I used data in two different ways - one to estimate casualty figures and one to "estimate" start dates. I use data to estimate these figures for 118 communes across the 12 prefectures for the days from 4/5/1994 through 7/31/1994 - or roughly 13,924 commune days. At least one source used to estimate casualty figures has non-zero values on 857 of these 13,924 commune-days (or about 6% of the commune-days). I think it is important to recognize that there are very likely other events going on that were not captured by these surveys and that in the aggregate, they could amount to a sizable number of casualties. I say all of these things to highlight the fact that these are estimates, and as such, while hopefully a good estimate of the truth, these do not "reveal" the truth; rather, they provide us with a best guess of what the "truth" is based on the data. To the extent that other events not captured in these data sources were ongoing at the time, these estimates are more likely to be under- rather than over-estimates of the actual conflict-related deaths.

1 Casualty Estimates

I used 5 sources to estimate casualty figures - Africa Rights, Human Rights Watch, Ministry of Youth, Culture and Sport, Ministry of Education and Ibuka. I start by assuming that if no source mentioned activity on a particular day, that no activity happened on that day. Next, I assume that if at least one source mentioned activity, that something did happen on that particular day. I only include in the analysis, days for which some source mentioned an event. The data then would look something like the following:

Table 1: Hypothetical Observations

AR	HRW	YCS	ED	Ibuka
100	NA	NA	150	NA
NA	NA	NA	NA	10

Table 1 shows two hypothetical observations. In the first, two sources contained information for the day in question and in the second, only one source contained information. The sources that do not mention any activity are coded as missing (`NA`), which gives the model more flexibility than if they had been coded as zero.

These various sources represent considerably different aggregate figures. The aggregate counts from the observed data are provided in Table **??**. Sometimes, different estimates were given regarding the magnitude of killings at various locations. When this happened, we recorded both the lowest and highest values as well as the mean value. The figures in the "Total" column of table **??** represent the aggregation of the means. Those values in the "Lower" and "Upper" columns represent the aggregation of these upper and lower values.

Table 2: Observed Aggregate Casualty Figures 4/5/94 - 7/31/94

Source	Total	Lower	Upper
Africa Rights	138825	121651	179586
Ministry of Education	760317	752702	767941
Human Rights Watch	40557	31778	63406
Ibuka	25703	25445	25729
Ministry of Youth, Culture and Sport	823593	803313	854319

The casualty estimates are generated using a Bayesian latent variable model. The first step in the model is to draw the "observed" casualty figures from a uniform distribution with lower and upper bounds equal to the smallest and biggest counts mentioned in the source material, respectively. When the source did not provide contradictory information, the lower bound was the observed figure minus 0.5 and the upper bound was the observed figure plus 0.5.[1] This allows us to propagate uncertainty about the magnitudes of events through the latent variable model which will result in increasing the uncertainty on the latent variable point estimates. One benefit of using a Bayesian model is its ability to easily incorporate known or hypothesized sources of variability.

Next, I built a model for these "observed" counts.

$$Y_{cds} = \gamma_{0s} + \gamma_{1s}\theta_{cd} + \nu_{cds} \tag{1}$$

where Y_{cds} indicates the activity in commune c on day d from source s. In words, I am saying that there is one true value θ_{cd} for each commune day and that every account of activity across the four sources is an estimate related to θ_{cd}. The model was estimated through Markov Chain Monte Carlo simulation. As such, priors are required on each of the model parameters. The residuals variances for the Y_{cds} were given Inverse Gamma distributions with $a = 1$ and $b = 1$. The coefficients (γ's) were given normal priors with mean equal to zero and variance equal to 10000, thus essentially flat. The latent variable scores were given normal priors with mean zero and an estimated variance.[2]

1.1 Results

The model was run for 30000 burnin iterations with 1250 monitored iterations on 2 chains. The results show strong evidence of convergence both according to visual methods (e.g., examination of trace-plots, Brooks, Gelman and Rubin) and numerical methods (e.g., Brooks and Rubin, Geweke). The results presented below are functions of the 2500

[1] This was done to accommodate the requirements of the software and is unlikely to create any real problems. The only thing it *could* do is to increase the variance on the latent variable estimates, though only marginally.

