Preliminary Statistical Analysis of Documentation of Killings in the Syrian Arab Republic[*]

Megan Price, Jeff Klingner, and Patrick Ball

The Benetech Human Rights Program

2 January 2013

[*]This analysis was commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR).

Executive Summary

This report presents findings integrated from six databases built by Syrian human rights monitors and one database collected by the Syrian government. The databases collect information about conflict-related violent deaths — killings —- that have been reported in the Syrian Arab Republic between March 2011 and November 2012. Although conflict conditions make it difficult to identify an accurate record of events, governmental and non-governmental monitors are persevering in gathering information about killings through a variety of sources and venues. The purpose of the report is to explore the state of documentation, the quantitative relationship of the sources to each other, and to highlight how understanding of the conflict may be affected due to variations in documentation practices.

This report examines only the killings that are fully identified by the name of the victim, as well as the date and location of death. Reported killings that are missing any of this information were excluded from this study. This report finds that when the fully identified records were combined and duplicates identified, the seven databases collected here identified **59,648** unique killings.

It should be noted that this count is not the number of conflict-related killings in the Syrian Arab Republic. The statistics may include a small number of undetected duplicates among the unique killings, thus, this count may be slightly too high. More significantly, there is an unknown number of killings which have not yet been documented by any of these seven projects. As each additional dataset has been added over the past few months, previously undocumented deaths have been reported. The statistics presented in this report should be considered minimum bounds.

This report provides comparative statistical analyses of all seven datasets, including patterns of documented killings over time, as well as by geography, sex and age of the victims (in Section 2). A detailed analysis of how the datasets overlap with each other is presented in Section 3; the overlap analysis helps explain how the various data sources each capture distinct aspects of the total universe of killings.

Methodology

This report begins with 147,349 records of reported killings of fully identified victims from seven datasets. Many of these records are duplicates. An expert whose native language is Syrian Arabic and who is fluent in English reviewed 8,280 pairs of reported deaths. He classified the reports as either referring to the same victim or to different victims. Benetech used the expert's classifications with a computer algorithm called an Alternating Decision Tree to build a model to classify the remaining records as either matches or non-matches. The resulting records were merged into a combined dataset which, with duplicates removed, includes 59,648 records of documented killings (more detail on matching is available in Appendices A.1 and A.2).

1 Documented Killings

This report presents an analysis of killings that have been reported in the Syrian Arab Republic between March 2011 and November 2012, based on seven datasets: 1) the Violations Documentation Centre[1] (VDC), the documentation arm of the Local Coordination Committees; 2) the Syrian Network for Human Rights[2] (SNHR); 3) the Syrian Revolution General Council (SRGC), which was combined with the SNHR (see below); 4) the Syria Shuhada Website [3] (SS); 5) the March 15 Group (15Mar); 6) the Syrian Observatory for Human Rights[4] (SOHR); and 7) the Syrian government (GoSY). For brevity, each list will be referred to by its acronym in the tables and figures throughout this report.

Benetech is aware of other organizations collecting data on killings in the Syrian Arab Republic, for example the Strategic Research and Communication Centre[5] and Syria Tracker[6], among others. Unfortunately, Benetech has not yet been able to obtain copies of data from these sources.

The first step in this analysis involves close examination of each individual record in each dataset in order to identify multiple records that refer to

[1] http://www.vdc-sy.org/

[2] http://www.syrianhr.org/

[3] http://syrianshuhada.com/

[4] www.syriahr.com, www.syriahr.net

[5] http://www.strescom.org/

[6] https://syriatracker.crowdmap.com/

the same death. Sometimes these records occur within a single dataset (duplicate records) and other times they occur in multiple datasets (matched records). See Appendix A for a description of this process.

Each dataset covered slightly different periods of time (see Section 2 for more detailed descriptions of each individual dataset) so this comparison of records was conducted over three time periods. For March 2011-December 2011, Benetech examined March 15, GoSY, SOHR, SS, VDC, and SNHR. The March 15 group stopped collecting data in December 2011, so records from this source were only included in the first ten months of analyses. Similarly, the government data only extended until March 2012, so for the period of January 2012-March 2012 Benetech examined GoSY, SOHR, SS, VDC, and SNHR. Finally, the four remaining datasets (SOHR, SS, VDC, and SNHR) include records through November 2012.

Benetech also examined data from the Syrian Revolution General Council (SRGC). Data from this group covered the period from March 2011 to January 2012. Benetech learned that the Syrian Network for Human Rights was a spin-off of the Syrian Revolution General Council, so the records of these two groups were compared before comparing them with March 15, GoSY, SOHR, SS, and VDC. From the time period covered by SRGC, 90.2% of killings recorded by SRGC were also recorded by SNHR. Considering the high level of overlap, the contextual knowledge that SNHR was originally a part of SRGC, and the fact that SNHR's dataset covers a longer period of time, Benetech chose to combine the SNHR and SRGC datasets into a single dataset, referred to in the following sections as only SNHR.

This comparison of records has only been possible for records with sufficient identifying information - the name of the victim, plus the date and location of death. Each dataset considered in this study included a number of records which lacked this information. Table 1 lists the number of records from each dataset included in the analyses presented in this report (those with sufficient identifying information) and the number of records excluded from these analyses (those lacking sufficient identifying information).

It is worth noting that none of the included counts in Table 1 match the total number of documented killings — 59,648 — because each dataset contains records that none of the other groups documented, duplicates within the dataset, as well as records that are common to two or more datasets.

Table 1: Number of Records Included and Excluded in Analyses

Dataset	Identifiable Records	Unidentifiable Records
GoSY	2,539	10
March 15	4,195	165
SOHR	29,521	232
SS	33,617	9,769
SRGC	6,206	369
SNHR	33,151	5,397
VDC	38,120	1,984

Based on a comparison of records from March 15, SOHR, GoSY, SS, VDC, and SNHR (combined with SRGC), Benetech found that the seven datasets document a total of 59,648 unique records of killings between March 2011 and November 2012. Of those documented killings, 76.1% are male victims, 7.5% are female victims, and 16.4% of records do not indicate the sex of the victim.

2 Descriptive Statistics for Individual Datasets

Prior to matching and comparing specific records across datasets, Benetech examined summary statistics for each individual dataset. This section presents those basic summary statistics. It must be noted that the analyses presented in this section describe only identifiable victims documented by each individual dataset; unobserved and unidentifiable killings are not considered. Therefore the analysis is affected by selection bias, that is, differences between what can be seen in the analysis and the true patterns that result from patterns in common among the unobserved killings. Selection bias is an inevitable outcome when certain events are more or less likely to be observed and recorded based on the characteristics of both the event and the data collection organization. For example, one group may have better contacts within a certain ethnic group or region, whereas another may have access to government personnel records. Another group may have excellent sources one week and be unable to contact these sources at other times. And of course, there are also violent events that occur, but are not reported to any source, either because only the perpetrators survived the event, or because surviving witnesses were unable or chose not to report the incident. Individual datasets are useful for case studies and as inputs to aggregated analyses (like the ones presented in the following sections), but on their own they are not suitable for drawing conclusions about statistical patterns.

Nonetheless, analysis of the individual datasets explores what has been seen. This analysis is called "descriptive" because it describes the data. Although this may not provide much insight into the unobserved true patterns, descriptive analysis shows what the datasets have in common, and how they differ.

These descriptive statistics only include records of identifiable victims. Records of identifiable victims include the victim's name, plus date and location of death. The full identifying information is essential for the record comparisons required to match records across different datasets. Records lacking the complete information are considered 'anonymous' and were excluded from the integration and analysis. The anonymous records describe victims of violence in the Syrian Arab Republic who deserve to be acknowledged. However, they cannot be included in the analysis as observed victims because it is impossible to determine if the records with partial information refer to killings also described by other records. That is, anonymous records cannot be matched or de-deduplicated. Records with partial information provide hints about the existence of killings which have not been fully documented; a full accounting of the undocumented killings will require more information and additional data analysis.

Figure 1 shows the frequency of reported killings by week for each dataset. Four datasets, SOHR, VDC, SNHR, and SS indicate roughly comparable patterns of violence over time (as also indicated in Figures 13–16 in Section 3). Note though that VDC reports more killings than SNHR and SOHR; VDC, SNHR, and SOHR report more killings than SS. The patterns of violence recorded by the remaining two datasets, March 15 and GoSY, look quite different. The pattern shown by March 15 approximately tracks SOHR, VDC, SNHR, and SS, but the similarity is difficult to see in these graphs because March 15 documents so many fewer cases. The variation in 2011 in SOHR, VDC, SNHR, and SS is much smaller than the variation in 2012. Because March 15 stopped documenting killings in December 2011, its pattern seems different. Data from the Syrian government includes very few records after March 2012 and shows a February 2012 peak that is not found in the other datasets.

Although four of the datasets (SOHR, VDC, SNHR, SS) indicate a substantial increase in documented killings over time, it is important to note that these are recorded killings and this increase may reflect an overall increase in violence or an increase in documentation efforts and therefore in *records* of violence. Alternatively, it may be that documentation has weakened over time, which would mean that violence has increased even more than shown in Figure 1. Because this report includes only the fully-identified reported deaths, it is impossible to rigorously distinguish between these alternatives.

