Final Project; Categorical Data Analysis

Abby Smith, Miruna Baronschi, Martha Eichlersmith

2019-12-05

1 Estimates of Killing in Casanare

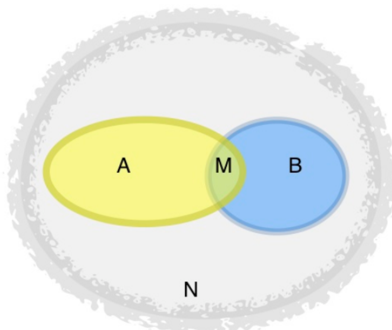
Casanare is a large, rural state in Colombia that includes 19 municipalities and a population of almost 300,000 inhabitants. Located in the foothills of the Andes, Casanare has a history of violence. Casanare (alongside other states, including Arauca) became pivotal to the Colombian government in the 1980s and 1990s when oil reserves were discovered there. Because of this natural resource, the government targeted opposition to what was perceived as the Colombian interest and modernization and the Union Patriotica (UP) in Casanare (as well as Arauca), who supported people's access to land and workers' rights, became the victim. Multiple armed groups operated in Casanare, including the Colombian military, paramilitaries, and guerrillas, going after "members" (both actual members and merely perceived supporters) of the UP. As a result, many Casanare citizens suffered violent deaths and disappearances. It can be argued that what was seen in Casanare (among other places) in Colombia was "political" genocide (Gomez-Suarez, 2007). But, one of the contentions against framing this episode in Colombian history as a genocide is the low death and disappearances toll.

So how many people have been killed or disappeared? We review the Human Rights Data Analysis Group (HRDAG)'s reporting on this issue of population estimation (Guzman et al., 2007). In this study, the authors used information about victims of killings and disappearances provided by 15 datasets .Any accounting of lethal violence will be incorrect if we assume that any one dataset or combination of datasets contains a comprehensive count of violent acts and disappearances. Registries of violent acts kept by governmental and non-governmental institutions contain some, but not all, of the records of lethal violence. Organizations collecting this data may only have access to certain subsets of a population or geographic areas

The datasets come from state agencies – including government, security, forensic and judicial bodies – and from civil society organizations. Across these 15 datasets, there are individuals that have been "captured" only by one dataset, and some that have been captured by multiple. How can we disambiguate the patterns of violence?

2 Multiple Systems Estimation (MSE), Capture Re-Capture

The goal is to estimate the overall population of victims by first estimating the victims who are not captured by any of the datasets. MSE estimates the total number of disappearances by comparing the size of the overlap(s) between lists to the sizes of the lists themselves. If the overlap is small, this implies that the population from which the lists were drawn is much larger than the lists. If, on the other hand, most of the cases on the lists overlap, this implies that the overall population is not much larger than the number of cases listed.



Equations for deriving unknown population N, two systems.

$$\frac{A}{N} \times \frac{B}{N} = \frac{M}{N}$$

$$\frac{AB}{N^2} = \frac{M}{N}$$

$$ABN = MN^2$$

$$AB = MN$$

$$\hat{N} = \frac{AB}{M}$$

Figure 1: Multiple System Estimation[1]

2.1 MSE Assumptions

There are several MSE assumptions that are need for when there are two “systems” (lists), but not more than that. The assumptions are:

1. *Closed system*: The population of interest does not change during the measurement period. This means that the object of measurement, whether that is a population of persons in a country or a population of violent events that occurred in a state, is a closed system: the target population does not change during the period of measurement. This assumption is generally un-problematical for data on violent events, because events that occurred cannot “un-occur” later.
2. *Perfect matching (record linkage!)*: The overlap between systems (i.e., the group of cases recorded in more than one list) is perfectly identified.
3. *Equal probability of capture*: For every data system, each individual has an equal probability of being captured. For example, every death has probability X of being recorded in list 1, every death has probability Y of being recorded in list 2, and so on. This assumption, the homogeneity of capture probability, is unlikely to hold for any type of violence data. For example, persons with fewer social connections may be both more likely to go missing and less likely to be reported missing; rural locations are more difficult to access than urban ones. Constructing two-sample estimates without accounting for different probabilities of capture leads to conclusions that may be biased.
4. *Independence of lists*: Capture in one list does not affect probability of capture in another list. For example, being reported to one NGO does not change the probability that an individual is reported to another. The third assumption, independence of systems, is similarly difficult to meet.

Like differences in capture probability, dependencies between systems are impossible to account for in the the two-system setting. A common example here is the difference between governmental and non-governmental organizations. Because different populations may have different levels of trust in the two organizations, reporting to one type of organization may imply that the witness is very unlikely to report to the other.

[1] Green, Amelia Hoover (2013) Multiple Systems Estimation: Stratification and Estimation. Available HTTP: https://hrdag.org/2013/03/20/mse-stratification-estimation/ (accessed 4 December 2019)

3 Overview of Data

Table 1: Contingency Table

	Organization	Total Captures	Unique	Type
d_CCJ.n	Colombian Commission of Jurists	214	48	judicial
d_EQU.n	Equitas	22	0	civil
d_FON.n	Fondelibertad	304	67	security
d_IMLD.n	National Institute of Forensic Medicine Disappearances	153	9	forensic
d_PN0.n	Policía Nacional	825	221	security
d_CIN.n	CINEP	267	91	civil
d_FAM.n	Families of Victims' Organizations	51	1	civil
d_FSR.n	Prosecutor General of Santa Rosa	151	0	security
d_IMLM.n	Instituto Nacional de Medicina Legal	1878	1219	forensic
d_VP.n	Vice Presidency Office	501	284	judicial
d_CCE.n	Colombia-Europe	72	30	civil
d_CTI.n	Technical Investigative Body of the Prosecutor Generals Office	36	0	judicial
d_FDC.n	Prosecutor General list of the Disappeared	623	376	forensic
d_GAU.n	Gaula	110	1	security
d_PL.n	Pais Libre	9	0	civil

3.1 Overall Trends

The following graph shows that violence has been concentrated in Yopal. We can see that the *reported* violence (i.e. the counts in the 15 datasets) appears to have intensified across all municipalities in Casanare in 2003-2004 and then dropped.

