Final Project; Categorical Data Analysis

Abby Smith, Miruna B., Martha Eichlersmith

11/10/2019

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. Multiple armed groups have operated in Casanare including paramilitaries, guerillas and the Colombian military. Many Casanare citizens have suffered violent deaths and disappearances.

But *how many* people have been killed or disappeared? We review the Human Rights Data Analysis Group (HRDAG)'s reporting on this work. In this study, the authors used information about victims of killings and disappearances provided by 15 datasets. 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 disambguiate the patterns of violence?

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

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.

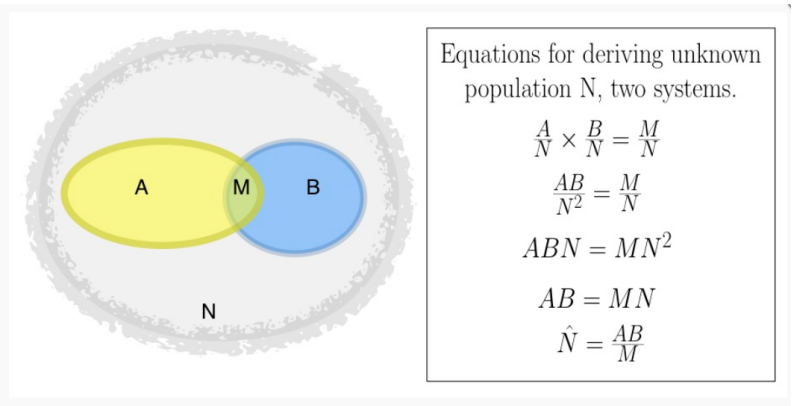


Figure 1: Multiple System Estimation

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 unproblematic 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. TThis 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 reportedmissing; 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, dependences 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.

3 Overview of Data

Table 1: Contingency Table

	Organization	Total Captures	Unique	Type
d_CCJ.n	Colombian Commision 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 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 diffrent 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.

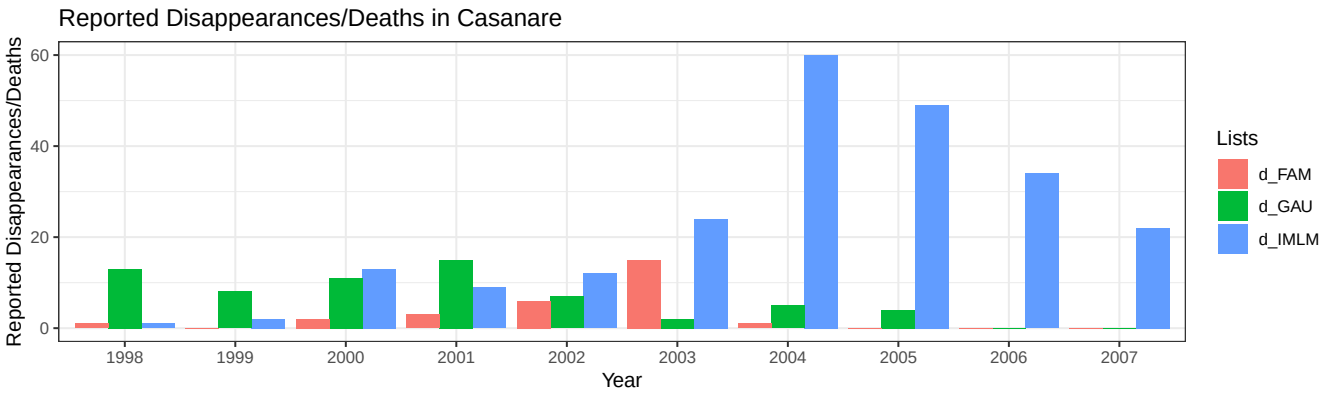
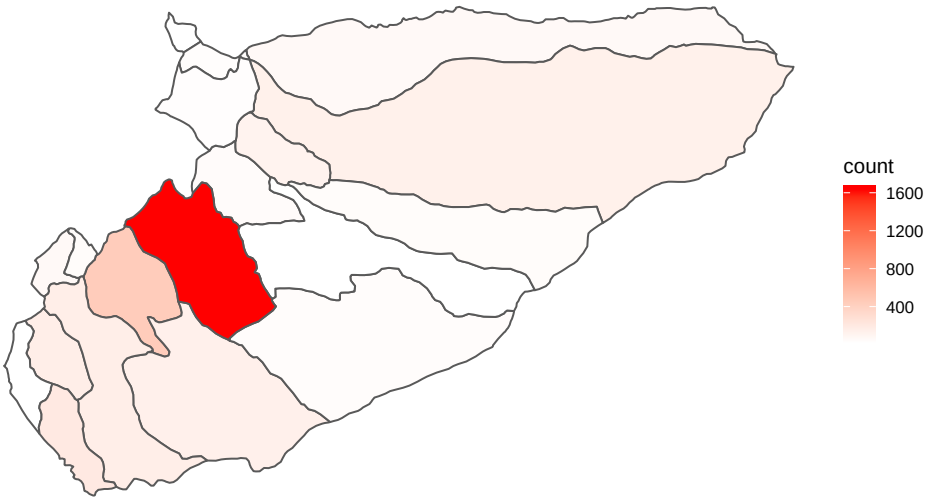