Figure 1: Distribution of Reported Death Dates by Week

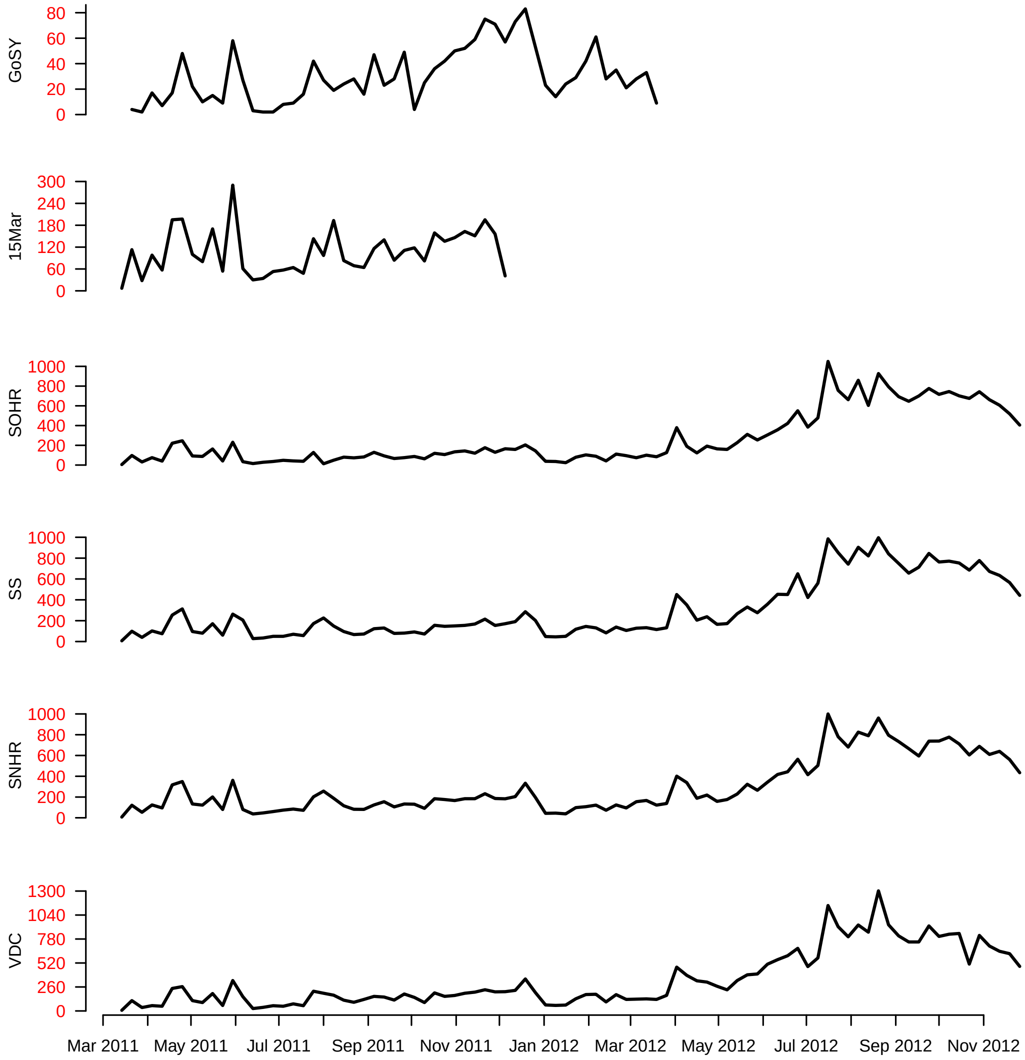
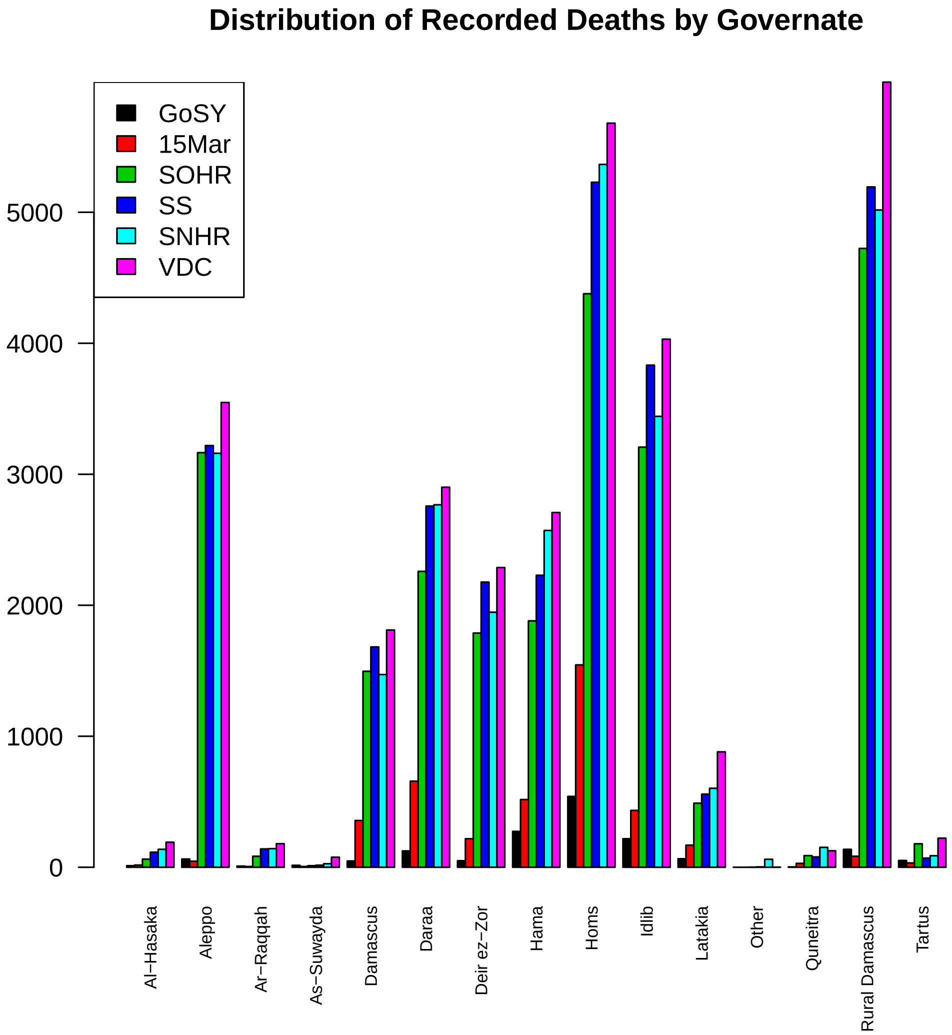


Figure 2 compares patterns of violence over geographic area across all six datasets. As in Figure 1, SOHR, SS, SNHR, and VDC indicate comparable patterns. All four groups record the highest number of killings in Homs, followed by Rural Damascus and Idlib, Aleppo, Daraa.

March 15 and the GoSY dataset also report the highest number of recorded killings in Homs. However, March 15 reports the next highest number of recorded killings in Daraa, Hamaa, Idlib, and Damascus. The government dataset reports the highest number of recorded killings in Homs, followed by Idlib, Hama, Rural Damascus, Daraa, and Aleppo. However, the pattern in the GoSY data may be distorted: Benetech cannot be certain, but it seems that in some cases, the location recorded in the GoSY dataset corresponds to the governorate of birth of the victim, rather than the location of the death, as it is recorded in most of the other datasets. This may account for differing patterns of violence recorded by governorate in the GoSY dataset as compared to the others. There is the possibility of similar location confusion in the SOHR data, however this dataset included an additional variable which indicated location of death if it differed from location of birth.

Figure 2: Distribution of Recorded Deaths by Governate



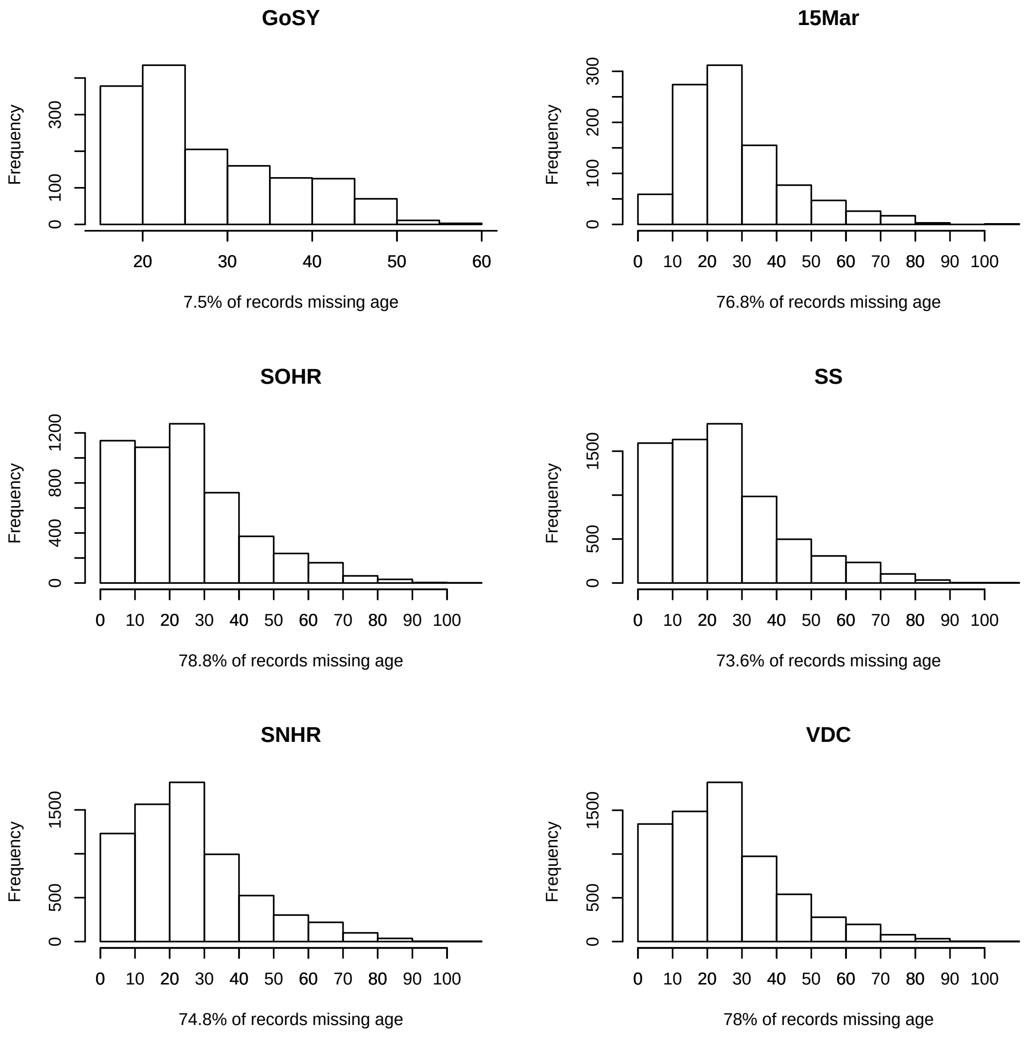
All six datasets include information about the sex (Table 2) and age of victims (Figure 3). There seems to be general agreement across the datasets that the vast majority of victims are male, however March 15 and SOHR contain the most missing information with regard to sex.