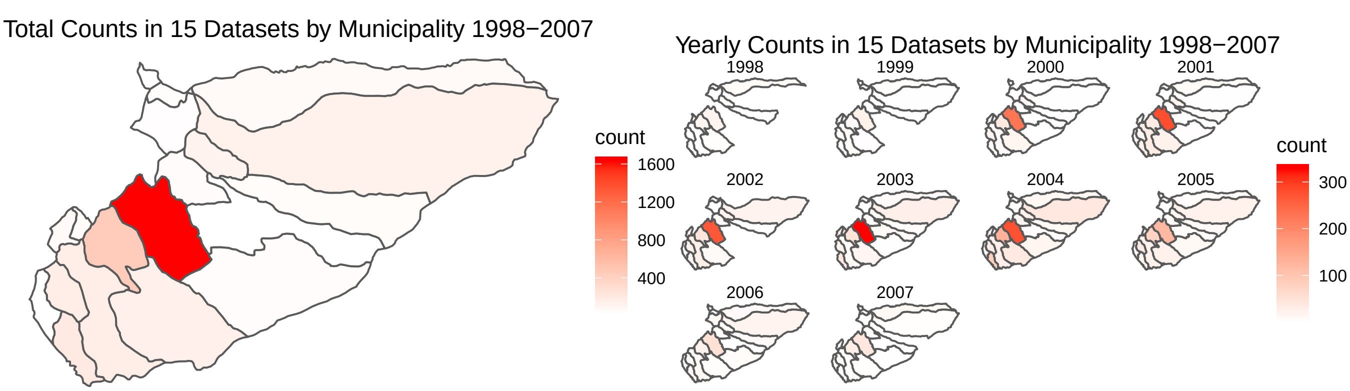


Figure 2: Overall Trends in 15 Datasets by Municipality

Looking just at the totals of the 15 datasets, that data suggest the violence peaked in 2003-2004 and the total number of victims is 3501. However, just looking at the trends among the totals of the 15 datasets does not give the whole picture.

3.2 Heterogeneity Issues

One of the requirements for MSE is for each individual has equal probability of capture for a given list. However, we can see from the figure below this does not seem to be the case. In 2004, the violence was centered around Villanueva, but all three of these lists have very little counts in that location. Since these individual list charts have different gradient trends than the total, that suggest that there is heterogeneity in capture probability. This will make our results more uncertain.

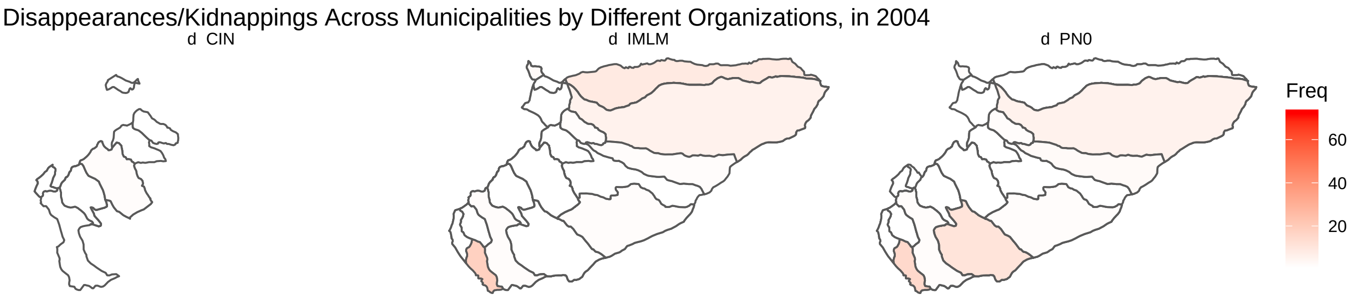


Figure 3: Counts from Three Different Datasets in 2004

3.3 Different Datasets - Different Stories

We motivate our need for a more sophisticated analysis by showing the reporting patterns of 3 different organizations across 1998-2007. If we relied on just one organization or even a combination of two we could tell different stories regarding the violence. Relying on FAM shows a peak in violent incidents in 2003, while reports by IMLM show a peak in 2004 . Using multiple systems estimation allows us to parse both the reported and unreported violent in Casanare.

