Explanation of Rwanda Data Methodology

Dave Armstrong
University of Wisconsin-Milwaukee
e: dave@quantoid.net

October 17, 2014

We used the five data sources and a latent variable model to try to figure out the extent of casualties in the Rwandan genocide.[1] By latent variable model, what we really mean is that we assume a) that there is a true number of casualties in each commune-day and b) that the observed data are error-laden (i.e., lower quality) estimates of those true values. By lower quality, we mean that the observed value is the true value contaminated with some random noise. Even though this sounds complicated, we are all well acquainted with the idea. Imagine you want to figure out the level of reading comprehension for each student in a class. You could give everyone the same passage to read and then ask all of the students to answer questions about it. We wouldn't want to ask one question because people could get lucky, could be distracted for the moment, etc... There are lots of reasons why a single estimate might not work. Instead, we ask lots of questions about the passage. The extent to which each person answers the questions correctly is a function of underlying latent ability and whatever other idiosyncratic stuff is happening at that particular time (i.e., random noise). By averaging across the questions (i.e., calculating the proportion correctly answered), we get a better sense of each student's reading ability. The idea is exactly the same here. It might be that African Rights missed or over-counted (due to random factors) victims in any particular locale. Same for other sources, too. By averaging over the different sources, we hope to come up with a better measure of the number of casualties (and the pattern of killings over space and time).

The model here is a bit more complicated than that. What we do is say that the true level of killings (call this μ_i where i stands in for each commune-day) has a linear relationship with each of the different measures. So, we can predict the African Rights casualty figures as a linear function of μ_i (e.g., with $\text{AR}_i = a + b\mu_i$).[2] We replicate this process for all of the five different measures, allowing each to have its own linear relationship to the true casualty estimates. The model is complicated by the fact that we do not have data for every commune-day. We assume that if something happened in a particular commune on a

[1] This is a very simple version of a Bayesian Factor Analysis Model. These models and derivative varieties have been used extensively in political science. To read more about them, see Congdon (2002, pp. 323-334), Jackman (2009, pp. 435-453) or Armstrong (2009) for an applied example.

[2] Here, a represents the expected casualty figure when our estimated true casualty figure is zero and b gives the amount by which we expect the African Rights casualty figures to increase with each extra "true" casualty.

particular day, at least one of the sources would pick it up. If no source identified casualties on a commune day, we assume the true value to be zero and do not estimate it. Figure 1 (below) identifies the observed and missing observations by data source, commune and day. Blocks are colored white if no source identified activity in that commune day, gray if a source (but not the one represented in the figure) identified activity as happening in that commune on that day and black if the source represented in the figure identified activity in that commune on that day.

We use the model identified above to estimate the casualty figures in each of the gray blocks.[3] We do that by simply taking our estimate of the true value on the particular commune day, multiply it by the source's coefficient and then add the source's intercept value. This gives us a predicted value for each source for each commune-day represented in the data. Figure 1 on the GenoDynamics site (`http://genodynamics.weebly.com/data-on-violence.html`) these averages across all different combinations of the five sources. For example, the biggest predictions we get come from using just the Ministry of Education and/or the Ministry of Youth, Culture and Sport figures. We get the smallest figures by using just the Ibuka data. Not surprisingly, the figures using all five sources simultaneously are in the middle. Figure 2 on the GenoDynamics website is just a re-ordered version of Figure 1.

This brings us to Figure 3. The Ministry of Local Affairs (MINALOC) did a census of victim casualty figures (aggregated at the commune level, not disaggregate over time). They identified the number of victims that were declared to be dead by people they interviewed and also the number of victims that were actually accounted for. We wanted to know how closely our aggregated figures were, on average, to those numbers identified by MINALOC. That is what Figure 3 provides. The two different shapes, circles and squares, represent the relationship between our aggregate casualty figures by commune and the MINALOC number of victims declared measure or the MINALOC number of victims accounted for, respectively. What we are trying to figure out is to what extent does variation (not necessarily exact numerical value, but patterns of high and low values) in our measures correspond to variation in the MINALOC data (victims both declared and accounted for). The black symbols represent the correlation on the raw values and the red symbols represent the correlation on the log of the casualty figures. Since the casualty figured distributions are skewed (a few really big positive values create a skew to the right), a log transformation decreases the influence of those really big values. The higher the value, the greater the extent to which our values follow the MINALOC values. Consider the logged values for the highest (all indicators) and lowest (just Ibuka) combinations. Figure 2 (below) shows the plots of predictions versus the MINALOC figures. It is clear to see that the predictions using all indicators more closely parallel those of MINALOC than do the ones using only Ibuka.

[3]One other complication is that for some data, we only have bounds on the estimate. For example, African Rights gives a range of values between 18000 and 20000 for Mabanza on 4/17/94. We made sure to incorporate that uncertainty into our models so that the aggregate predictions would reflect the fact that sources were more certain or precise about some figures than others.