[2] The variance was an identified parameter because $\gamma_{0,AR}$ was set to zero and $\gamma_{1,AR}$ was set to 1 to identify the scale of the latent variable. This estimated variance was also given an Inverse Gamma distribution with $a = 1, b = 1$.

monitored iterations. Predicted counts are taken by aggregating (i.e., taking the mean of) $\hat{Y}_{cds}$, the model predicted counts for each source across various sources. Originally, we used all the sources, but found that the choice of sources to use figured prominently in how "big" the aggregate figures were. Considerable variation could be induced by simply choosing different sources. Thus, we present information from all possible combinations of the sources organized two different ways. Figure **??** shows them ordered by magnitude from highest to lowest. Figure **??** shows them organized by number of sources used.
The "meta-bounds" here are roughly 0-1.4 million. Thus, we can be relatively confident that the aggregate figure is somewhere in this range. Unfortunately, definitively narrowing it down any further would require making more assumptions

All of these figures are in the attached dataset. The variables labeled with `ar`, `ed`, `hrw`, `ibk`, and `ycs`, are the original data. The variables labeled with `A`, `E`, `H`, `I`, `Y`, are the estimates and the combination used to make the estimate is indicated in the variable name.

Figure 1: Aggregate Figures by Source Combination (organized by magnitude)