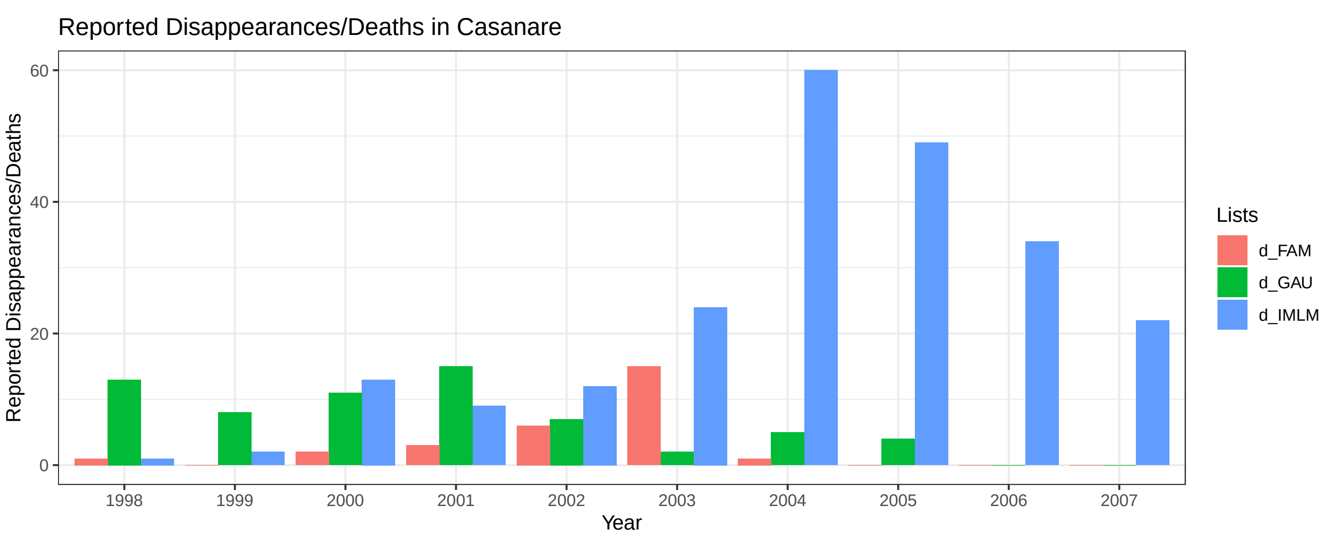


Figure 4: Count Trends for Three Organizations

4 Loglinear Modelling

In order to do our estimation, we will use loglinear modeling. We will describe the model as if we had two datasets (for simplicity). We know the victims "captured" by only dataset 1, only dataset 2, and both datasets. However, we don't know the amount of victims that are not captured by either list. Estimate this value allows us to estimate the total count of victims.

The count of victims captured into a dataset or combination of datasets is n_{11}, n_{10}, n_{01}, and n_{00}. Each of these cells is a count of victims captured. The subscripts denote which datasets a victim has been captured. The subscripts denotes if the victim has been captured (1), or not been captured by a certain data set. For n_{ij}, i is for dataset #1 and j is dataset #2.

Our estimates are primarily based on Poisson regression.

4.1 Estimating the Total Count of Victims

We are interested in estimating n_{00}, which also allows us to estimate the total number of victims.
The (log of the) expected cell count n_{00} is a function of the other observed cell counts, as shown in the equation below.

$$\log(n_{00}) = \alpha + \beta_1 \cdot \mathbb{1}(x \in n_{10}) + \beta_2 \cdot \mathbb{1}(x \in n_{01})$$

This is the saturated form of the log-linear models introduced in Bishop, Fienberg and Holland (1975). To quote from Agresti, "the saturated GLM has a separate parameter for each observation. It gives a perfect fit. This sounds good, but it is not a helpful model. It does not smooth the data or have the advantages that a simpler model has, such as parsimony. Nonetheless, it serves as a baseline for other models, such as for checking model fit."

When estimation of the total "population" of missing people in Casanare is the goal (as it typically is with multiple-systems estimation), the key value here is the intercept α. To estimate $\log(n_{00})$, all the other values in the model are zero, as the indicator functions for n_{ij} for zero. Therefore the only term that contributes to the estimate of $\log(n_{00})$ is α. The value of n_{00} is therefore the exponentiated value of a, that is, $\exp(\alpha)$. The total number of cases, N, is the sum of the observed cases plus $\exp(\alpha)$.

$$\begin{aligned}
\log(n_{00}) &= \alpha \underbrace{+\beta_1 \cdot \mathbb{1}(x \in n_{10}) + \beta_2 \cdot \mathbb{1}(x \in n_{01})}_{\text{each } = 0} \\
\log(n_{00}) &= \alpha \\
\hat{n}_{00} &= \exp\{\hat{\alpha}\} \\
\\
\hat{N} &= n_{11} + n_{10} + n_{01} + \hat{n}_{00}
\end{aligned}$$

As with any regression and data mining model, we want to avoid over-fitting. There is a trade off we need to balance between "goodness of fit", and simple (parsimonious) models.

We need to find the best model in order to get an accurate estimate of α. Thus, we should determine whether the full (saturated) model above, which assumes that all three two-way interactions between datasets are important, is actually necessary. There is one simpler models that assume the two-way interaction.

$$\log(n_{00}) = \alpha + \beta_1 \cdot \mathbb{1}(x \in n_{10}) + \beta_2 \cdot \mathbb{1}(x \in n_{01}) + \beta_{12} \cdot \mathbb{1}(x \in n_{11})$$

In this case, we can clearly write out the possible model. During model selection, we estimate these models and choose the model that minimizes the Bayesian Information Criterion (BIC), a test that weighs goodness of fit against degrees of freedom.

4.2 Challenges with Loglinear models

1. **Interpretation:** The inclusion of so many variables in loglinear models often makes interpretation very difficult.
2. **Independence Assumption:** The frequency in each cell is independent of frequencies in all other cells, which is not necessarily the case here. We attempt to model this.
3. **Sample Size Requirement:** With loglinear models, you need to have at least 5 times the number of cases as cells in your data. If you do not have the required amount of cases, then you need to increase the sample size or eliminate one or more of the variables.

4.3 Choosing our "Systems"

Our dataset encompasses 15 datasets, far too many to model with a loglinear model. We collapse these 15 datasets into 4 systems (i.e. groups) based on the type of organization that produced the dataset.