Table 2: Sex

Dataset	Female	Male	Unknown
GoSY	0	2,534	0
March 15	109	2,407	1,667
SOHR	2,517	21,357	5228
SS	3,012	30,476	0
SNHR	3,032	26,096	3,740
VDC	3,398	34,498	4

As indicated in Figure 3, these six datasets indicate a similar *reported* age distribution pattern; the majority of victims for whom age is reported are under 40 years old. While the March 15 data has relatively few children less than ten years old, the SOHR, SS, SNHR, and VDC datasets show substantial numbers of young children. It could be that more children have been affected in 2012, after the March 15 group stopped their documentation efforts. However, many records are missing indication of age. Consider the histograms in Figure 3. With the exception of GoSY, the remaining datasets are all missing information on age for over 70% of records. The records without ages could have substantially different ages than the records with reported ages. For example, the age of very young people and very old people is often relevant to their identity. "He was only four years old" or "he was over seventy years old" are common phrases, but there is no comparable salience for an adult's age. It may be that most or all of the records with missing age data are in fact adults, which would make most distributions look more like the GoSY or 15 March patterns. The high proportion of missing age data prevents us from drawing conclusions about the true distribution of the age of victims reported to each group.

Figure 3: Age Distribution


GoSY

7.5% of records missing age

15Mar

76.8% of records missing age

SOHR

78.8% of records missing age

SS

73.6% of records missing age

SNHR

74.8% of records missing age

VDC

78% of records missing age

3 Data Overlaps

Once records have been matched and deduplicated, it is possible to compare each pair of datasets (VDC and SS, SS and SNHR, etc.) according to the number of records they have in common, relative to the number of records documented by each dataset alone. Since each data source includes records from slightly different periods of time, the following sections describe comparisons of pairs of datasets organized according to the periods of time they cover.

Figures 4–16 show this overlap analysis over time and space (each panel is a single governorate, that is, a Syrian province; six governorates were selected to display as examples) for each combination of pairs from the datasets listed in the previous section. The light and dark blue portions of each bar indicate records in a single dataset and the pink portions indicate identical records shared by both groups.

For March–December 2011 datasets were compared over three to four month periods - from March to May, June to August, and September to December. Comparisons for 2012 were conducted for each individual month.

3.1 March 2011-December 2011

Figures 4–7 show that the March 15 group shares roughly the same proportion of records of killings with VDC, SS, SOHR, and SNHR. In three of these Figures, VDC, SS, and SNHR appear to document the majority of killings in five of the six governorates displayed (the lighter blue portions of the bars in each Figure, indicating killings recorded only by VDC, SS, or SNHR). Exceptions to this pattern are Damascus, where March 15 appears to be documenting the majority of recorded killings (the darker blue portions of the bars in the Damascus section of each Figure) and in Daraa between March and May. This pattern mostly holds for the overlap between March 15 and SOHR as well (Figure 7), however SOHR also appears to document fewer deaths in Hama than the other groups, resulting in roughly comparable numbers of deaths recorded in Hama by March 15 alone and by SOHR alone, with relatively few overlapping records. It is worth further investigation to determine if March 15 had sources of information in Damascus and Daraa that were not available to the other groups and if these other groups had sources of information in Hama not available to March 15 and SOHR.

In general, the records of killings documented by March 15 appear to be somewhat different from those documented by VDC, SS, SOHR, and SNHR. This can also be seen in Section 2 in terms of the period of time covered by March 15, the number of killings recorded, and the number of records in each governorate. In all of these attributes March 15 differs somewhat from VDC, SS, SOHR, and SNHR; Benetech underlines the complementary importance of the March 15 data for understanding 2011.