Figure 1: Patterns of Missingness Across Indicators

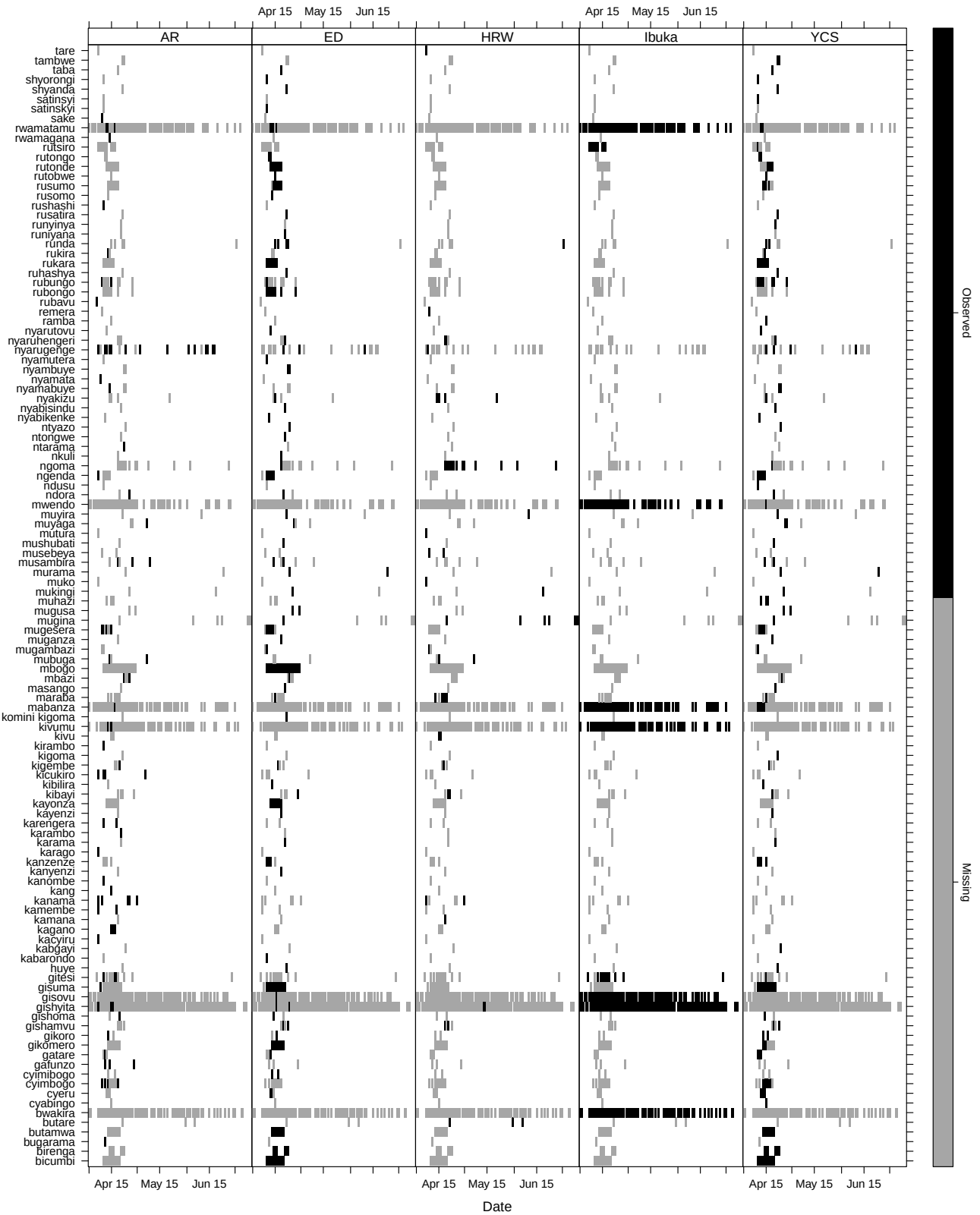
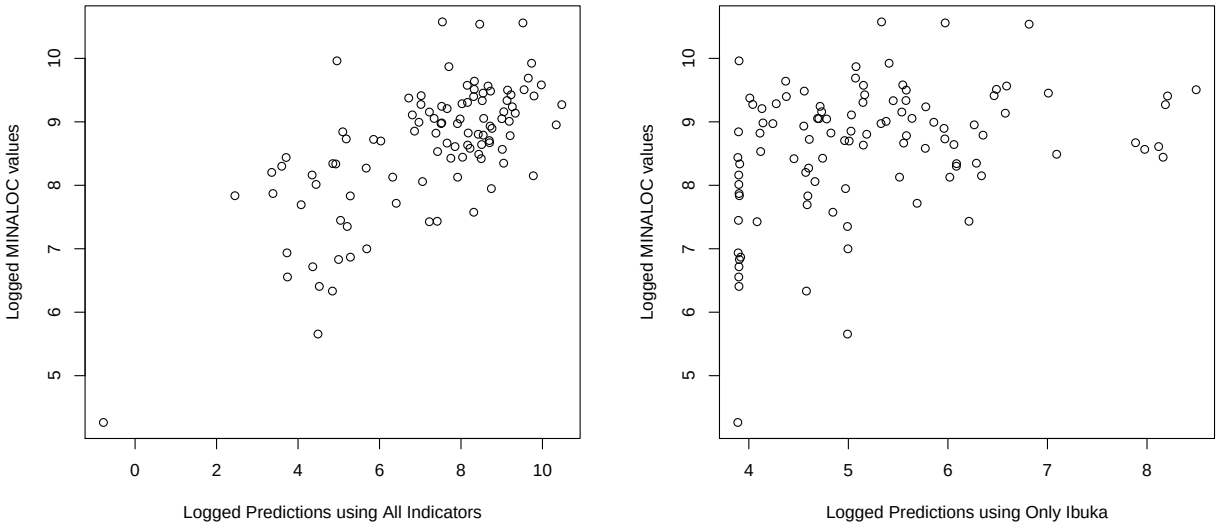


Figure 2: Plot of Comparisons Between Our Figures and MINALOC



(a) All Indicators

(b) Only Ibuka

References

Armstrong, David A. 2009. "Measuring the Democracy-Repression Nexus." *Electoral Studies* 28(3):403–412.

Congdon, Peter. 2002. *Applied Bayesian modelling.* New York; Chichester: Wiley.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences.* Chicester, UK: John Wiley & Sons, Ltd.