Table 2: Example of Information from Systems

security_ind	forensic_ind	judicial_ind	civil_ind	Freq
0	0	0	1	5
0	0	1	0	1
0	0	1	0	1
1	0	0	0	1
1	0	1	0	1
0	0	1	1	1

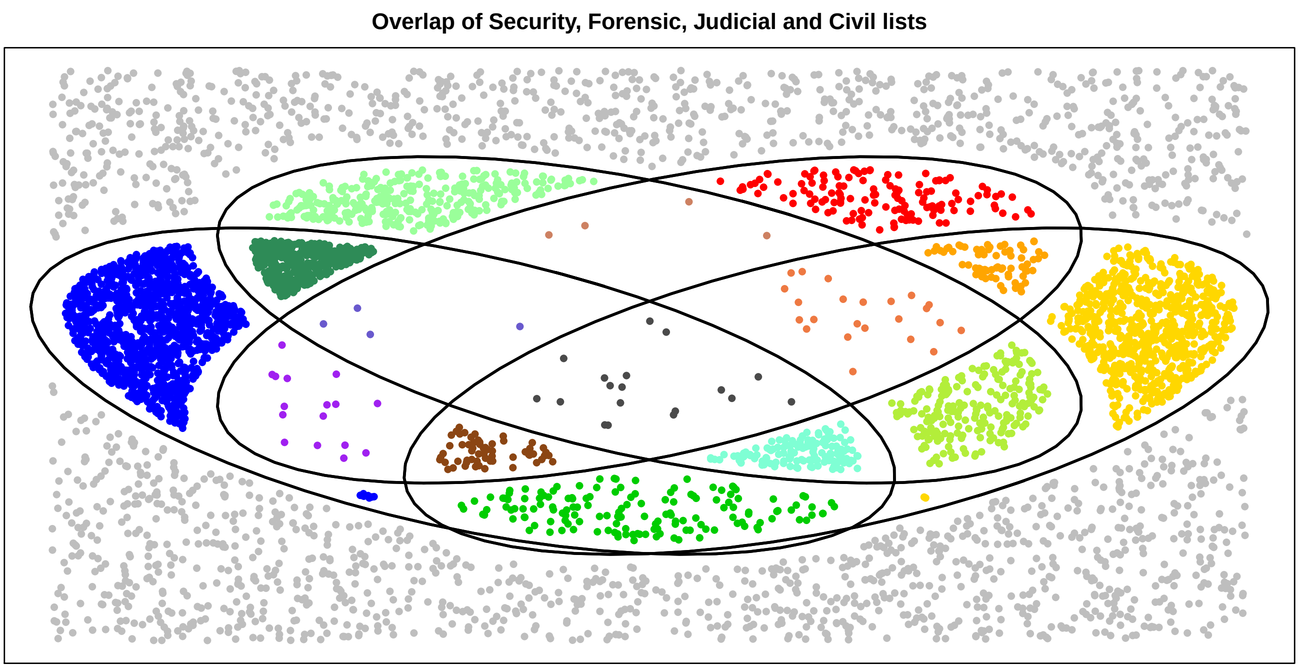


Figure 5: Estimated Venn Diagram of Systems

In the above Venn Diagram, the colored dots represent a specific victim. We are trying to estimate the victims that are not captured in one of the lists (shown by the graph dots). We can see there is not a lot of overlap, which suggests there may be lack of heterogeneity capture probability (i.e. the probability of a victim captured differs from list to list). In addition, this implies that there is likely a much larger population of victims than counted in the 15 datasets.

4.4 Model Definitions

4.4.1 Types of Models

Models will be denoted by M, and the subscripts will denote the type of model.

- M_0: The M_0 model is the simplest possible multiple source capture recapture model. It assumes that there is no heterogeneity and that all lists (civil, security, judicial, etc) have the same probability of capturing individuals. We know that this is not the case here.
- M_t: This model relaxes the M_0 model to allow for lists to have different capture rates.
- M_h: This model relaxes the M_0 model to allow for individual capture heterogeneity.
- M_{th}: This model allows for both list heterogeneity and capture events having different rates.

4.4.2 Types of Heterogeneity

When heterogeneity in capture probability is present (i.e. the probability of a list capturing a victims differs), there are different forms that this heterogeneity can take.

- Normal: The log odds of capture follows a Normal distribution.
- Darroch: The log odds of capture among those who were not captured follows a Normal distribution.
- Poisson: The log odds of capture among those who were not captured follows a Poisson distribution.
- Gamma: The log odds of capture among those who were not captured follows a Gamma distribution.

4.5 Non-Hierarchical Models

After collapsing the 15 lists into four systems, we fit several loglinear models. We see that the best fits clearly take into account both system and individual heterogeneity. We therefore choose the M_{th} model with the log odds of capture among those who were not captured follows a Gamma distribution. We see that the M_0 model demonstrates a clear lack of fit, which we would expect for this data. The models listed below are *not* hierarchical in nature.

We will need the hierarchical structure to perform model selection. It's important to note that a model is not chosen if it bears no resemblance to the observed data. The choice of a preferred model is typically based on a formal comparison of goodness-of-fit statistics associated with models that are related hierarchically (models containing higher order terms also implicitly include all lower order terms). Ultimately, the preferred model should distinguish between the pattern of the variables in the data and sampling variability, thus providing an intuitive interpretation.

The "number of captured units" is the number of observed elements, in this example, the number of people documented as missing/killed, we usually call this N_c (N for overall total, and c denoting captured. The "abundance" column shows the estimate of $\hat{N}$, the total population including the observed and the estimated unobserved deaths. The AIC and BIC columns show the "information coefficients" which balance the goodness of fit (shown in the "deviance" column) with the information used to estimate the model (degrees of freedom indicate this). Model selection often occurs based on the smallest AIC and BIC values, however this can only be done with hierarchical models which is not the case here.

Table 3: Summary of Models (Non-hierarchical models)

	abundance	stderr	deviance	df	AIC	BIC
M0	5604.929	100.607	2279.471	13	2376.736	2389.057
Mt	5232.814	87.309	529.823	10	633.087	663.891
Mh Chao (LB)	5970.687	139.893	2253.979	12	2353.244	2371.726
Mh Poisson2	6147.489	188.431	2262.122	12	2361.387	2379.869
Mh Darroch	6866.486	367.890	2257.937	12	2357.202	2375.684
Mh Gamma3.5	7725.102	634.443	2256.636	12	2355.901	2374.383
Mth Chao (LB)	5665.285	125.069	473.765	8	581.029	624.155
Mth Poisson2	6035.010	181.678	482.267	9	587.532	624.497
Mth Darroch	7204.688	411.304	476.057	9	581.322	618.287
Mth Gamma3.5	8782.672	810.694	474.715	9	579.980	616.945
Mb	4228.193	64.664	2107.184	12	2206.448	2224.931
Mbh	3643.007	47.707	1618.987	11	1720.252	1744.895

Each model controls for a subset of all the possible interactions among the models. In the context of MSE, the two- and three-way interactions estimate (and to some extent, control for) associations in the probabilities of capture between (and among) the lists. For example, is a certain person more likely to be seen on one list, and also more likely to be seen on a second list? If associations like this are present (and they usually are), they can bias the estimate.