Figure 4: Distribution of Records between March 15 and VDC

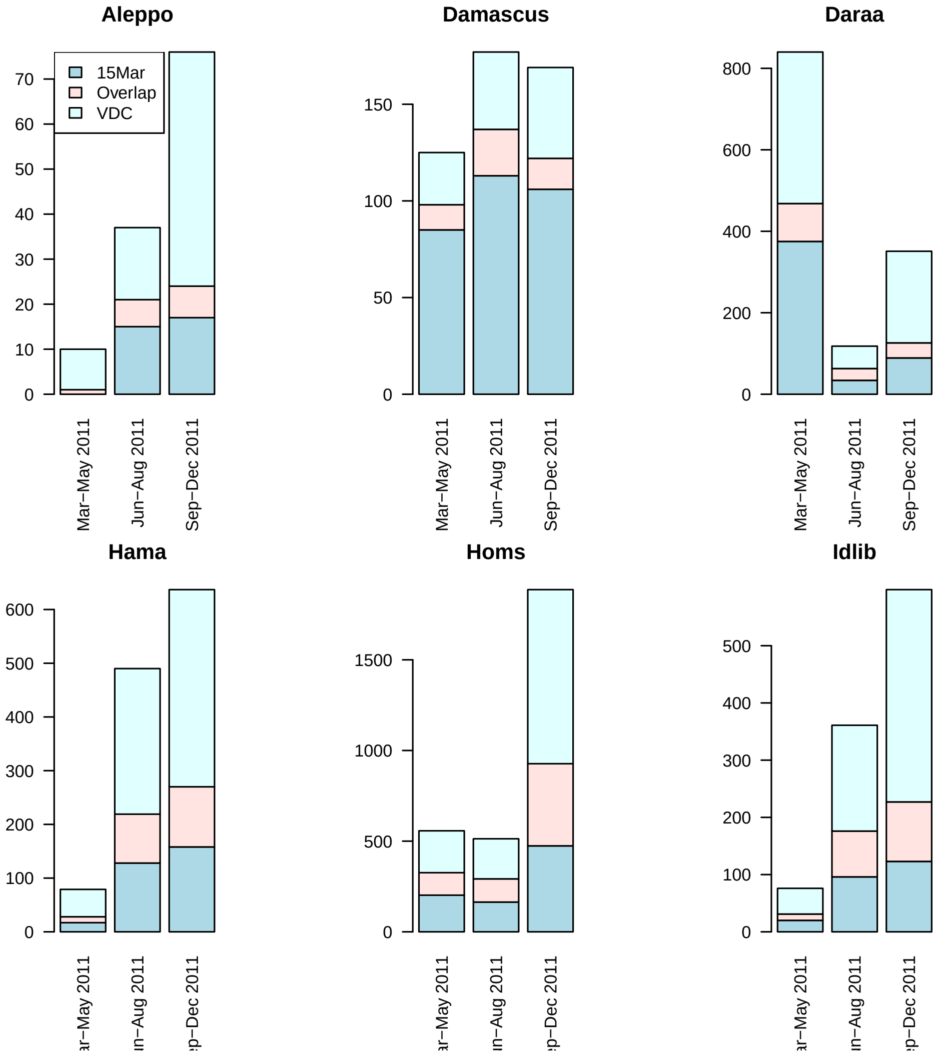


Figure 5: Distribution of Records between March 15 and SS

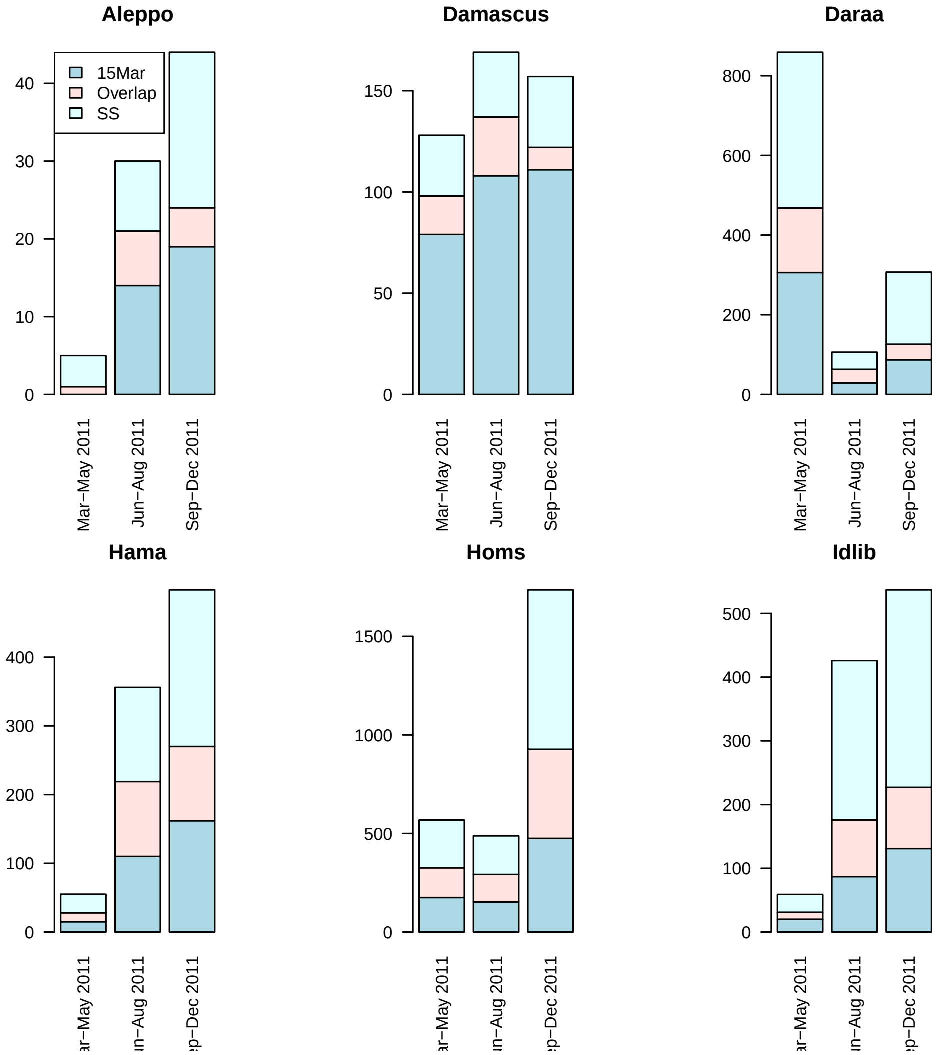


Figure 6: Distribution of Records between March 15 and SNHR

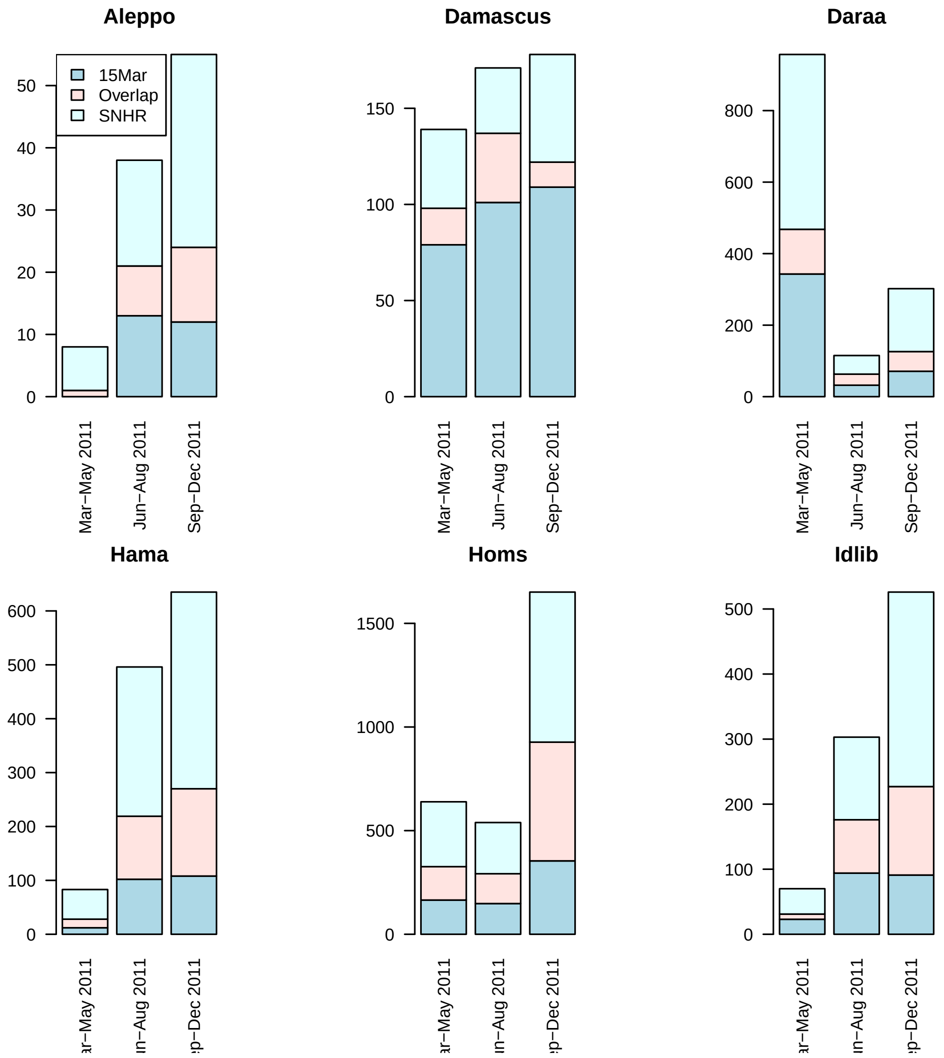
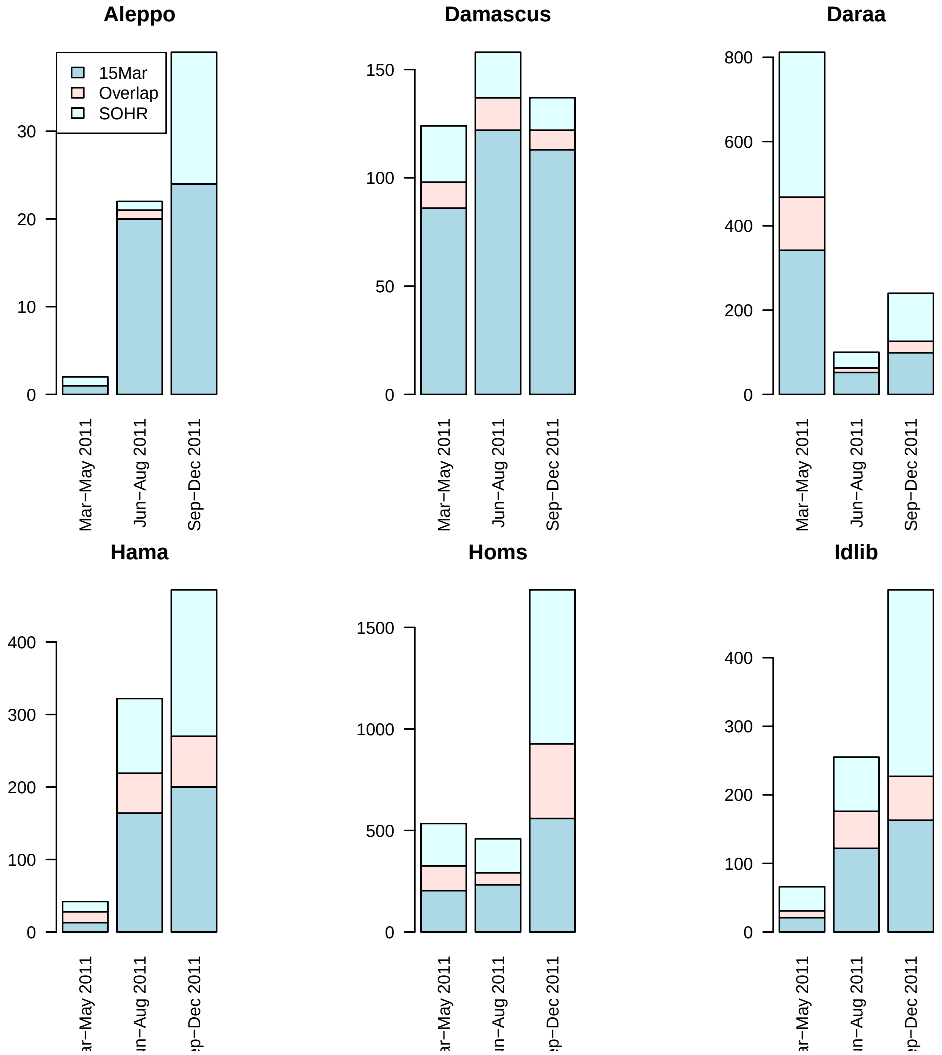
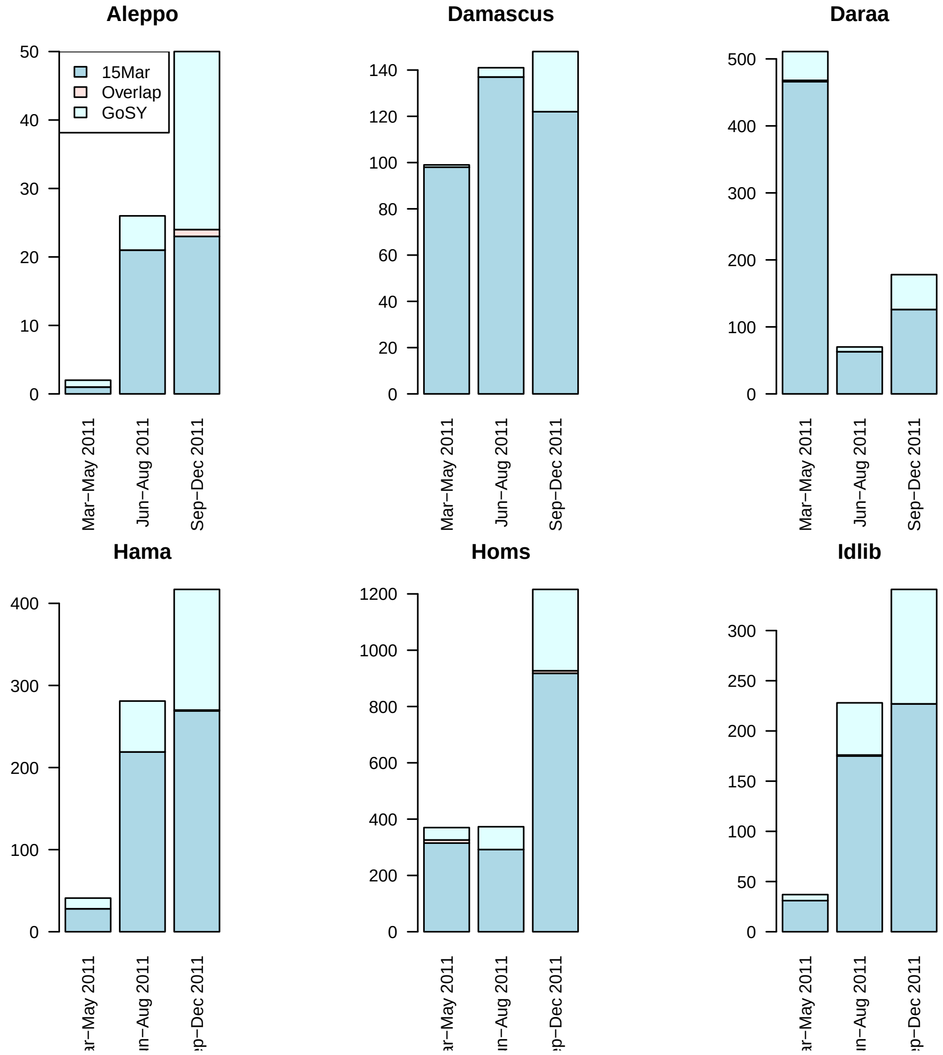