Figure 2: Count Trends for Three Organizations

3.2 Heterogenity

From the bottom graph of our heterogenity graph, we see differences in the capture probabilities across lists.

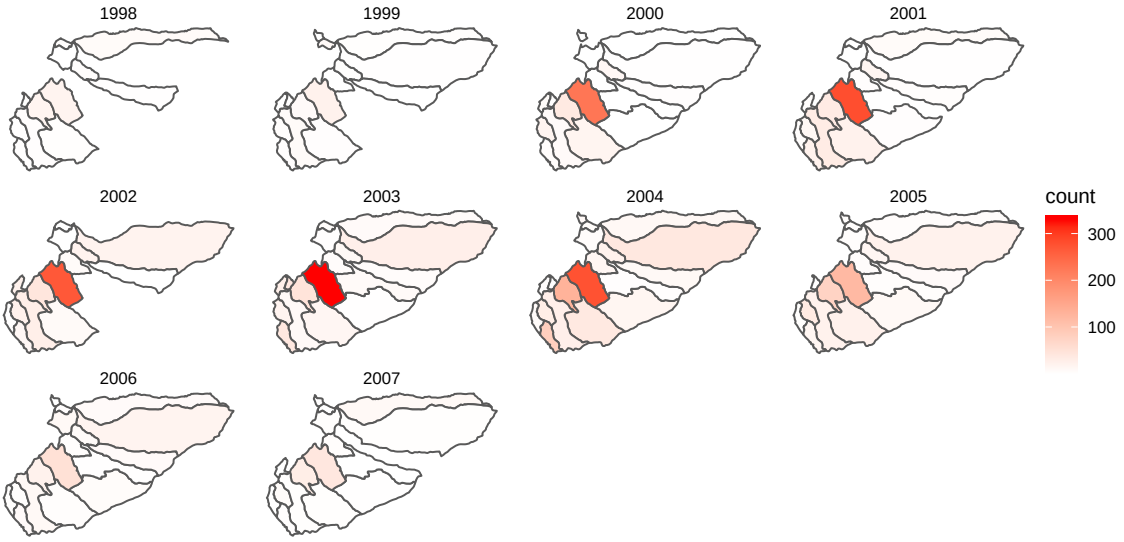
NEED MORE DESCRIPTION
- this all the lists for all years?? - need more info on this graph



3.3 Total Reported Victims Trends

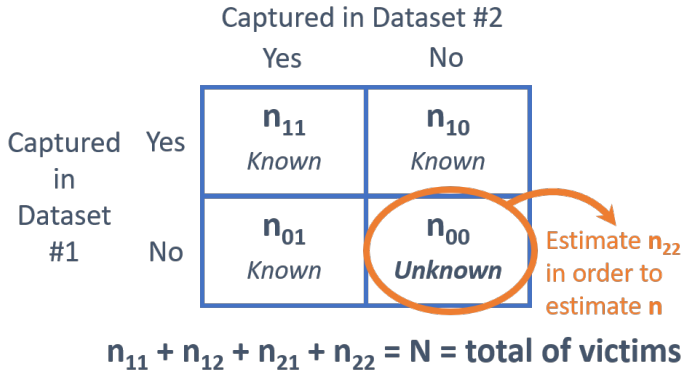
We can see that the *reported* violence appears to have intensified across all municipalities in Casanare in 2003-2004 and then dropped.