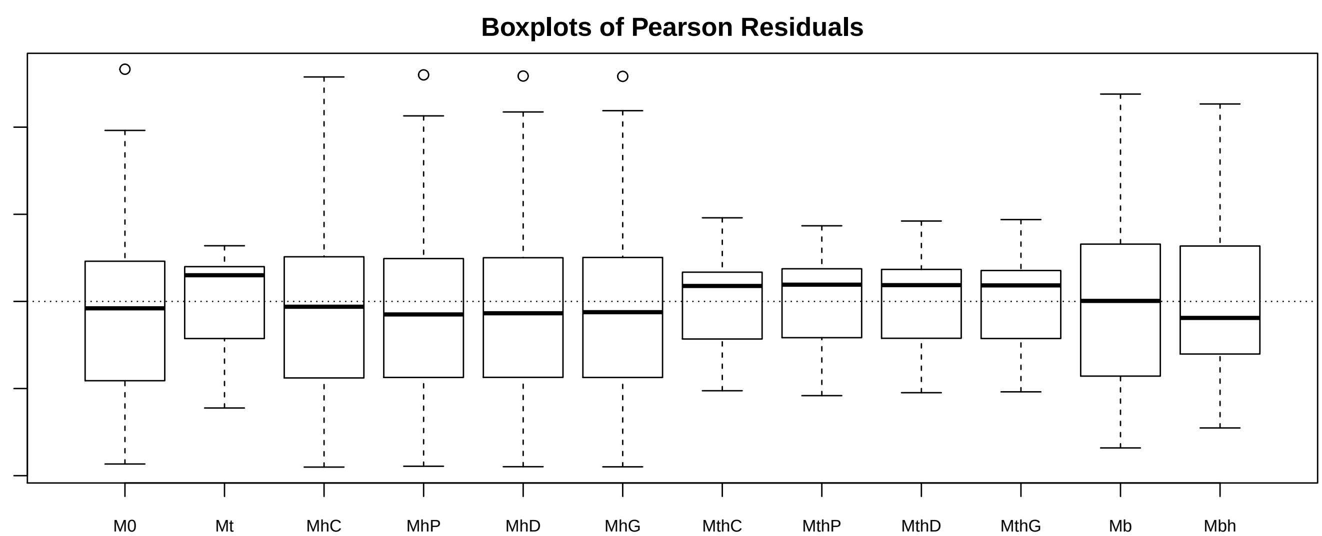


Figure 6: Boxplot of Residuals for Models

These boxplots of residuals offer a general assessment of model fit. These boxplots of residuals offer a general assessment of model fit.

Since the individual cell counts:

$$n_i \sim Pois(m_i)$$

$$E[n_i] = Var(n_i) = m_i$$

it follows that the Pearson residuals:

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{(m_i)}}$$

are approximately mean 0, variance 1. This is why they are sensible residuals to use.

The light dotted line represents zero, and ideally you want the residuals centered around zero. We see that there is significantly less variation in the M_{th} Models. These models account for list heterogeneity (probability of capture for a specific victim varies from list to list) and individual heterogeneity (probability of capture for a specific list varies from victim to victim). Models that account for heterogeneity offer less variation but also do not have residuals centered at zero.

