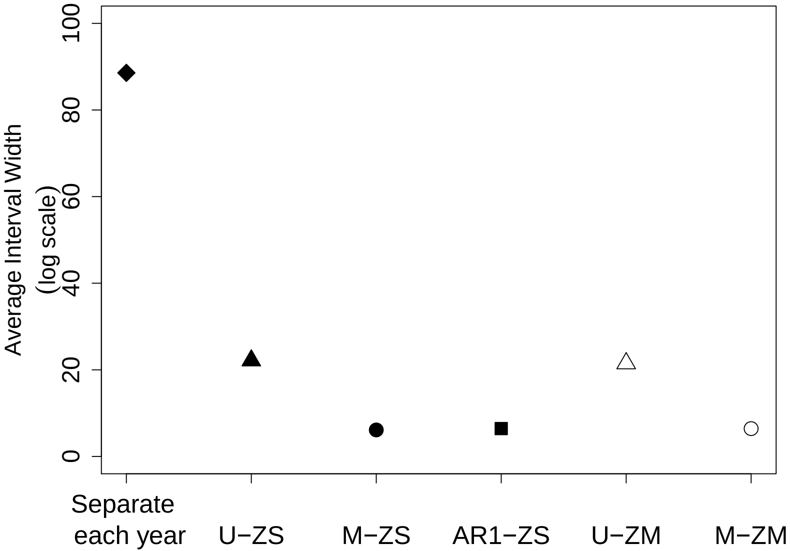
Generate data from AR1-ZS: Bias



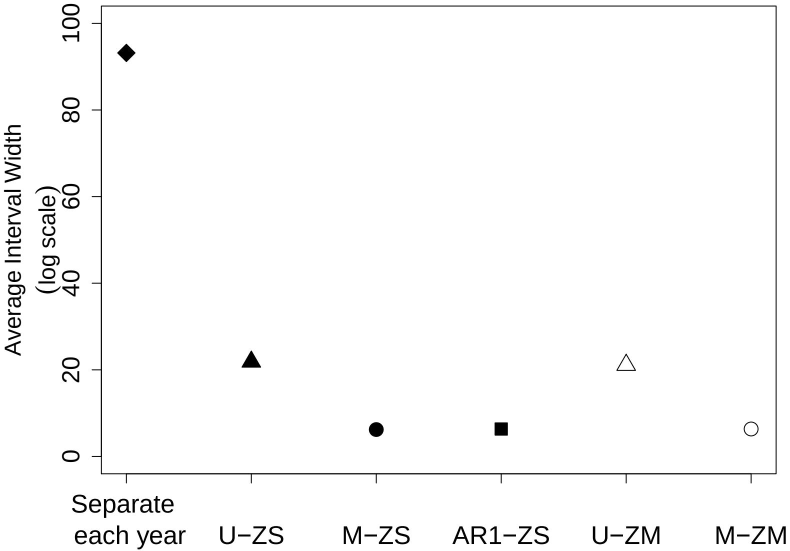
Generate data from M-ZM: Interval Width



Generate data from M-ZS: Interval Width



Generate data from AR1-ZS: Interval Width



Continuous Model Expansion

$$\log\left(\mu_k^{(t)}\right) = \lambda_{0t} + \lambda_{1,t} k_1 + \ldots + \lambda_{6,t} k_6 +$$

$$\sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govts} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govts} \omega_{mix} k_j k_{j'}$$

$$\omega_{j,j'} \sim N\left(\omega_{NGO}, \sigma^2_{NGO}\right)$$

Continuous Model Expansion: 3-way log-linear interactions

- Population heterogeneity $\Rightarrow$ higher-order interactions.
- Story for list cooperations?

Continuous Model Expansion: 3-way log-linear interactions

A Story:

Continuous Model Expansion: 3-way log-linear interactions

- Cauchy priors - regularization [Gelman et al., 2008]
- Exchangeable based on NGO/govt