Figure 7: Distribution of Records between March 15 and SOHR



In contrast, Figure 8 indicates virtually zero records shared between March 15 and GoSY - very little pink sections, indicating overlapping records, are visible in Figure 8. This is a pattern seen again in the following section comparing GoSY and SS, SOHR, VDC, and SNHR. Additionally, with the exception of the last four months of 2011 in Aleppo, Damascus, Hama, Homs, and Idlib, GoSY appears to record very few documented killings (the small light blue portions of each bar in Figure 8).

Figure 8: Distribution of Records between March 15 and GoSY



3.2 March 2011-March 2012

Five of the datasets (GoSY, SS, SOHR, SNHR, and VDC) include records for one year, from March 2011 to March 2012. Figures 9–12 show these comparisons. As in Figure 8 we see that the government data source shares very few records in common with the other data sources. Figure 9 demonstrates this most dramatically, with virtually no visible pink sections in any of the bar charts, indicating records in common between GoSY and SS.

Figures 10 and 11 show a slightly higher number of records shared between GoSY and SNHR and GoSY and SOHR (as compared to GoSY and March 15 or GoSY and SS) with notable overlaps (pink sections) in Aleppo, Damascus, Hama, and Homs. As with the comparison of March 15 and GoSY in Figure 8, overall the GoSY dataset records a much smaller number of documented killings than the other datasets (see in Figures 9 – 11). The dark blue portions of the bars in Figures 9–11 are quite small, indicating few killings documented only by GoSY. This can also be seen in the descriptive summary of GoSY in Section 2 - this dataset includes far fewer records than any of the other datasets.

Lastly, Figure 12 shows that although VDC also has relatively few records in common with GoSY, compared to March 15, SS, SNHR, and SOHR, VDC has the most overlap with GoSY. Figure 12 shows noticeable overlap (pink sections of bars) between the two groups in Aleppo, Damascus, and Hama, with slightly smaller overlap sections in Daraa, Homs, and Idlib.