4.6 Specific Model Results

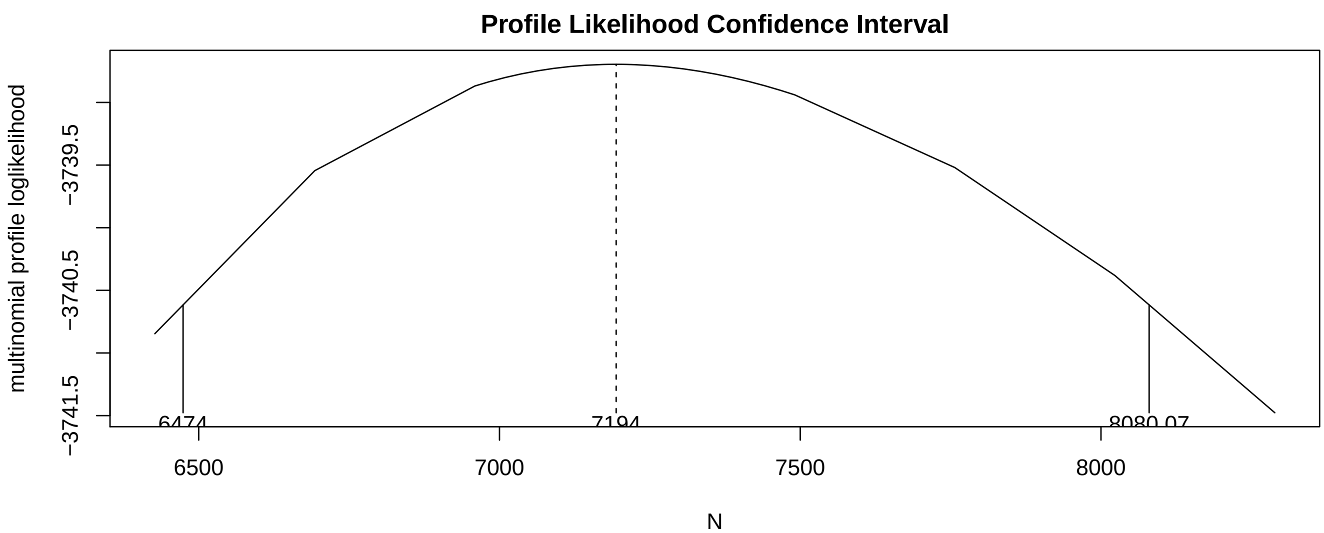


Figure 7: Profile Likelihood of Specific Model

Table 4: Confidence Interval of Specific Model

	abundance	InfCL	SupCL
Mth Darroch	7194	6474	8080.07

4.7 Capture Recapture

We display some basic capture-recapture frequency statistics to explore capture patterns. It displays, for $i = 1, ...t$, the number of people captured i times (f_i), the number of people captured for the first time on occasion i (u_i), the number of units captured for the last time on occasion i (v_i) and the number of units captured on occasion i (n_i). Here "occasion" should be interpreted as one of the four "systems"– where security is the 1st system, forensic is the 2nd, judicial is the 3rd, and civil is the 4th "occasion".

If the n_i statistics vary among capture occasions, there is a temporal effect– which we clearly see here. We would expect the top panel of the plot to be linear, while the bottom panel should be concave down or exhibit no pattern for capture patterns that are best fit with M_{th} models.

Table 5: Capture Recapture Statistics

	i = 1	i = 2	i = 3	i = 4
fi: number of units captured i times	2392	866	225	18
ui: number of units captured for the first time on occasion i	1128	1455	785	133
vi: number of units captured for the last time on occasion i	317	1640	1216	328
ni: number of units captured on occasion i	1128	2028	1387	328

$$\log\left(\frac{f_i}{\binom{t}{i}}\right) = \log\left(\frac{N \times P(i \text{ captures})}{\binom{t}{i}}\right) = \log(N(1-p)^{t-i}p^i = \log(N(1-p)^t) + i\log\left(\frac{p}{1-p}\right)$$

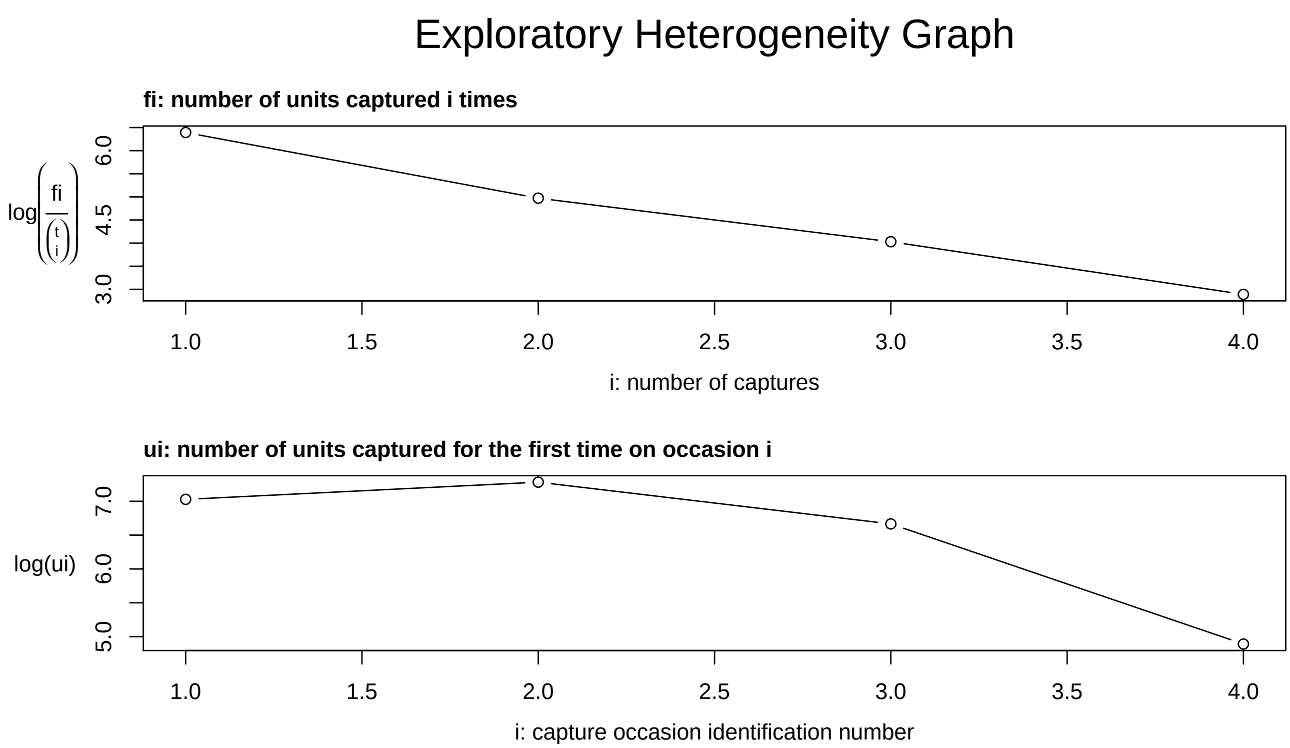


Figure 8: Capture-Recapture Frequency Statistics

4.8 Hierarchical Models

4.8.1 More Interactions: Better Fit

We start by fitting three simple models. The first, and simplest, is the model that assumes independence between the four systems. The second model looks at all possible two-way interactions. The third models looks at all three-way interaction between each of the systems. Recall that 1=Security, 2=Forensic, 3=Judicial, and 4=Civil.

$$\text{Model 1:} \quad \log(\hat{N}) = \alpha + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$$
$$\text{Model 2:} \quad \log(\hat{N}) = \alpha + \lambda_1 + \lambda_2 + \lambda_{13} + \lambda_{14} + \lambda_{23} + \lambda_{24} + \lambda_{34}$$
$$\text{Model 3:} \quad \log(\hat{N}) = \alpha + \lambda_1 + \lambda_2 + \lambda_{13} + \lambda_{14} + \lambda_{23} + \lambda_{24} + \lambda_{34} + \lambda_{123} + \lambda_{124} + \lambda_{134} + \lambda_{234}$$

We perform log-likelihood ratio test and see that the higher the order of the interaction, the better the fit (with saturated model signifying a perfect fit). The model deviance (LR statistic for testing against the saturated model) for each is defined as:

$$G^2 = 2\sum_i n_i \log(n_i/\hat{m}_i) = \sum_i d_i$$

We perform log-likelihood ratio test and see that the higher the order of the interaction, the greater the deviance explained. Each LRT for the nested models demonstrates that- the p-value for the reduction in deviance is significant. Note that these obey the **hierarchy principle**: If the k-way interaction is in the model then every lower order interaction and main effect is also in the model.