NEED MORE DESCRIPTION - is this one list? - all the lists together



4 Loglinear Modelling

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 combindation 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. Poisson regression treaets these cell counts - not the underlying individual cases data

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 $\exp(\alpha)$.

$$\log(n_{00}) = \alpha \underbrace{+\beta_1 \cdot \mathbb{1}(x \in n_{10}) + \beta_2 \cdot \mathbb{1}(x \in n_{01})}_{\text{each } =0}$$

$$\log(n_{00}) = \alpha$$

$$\hat{n}_{00} = \exp\{\hat{\alpha}\}$$

$$\hat{N} = n_{11} + n_{10} + n_{01} + \hat{n}_{00}$$

As with any regression and data mining model, we want to avoid overfitting. There is a tradeoff 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

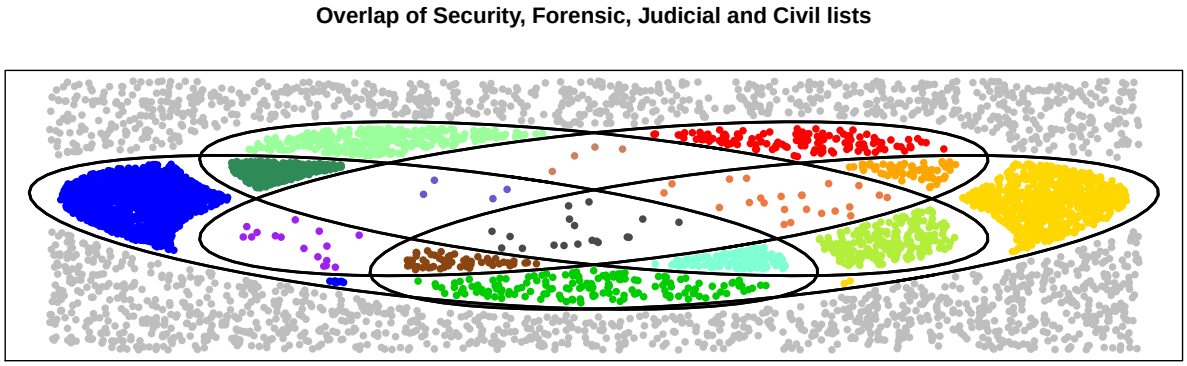
1. **Interpretation:** The inclusion of so many variables in loglinear models often makes interpretation very difficult.

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.