Figure 9: Distribution of Records between GoSY and SS

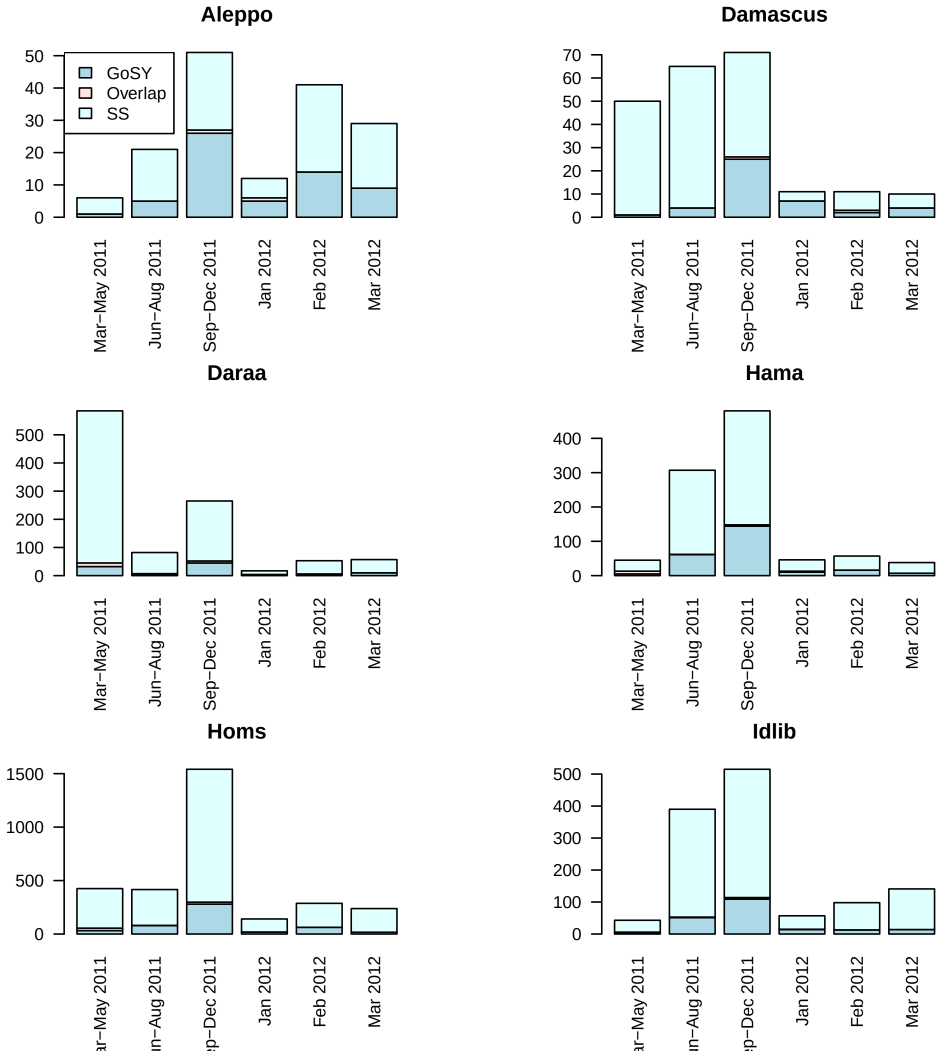


Figure 10: Distribution of Records between GoSY and SNHR

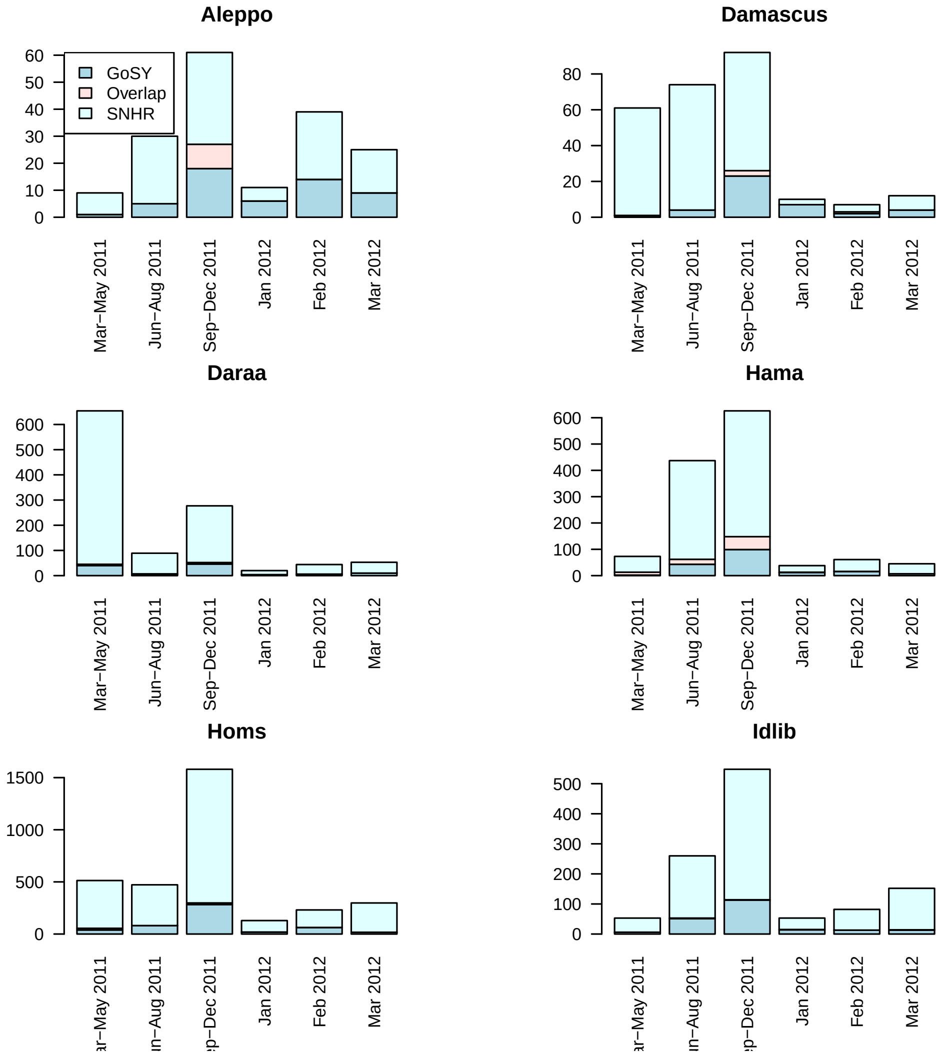


Figure 11: Distribution of Records between GoSY and SOHR

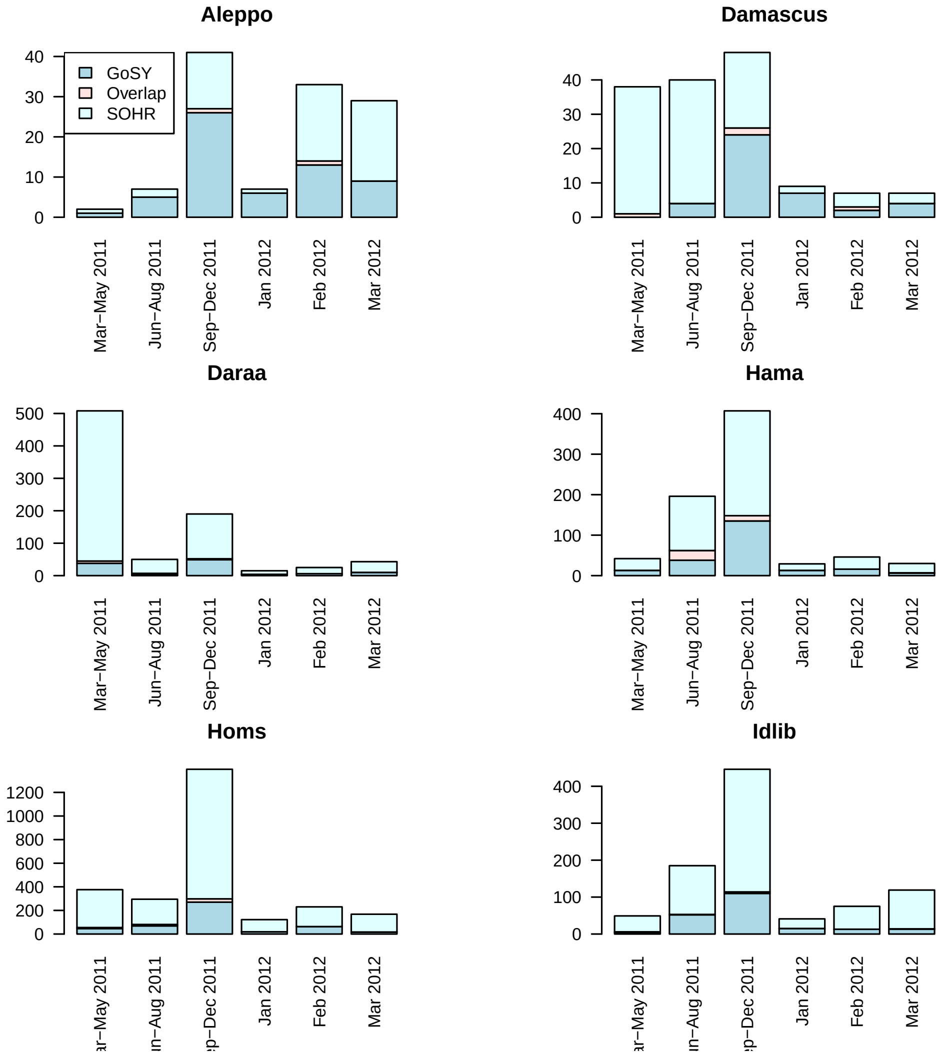
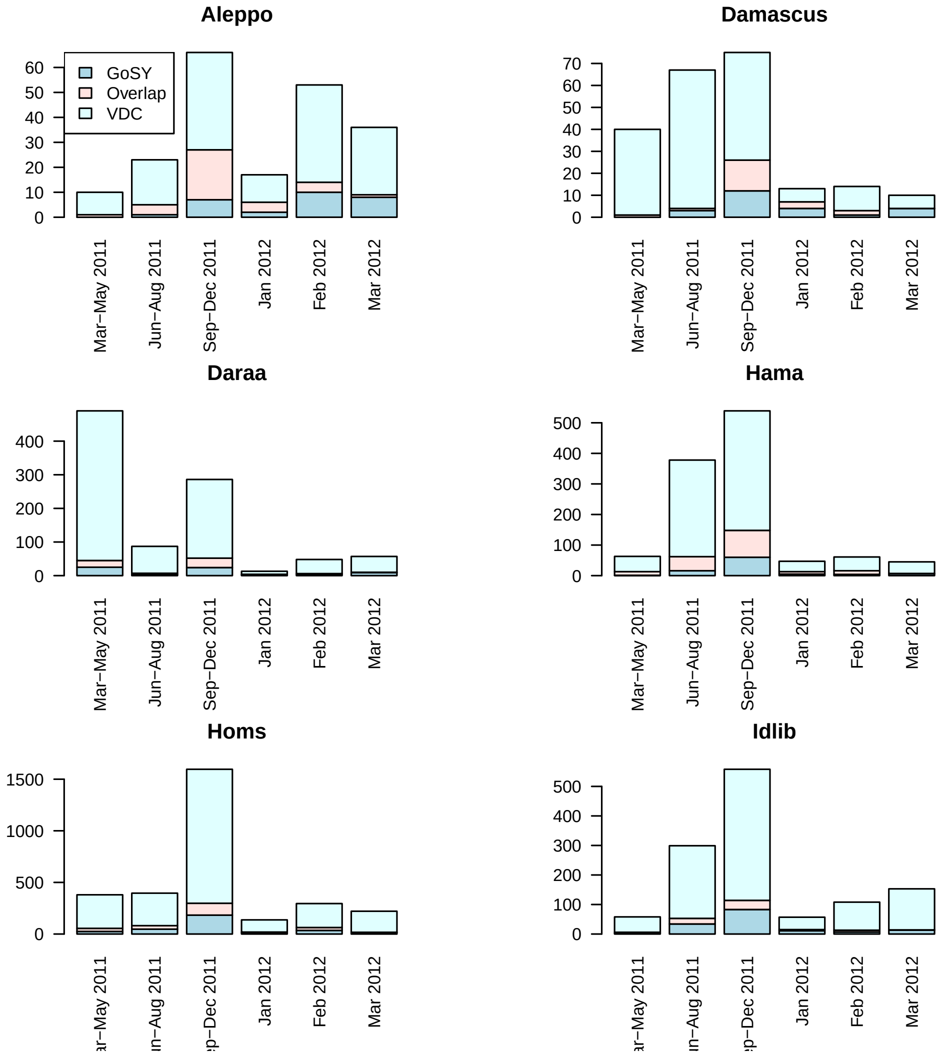


Figure 12: Distribution of Records between GoSY and VDC



These analyses present two possibilities in terms of the GoSY data. It is possible that the records included in the GoSY data are inaccurate, although there is no reason to believe this is the case. Alternatively, if the GoSY records are accurate, and we have no reason to believe they are not, they appear to be covering a different 'universe' from the other data sources.

That is, it may be that the GoSY dataset primarily includes records not included in any of the other five datasets, and it largely excludes records included in one or more of the other five datasets. This may indicate that the GoSY dataset has access to different sources, or has different criteria for recording a death, or is in some other way documenting a different universe of violence than the other five datasets.

3.3 March 2011-November 2012

Four datasets (SOHR, SS, SNHR, and VDC) cover the entire time period under study (March 2011 - November 2012). As Figures 13–16 show, there is broad agreement between these four datasets and this agreement appears to increase over time. In most governorates, the size of the pink portion of the bar graph, indicating records in common between each pair of datasets, increases over time. In general, these four data sources appear increasingly to be sharing sources documenting killings in the Syrian Arab Republic.

However, there are notable exceptions to this broad overlap, and these are worth highlighting as they may point to times or locations when one group had access to information that another group lacked. For example, Figure 13 indicates both SNHR (light blue) and VDC (dark blue) were documenting killings unrecorded by the other group in Daraa between March and May 2011. A similar pattern can be seen between SS and SOHR in Figure 15 - there is a notable lack of overlapping records, as compared to records documented by only one of the groups, in Daraa between March and May 2011, Hama between June and December 2011, Homs between September and December 2011 and Idlib between June and December 2011. Such reflections may help each group to further improve their documentation efforts.