Table 6: ANOVA for 3 standard models

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev)
Model 1	529.822654	10	0.000000	0	0
Model 2	72.792862	4	457.029793	6	0
Model 3	9.761127	0	63.031735	4	0
Saturated	0.000000	0	9.761127	0	0

4.8.2 Iterative Proportional Fitting

We will use iterative proportional fitting to estimate N. The iterative proportional fitting process generates maximum likelihood estimates of the expected cell frequencies for a hierarchical model. In short, preliminary estimates of the expected cell frequencies are successfully adjusted to fit each of the marginal sub-tables specified in the model.The un-adjusted data cells may be referred to as the 'seed' cells, and the selected totals may be referred to as the 'marginal' totals.
Here the marginal subtables would be:

1. Security by Judicial by Forensic systems
2. Security by Forensic by Civil systems
3. Judicial by Forensic by Civil systems
4. Judicial by Civil by by Security systems

To do this four-dimensional IPF:

1. Proportionally adjust each (three-dimensional) row of cells to equal the pre-determined totals of Marginal 1.
2. Proportionally adjust each column of cells to equal the pre-determined totals of Marginal 2.
3. Proportionally adjust each slice of cells to equal the pre-determined totals of Marginal 3.
4. Proportionally adjust each stack of cells to equal the pre-determined totals of Marginal 4. This is the end of the first 'Iteration'.
5. Repeat the above steps until the desired level of convergence is reached.

Under mild restrictions, we know that the cell-values at the end of this process that satisfy the fitted marginal totals are the MLEs.

4.8.3 Results

Using iterative proportional fitting, we fit all possible second-order interaction models (we restrict to second-order to aid interpretation). These will account for different dependency structures.

We choose he best model according to the BIC criteria, is below: estimates a total for missing/disappeared people as 8817, which is similar to the 8782 than that of the non-hierch., interactionless, model (M_{th}).

Table 7: 'Top' Five, 'Middle' Five, and 'Bottom' Five Models (Hierarchical models)

	abundance	stderr	bias	deviance	df	AIC
12,13,14,23,24,34	8817.112	1515.7607	237.8220039	72.33182	3	189.5966
12,13,14,23,24	5076.249	168.5601	9.6710160	93.42304	4	208.6879
12,13,14,23,34	7308.085	455.7503	36.3107669	73.99807	4	189.2629
12,13,14,24,34	8695.570	493.7868	32.6905220	72.33923	4	187.6040
12,13,23,24,34	9267.442	876.2745	91.6873828	72.44539	4	187.7102
12,24,34	6990.394	275.9882	8.1501334	251.01919	6	362.2840
13,14,23	5388.545	160.4725	6.9759022	180.47969	6	291.7445
13,14,24	6408.681	225.7425	8.3355958	397.52853	6	508.7934
13,23,24	5052.802	130.3749	3.2161351	209.76658	6	321.0314
13,23,34	5178.863	140.5477	3.1073571	223.51518	6	334.7800
14,2,3	6349.780	217.6530	6.4628534	444.04882	8	551.3136
23,1,4	5076.432	127.2073	1.1770391	278.46874	8	385.7336
24,1,3	6089.940	188.3562	2.6692553	450.63028	8	557.8951
34,1,2	5903.270	172.5871	1.4531676	404.10047	8	511.3653
1,2,3,4	6035.010	181.6780	0.7620688	482.26705	9	587.5319

We can see that the independence model listed as "1,2,3,4" has the worst fit– this is to be expected, as we have demonstrated in previous sections the list dependencies

The shorthand "14,2,3" means that lists 1 and 4 are independent of 2, and also independent of list 3, i.e. $(1,4) \perp\!\!\!\perp 2$ and $(1,4) \perp\!\!\!\perp 3$. It is represented by:

$$\log(\hat{N}) = \alpha + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_{14}$$

The shorthand identifies both the model and the margins that must be fitted to obtain MLEs using IPF. Not all shorthands have a conditional independence interpretation. The top models, all with 5-6 two-way interactions cannot be interpreted in this way.

These the hierarchical models have different estimates, ranging from 5053 to 9268.

We also plot the BIC values for different models and their accompanying estimates of $\hat{N}$. When we only look at estimates with a corresponding BIC value less 300, the estimates range from 5077 and 9268 (almost the exact same range of estimates when considering all the models).