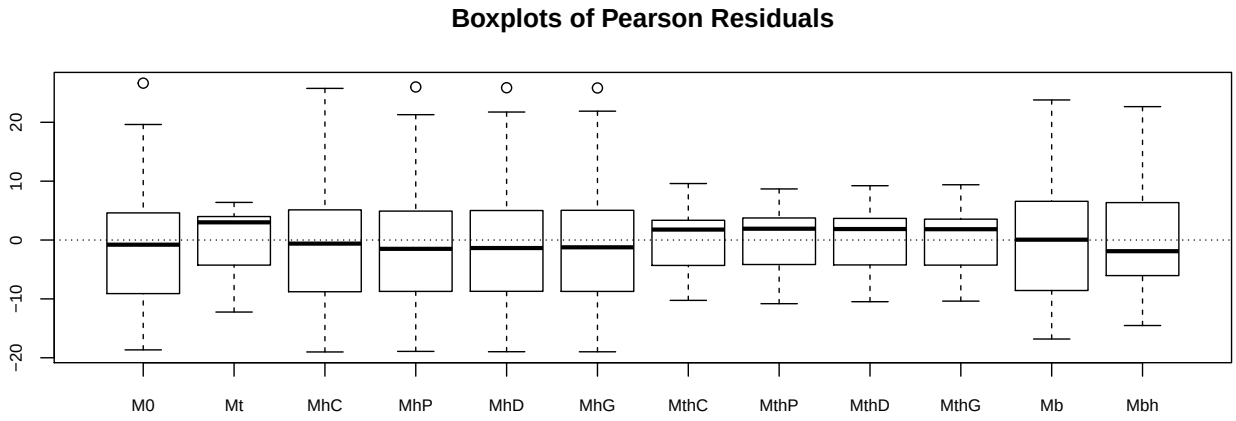


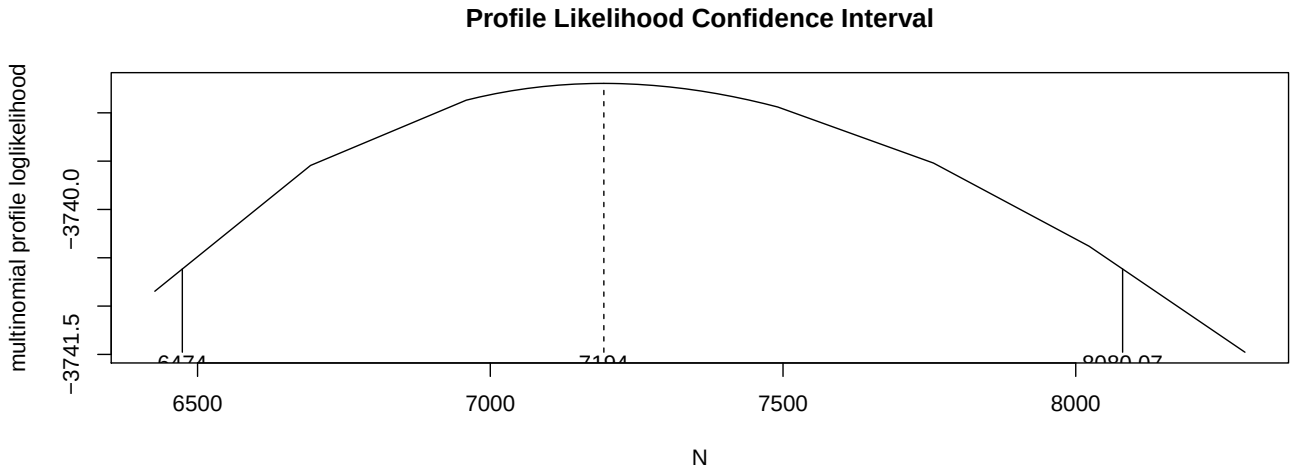
5 Results of Loglinear Modelling

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* hierchachal in nature.

We will need the hierchachal 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.

```
##
## Number of captured units: 3501
##
## Abundance estimations and model fits:
##                abundance stderr deviance df      AIC      BIC infoFit
## M0                5604.9  100.6 2279.471 13 2376.736 2389.057      OK
## Mt                5232.8   87.3  529.823 10  633.087  663.891      OK
## Mh Chao (LB)      5970.7  139.9 2253.979 12 2353.244 2371.726      OK
## Mh Poisson2       6147.5  188.4 2262.122 12 2361.387 2379.869      OK
## Mh Darroch        6866.5  367.9 2257.937 12 2357.202 2375.684      OK
## Mh Gamma3.5       7725.1  634.4 2256.636 12 2355.901 2374.383      OK
## Mth Chao (LB)     5665.3  125.1  473.765  8  581.029  624.155      OK
## Mth Poisson2      6035.0  181.7  482.267  9  587.532  624.497      OK
## Mth Darroch       7204.7  411.3  476.057  9  581.322  618.287      OK
## Mth Gamma3.5      8782.7  810.7  474.715  9  579.980  616.945      OK
## Mb                4228.2   64.7 2107.184 12 2206.448 2224.931      OK
## Mbh               3643.0   47.7 1618.987 11 1720.252 1744.895      OK
```






```
##
## Number of captured units: 3501
##
## 95% profile likelihood confidence interval:
##              abundance     InfCL     SupCL
## Mth Darroch       7194  6473.997  8080.067
```