Figure 13: Distribution of Records between VDC and SNHR

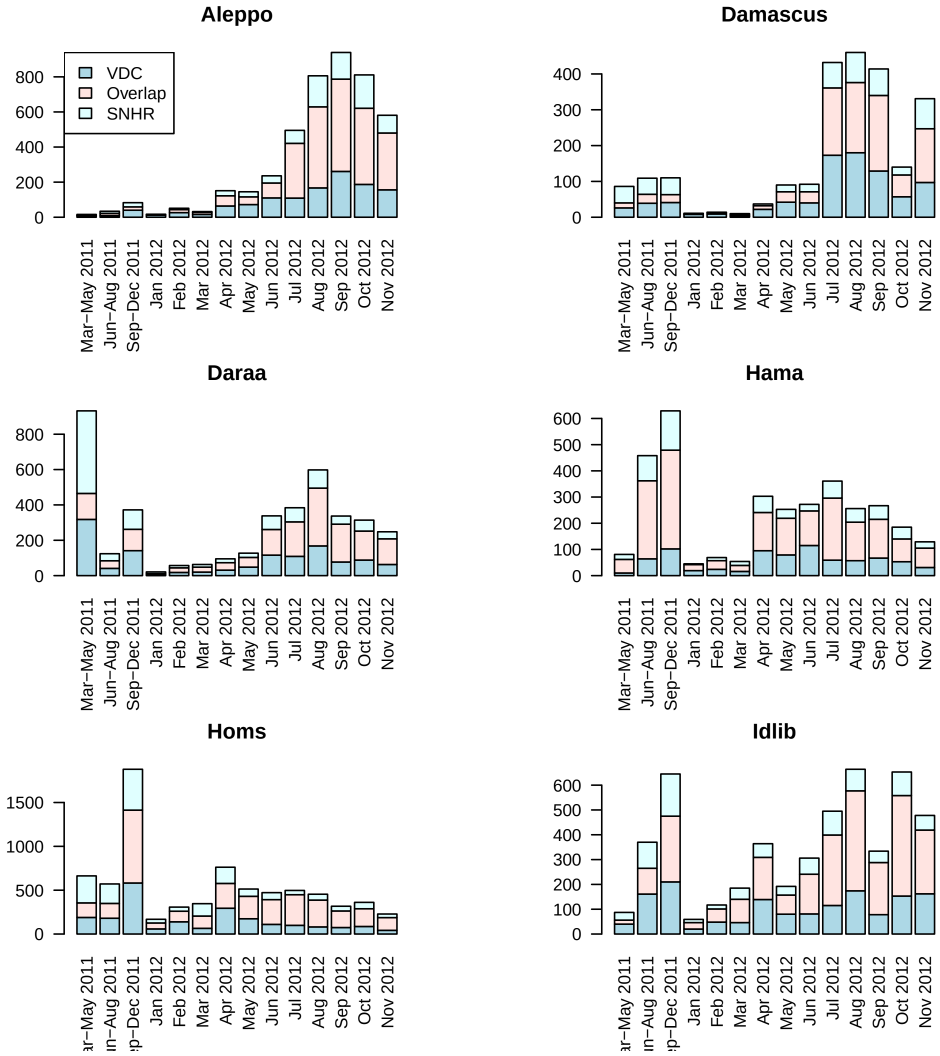


Figure 14: Distribution of Records between VDC and SS

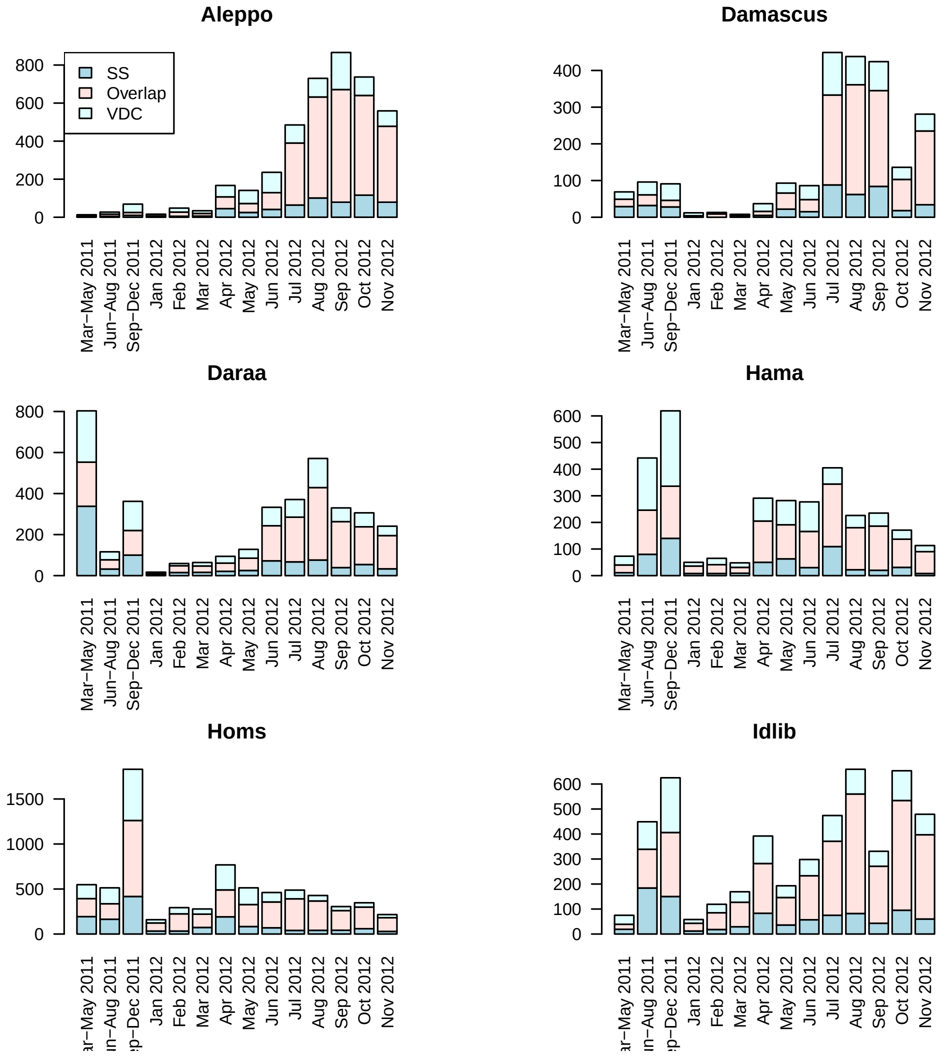


Figure 15: Distribution of Records between SOHR and SS

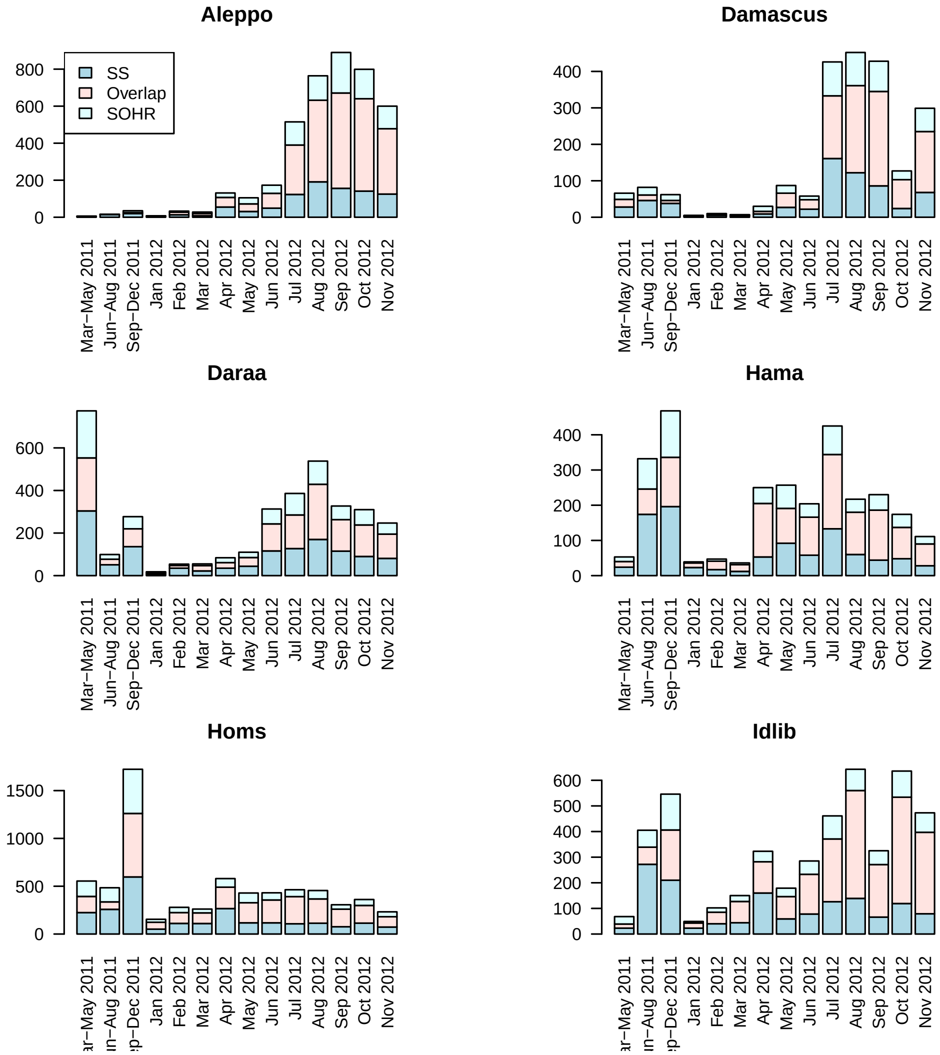
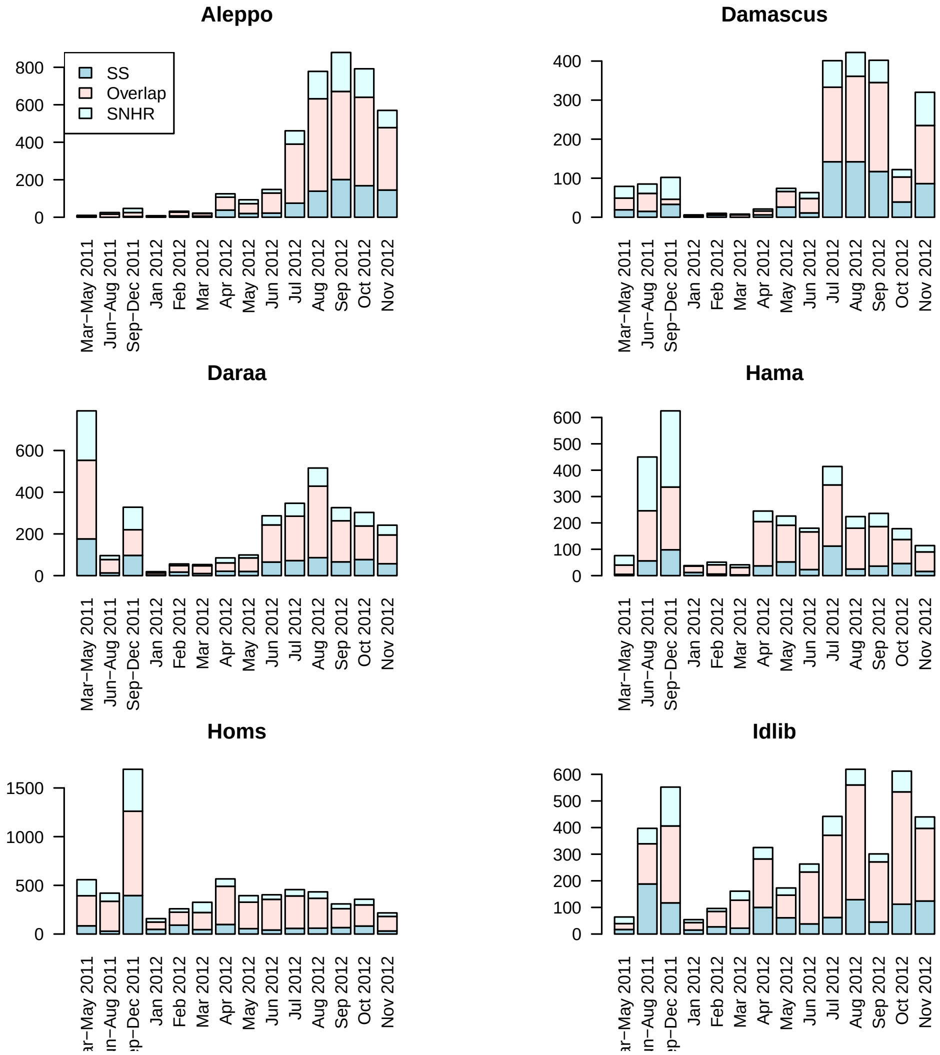