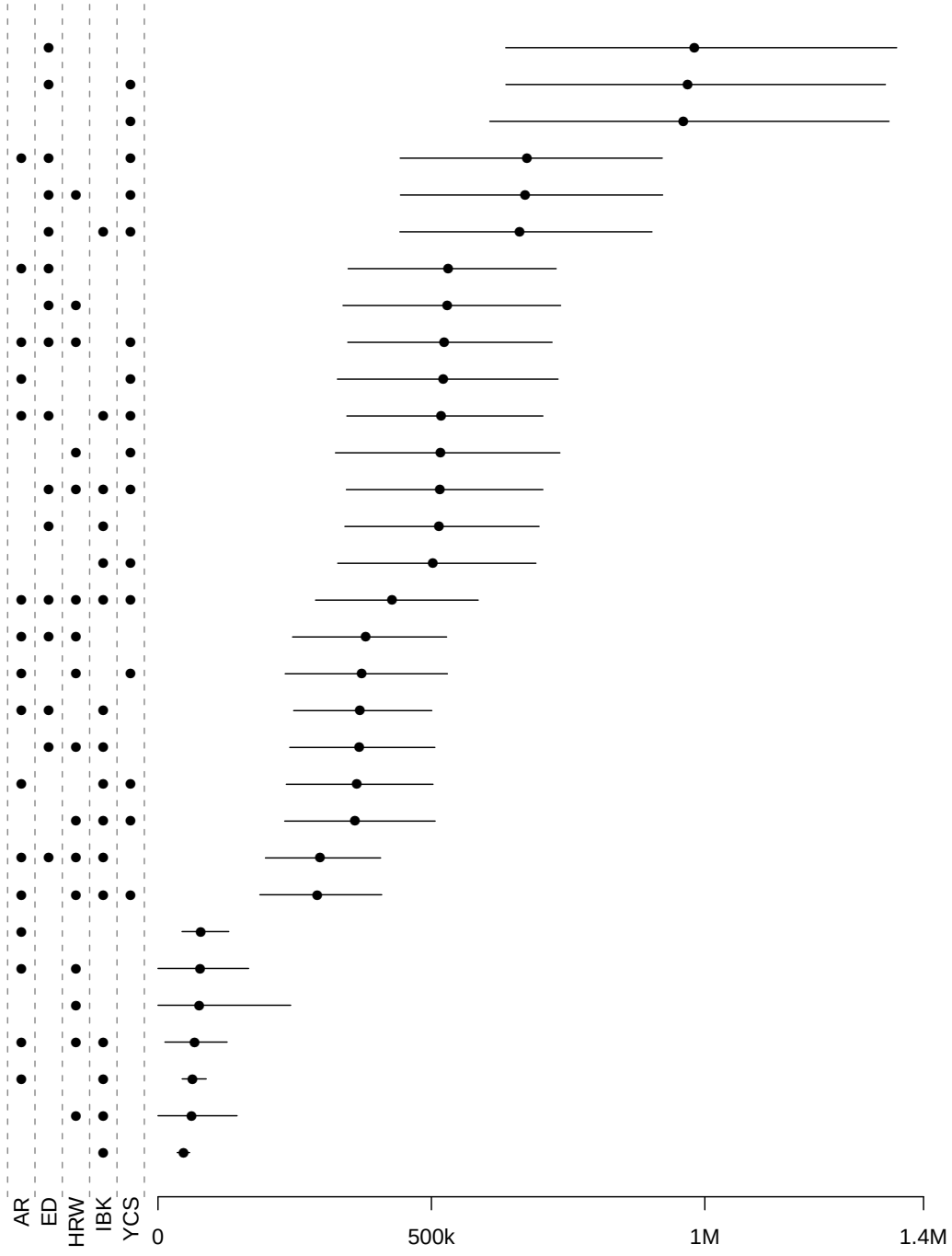
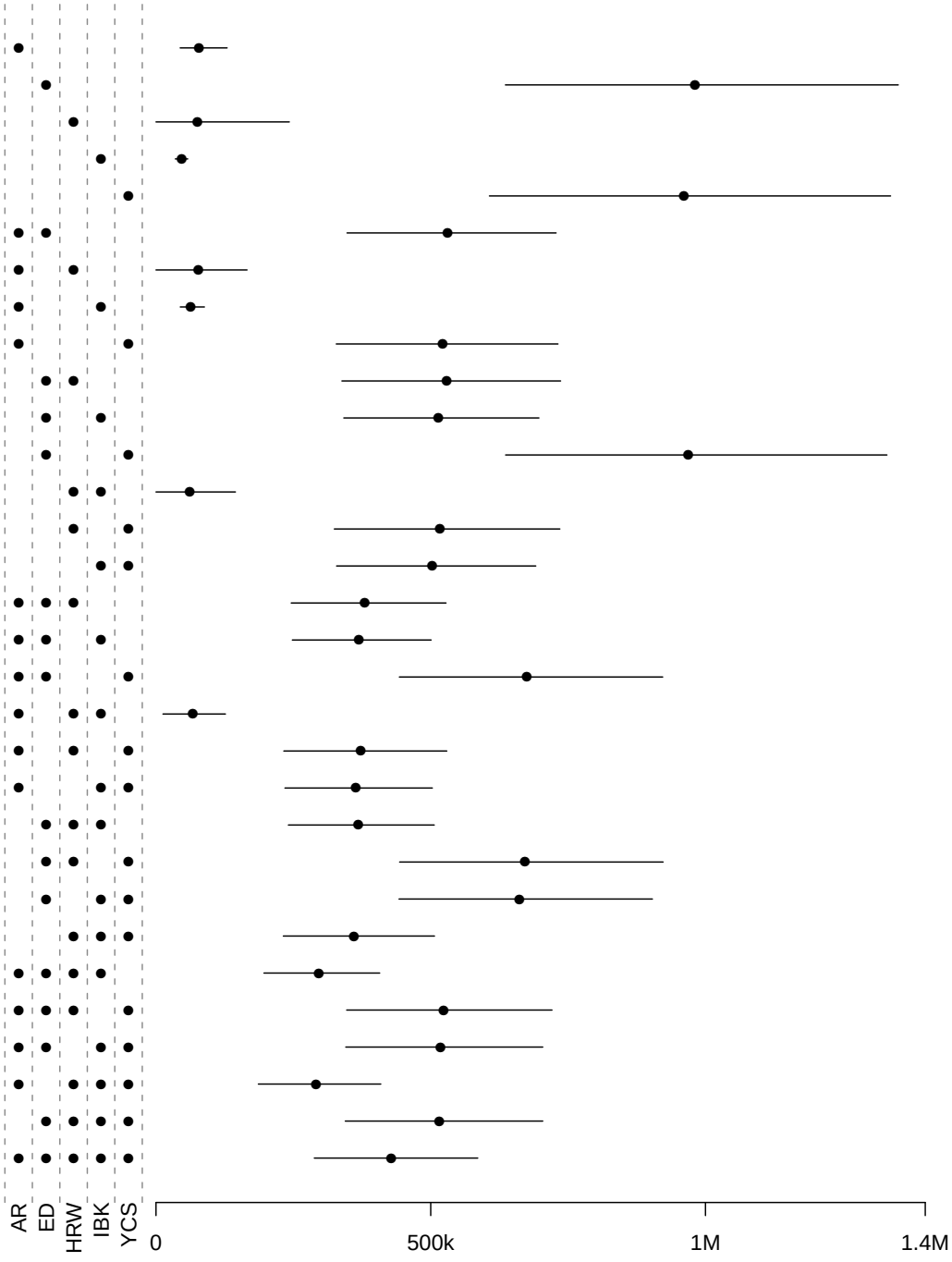


Figure 2: Aggregate Figures by Source Combination (organized by # sources used)



2 Start Dates

There are two strategies to calculating or estimating start dates. One would be with a predictive model that would allow the possibility that activity started *before* any organization indicated activity. While this is possible, no predictive model exists that is sufficiently well justified to permit this type of statistical conjecture in this case. The other alternative is to consider all available data and to identify the earliest day that any source indicated activity in a particular location. This is exactly what I did. For this, I used not only the four sources from above, but also the ICTR eye-witness testimonies and the figures from Ruzibiza's book on the subject. Together, these are constitute the majority of the data that has been collected on the events in Rwanda. Start date estimates are in the attached spreadsheet.