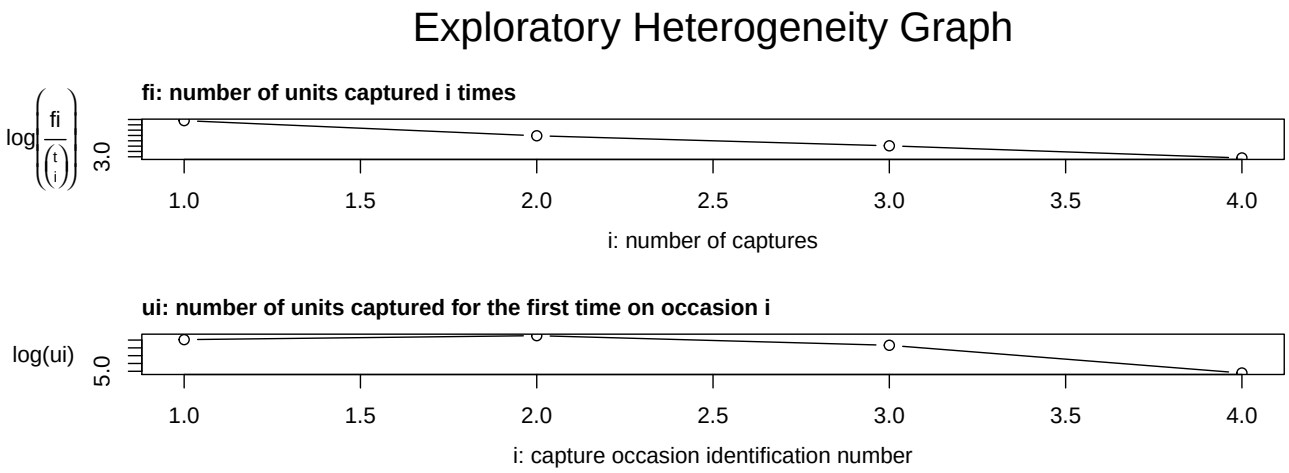
The "number of captured units" is the number of observed elements, in this example, the number of people documented as missing/killed, we usually call this N_k. The "abundance" column shows the estimate of $\hat{N}_k$, the total population including the observed and the estimated unobserved deaths. The 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).

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.

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 capturedfor the last time on occasion i (v_i) and the number of units captured on occasion i (n_i). 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.

```
##
## Number of captured units: 3501
##
## Frequency statistics:
##        fi    ui    vi    ni
## i = 1  2392  1128   317  1128
## i = 2   866  1455  1640  2028
## i = 3   225   785  1216  1387
## i = 4    18   133   328   328
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
```



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

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.

When heterogeneity is present, 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.

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.

For example, in the model $list_1$, $list_{12}$, $list_{123}$, the initial estimates are adjusted to fit $list_{12}$ then $list_{23}$ and finally to equal the $list_{123}$, observed frequencies. The previous adjustments become distorted with each new fit, so the process starts over again with the most recent cell estimate. This process continues until an arbitrarily small difference exists between the current and previous estimates.