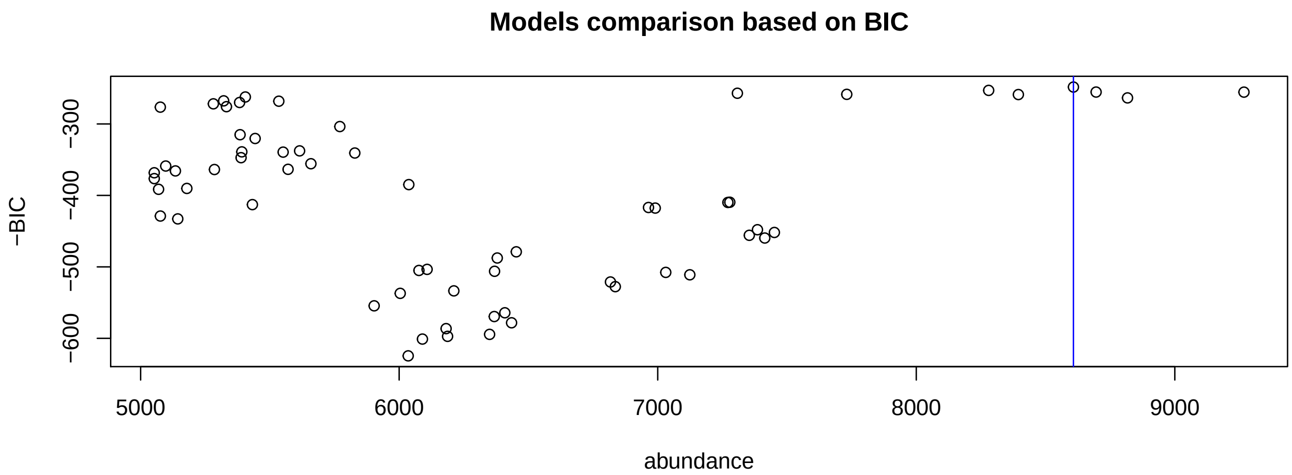


Figure 9: Hiercharical Models BIC Plot

This is the fundamental problem of the frequentist approach. We could just pick the "best" model, i.e., the one with the lowest BIC (note the y-axis is inverted).

Unfortunately, just picking one model ignores the error that we introduce by the selection itself. It also forces us to decide which dependencies among the systems we will control for, forcing us to decide which dependencies will not be included in the model. In the future, we would look into model averaging procedures and Bayesian approaches.

5 Conclusions

Our findings show that the total number of deaths and disappearances in Casanare is far greater than previously assessed. The initial analysis made by Guzman et al. for the Human Rights Data Analysis Group (HRDAG), which used thirteen datasets from 1986-2007, estimated a total of 2,533 deaths and disappearances in Casanare (95% CI: 2239-2867) from 1986-2007, although only about 1,500 were reported as such (Guzman et al., 2007). We only used datasets for a subset of those years, namely 1998-2007 (fifteen datasets in total). Looking at just the raw total from the fifteen datasets, there were 3,501 incidents of deaths or disappearances in Casanare from 1998-2007. Had our analysis also included data that went back to 1986, the raw total would be even higher.

Importantly, given estimation techniques used in our analysis, there is strong evidence that the true total of victims is actually much higher than 3,501. We used multiple systems estimation (MSE) and loglinear modeling to show that the best hierarchical models of the data produce estimates ranging from 5,077 to 9,268. While our estimation of the total number of deaths and disappearances indicates a far greater number of victims, there are some concerns estimating the true total of victims. First, the models themselves indicate a large range of possible numbers. Second, the data may not meet all the MSE and loglinear assumptions (especially the independence assumption), which makes point estimation less valid.

As can be seen, there is a lot of uncertainty about the true total of deaths and disappearances. Nevertheless, the number of deaths and disappearances within any given country is one important human rights measure that serves six inter-related functions: (1) contextual description and documentation, (2) classification, (3) monitoring, (4) mapping and pattern recognition, (5) analysis and policy recommendation, and (6) advocacy and political dialogue (Landman and Carvalho, 2009). While MSE and loglinear modeling can only go so far in terms of the first four functions, these statistical tools for better estimating the total of deaths and disappearances is nevertheless key for, on the one hand, analysis and policy recommendation, and on the other, advocacy and political dialogue. International NGOs like Amnesty International and Human Rights Watch respond based on the severity of human rights violations in a country, opting to use higher estimates of deaths and disappearances for their analysis of the situation in Colombia and thus their recommendations and advocacy, while the Colombian think tank, CERAC, have lower estimates (Ballesteros et al., 2007). Unfortunately, different estimates can be used as fact by different groups to proffer differing messages about the extent of human rights violations in Colombia. Nevertheless, utilizing statistical tools to show that the true total of victims is markedly higher than previously reported is particularly important for uncovering the extant nature and degree of violence in Colombia.

6 References

Agresti, A. (2003). Introduction to Categorical Data Analysis. 394.

Ballesteros, A., Restrepo, J., Spagat, M., and Vargas, J. (2007) The Work of Amnesty International and Human Rights Watch: Evidence from Colombia, Bogota, Colombia: Centro de Recursos para el Analisis de Conflictos. Online. Available HTTP: http://cerac.org.co/assets/pdf/Other%20publications/CERAC_WP_4.pdf (accessed 4 December 2019).

Christensen, R. (2006). Log-Linear Models and Logistic Regression. Springer Science & Business Media.

Darroch, J., Fienberg, S., Gary F. V. Glonek, & Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. Journal of the American Statistical Association, 88(423), 1137-1148. doi:10.2307/2290811

Gomez-Suarez, Andre (2007). “Perpetrator blocs, genocidal mentalities and geographies: the destruction of the Union Patriotica in Colombia and its lessons for genocide studies.” Journal of Genocide Research9:4, 637-660.

Guzman, D., Guberek, T., Hoover, A. and Ball, P. (2007) Missing People in Casanare, Palo Alto, CA: Human Rights Data Analysis Group. Palo Alto, CA: Benetech Initiative. Online. Available HTTP: <www.hrdag.org/resources/publications/casanare-missing- report.pdf> (accessed 4 2019).

Junker, B. (2007). Carnegie Mellon Lecture Notes: 36-720: Log-Linear Models: Three-Way Tables. Retrieved from http://www.stat.cmu.edu/~brian/720/week03/04-loglinear.pdf

Landman, Todd and Carvalho, Edzia (2009). Measuring Human Rights. London: Routledge.

Mitchell, Shira (2014). Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation. Retrieved from https://dash.harvard.edu/bitstream/handle/1/12274610/Mitchell_gsas.harvard_0084L_11586.pdf?sequence=4&isAllowed=y