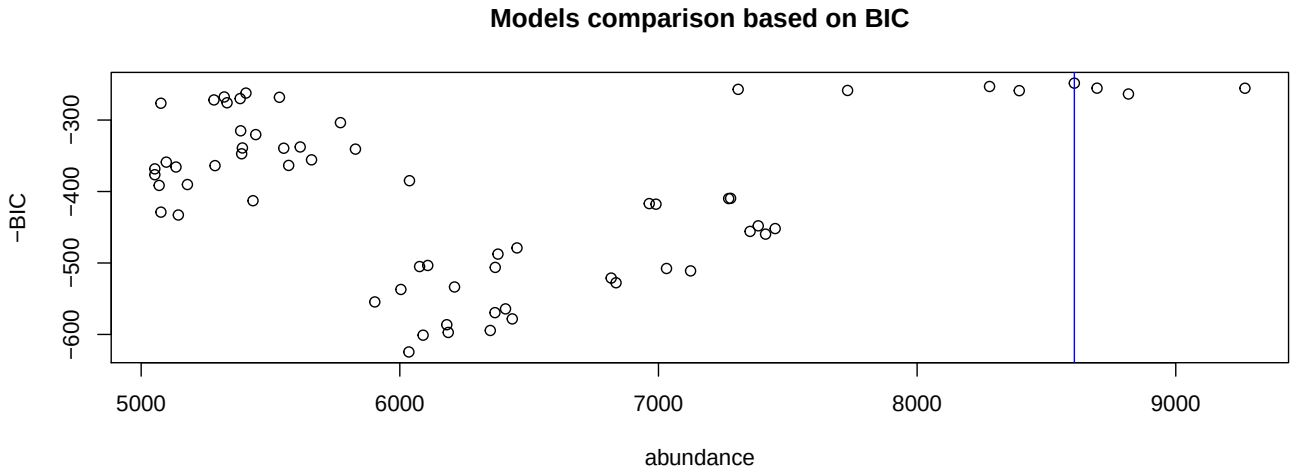
We fit hierchachal models, restricting to second-order (no three-way interactions) for sake of interpretation.The best model according to the BIC criteria, is below: estimates a total for missing/disappeared people as 8607, which is similar to the 8782 than that of the non-hierch., interactionless, model.

$$\log(\hat{N}) = \mu + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_{12} + \lambda_{13} + \lambda_{24} + \lambda_{34}$$

NEED TO TALK ABOUT INTERPREATION, DESCRIBE NOTATION

We also plot the BIC values for different models and their accompanying estimates of $\hat{N}$.

```
##
## Number of captured units: 3501
##
## Abundance estimations and model fits for the models with the smallest BIC:
##                       abundance stderr  bias deviance df     AIC     BIC
## [12,13,24,34]            8607.9  476.5  23.7   73.490  5 186.755 248.363
## [12,13,34]               8280.0  430.3  15.9   86.315  6 197.580 253.027
## [12,13,14,24,34]         8695.6  493.8  32.7   72.339  4 187.604 255.373
## [12,13,23,24,34]         9267.4  876.3  91.7   72.445  4 187.710 255.479
## [12,13,14,23,34]         7308.1  455.8  36.3   73.998  4 189.263 257.032
## [12,13,23,34]            7731.2  494.3  33.2   83.697  5 196.962 258.570
## [12,13,14,34]            8394.9  449.2  24.4   83.983  5 197.248 258.856
## [14,23,24,34]            5405.1  161.7   7.4   87.392  5 200.656 262.264
## [12,13,14,23,24,34]      8817.1 1515.8 237.8   72.332  3 189.597 263.526
## [14,23,24]               5321.2  153.4   6.7  100.963  6 212.228 267.675
##                       infoFit
## [12,13,24,34]              OK
## [12,13,34]                 OK
## [12,13,14,24,34]           OK
## [12,13,23,24,34]           OK
## [12,13,14,23,34]           OK
## [12,13,23,34]              OK
## [12,13,14,34]              OK
## [14,23,24,34]              OK
## [12,13,14,23,24,34]        OK
## [14,23,24]                 OK
```



Note that these the hierch. models have different estimates ,ranging from 5231 to 8607, all within very similar BIC values Which one should we use?

This is the fundamental problem of the frequentist approach. We could just pick the "best" model, i.e., the one with the lowest BIC. 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.

6 Goodness of Fit

I will describe here how likelihood ratio test is better for model comparison than Pearson's Chi-square??

7 Bayesian Model Averaging (??)

This method allows the analyst to flexibly account for list dependency by creating models for all possible dependencies, and averaging over them in a way that is proportional to the probability that the dependence is correct.

The first step in the analysis is to formulate a prior for population size. This is represents the analyst's prior knowledge about population size along with uncertainty. By default, a "non-informative" prior is used.

8 Conclusions