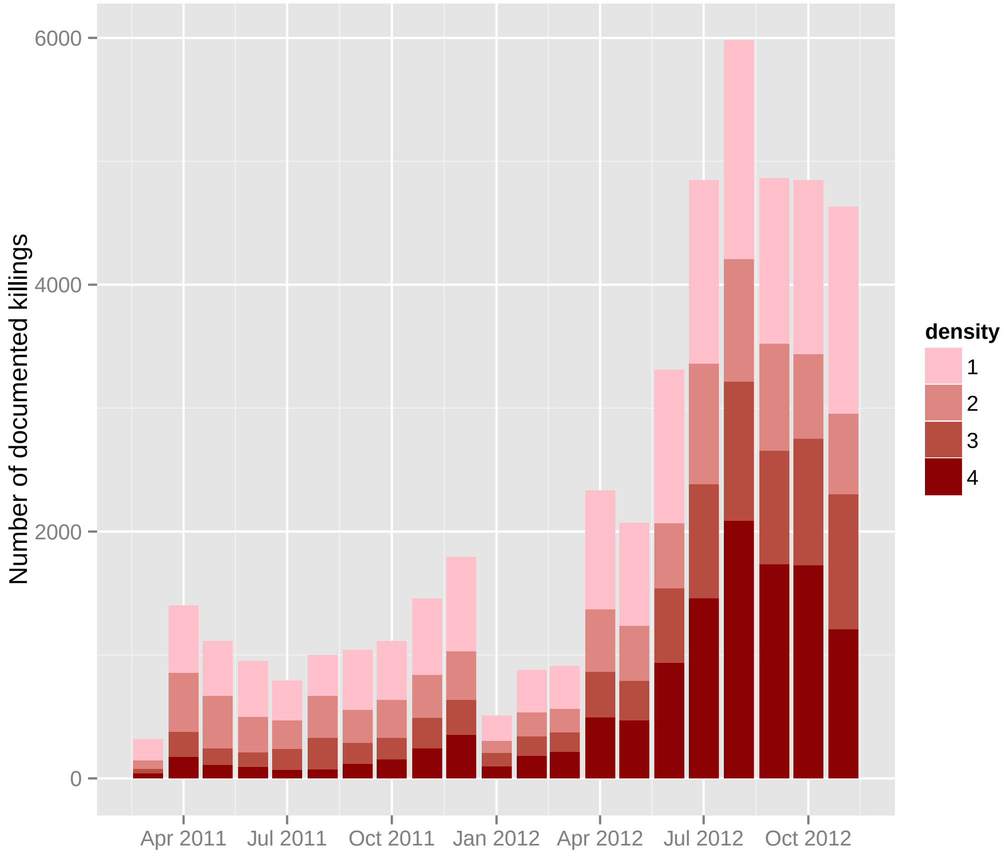
Figure 16: Distribution of Records between SNHR and SS



3.4 Patterns of Overlap over Time

This report began with a warning that despite the enormous efforts by the data collecting groups, many killings in the Syrian Arab Republic are still undocumented. One way to imagine that is to consider that in any particular month, some killings are documented by four groups, other killings are documented by three groups, others by two groups, and some killings are reported by only one group. The question this observation raises is: how many killings are reported by zero groups?

Figure 17: Documented Killings by Month and by Number of Sources per Killing



One way to visualize the intensity of reporting is shown in Figure 17 . This graph includes only the killings documented by the four datasets that cover the entire period (SS, VDC, SOHR, and SNHR). In a very informal sense, as more killings are documented by all four groups, the intuition is that there are fewer undocumented killings. By contrast, when a greater proportion of killings are documented by only one group, the intuition is that there are probably relatively more killings that have not been documented at all. The key observation from Figure 17 is that in all months, at least some killings are reported by only one group (represented by the light pink part at the top of each bar). It is therefore very likely that there are substantial numbers of killings undocumented by these four groups; in practice, many killings may be undocumented by any project. Comparing October and November 2012, note that a higher proportion of killings were documented by four projects in October than in November; similarly, a higher proportion of killings were documented by one project in November than in October. The implication is that there may be more undocumented deaths in November than in October; the true estimate is affected by many additional factors which limit the scope of this simple comparison. In a subsequent report, Benetech will address the question of undocumented killings directly through statistical modeling.

By comparing pairs of datasets over time and space, and by considering the density of overlap over time (and in the future, over space), an analyst can get a better sense of the state of data collection in the Syrian Arab Republic. Statistical modeling and estimation could provide insights about the total magnitude and true pattern of all killings, including those that have not been documented. These comparisons can be used to help individual groups improve their data collection. At a conference in Brussels in early July 2012, hosted by the Euro-Mediterranean Human Rights Network[7], Benetech heard repeatedly that monitoring groups are eager to coordinate their documentation efforts. Benetech is optimistic that most monitoring groups will welcome comparative information like that presented in this report.

[7] http://www.euromedrights.org/en/

A Matching

As mentioned in Section 1, to use the records described in this report, they must be linked together, identifying the records which refer to the same people. This is challenging, since each data source records slightly different information (as indicated by Section 2), not to mention each data source is working to overcome the difficulties inherent in collecting complete, accurate information in the midst of a conflict.

A.1 Non-technical matching overview

The linking together of records within each system is called de-duplication, and identifying the same death across different sources is called record linkage. Both are performed together, by starting with a single list of all records with sufficient information, including sex, age or date of birth, name, and date and location of death.

The records were divided in three groups, called partitions. The first includes data from seven sources (SOHR, SS, VDC, 15 March, SNHR, SRGC, and GoSY) during March to December 2011. The second partition includes six sources (SOHR, SS, VDC, SNHR, SRGC, GoSY) for January to March 2012. The third partition includes four sources (SOHR, SS, VDC, SNHR) for April to November 2012.

From the full list, all possible pairs of records are generated. There are hundreds of millions of possible pairs, and this is much more data than can be processed. To reduce the number of candidate pairs (that is, the set of pairs from which the computer algorithm is identifying plausible pairs) Benetech limits pairs to plausible pairs of records by excluding records that cannot plausibly represent the same person. Rejection rules include differences between records that preclude them from being considered a candidate pair, such as records with different sex, with locations of death that are geographically distant, or with dates that are widely separated in time. The remaining pairs are called candidate pairs: there are 3.8 million candidate pairs in the first partition, 3.0 million in the second partition, and 11.5 million candidate pairs in the third partition.

In the next step, Benetech generated numeric comparisons among all the candidate pairs and all the training set pairs, which is a way to summarize how 'similar' records are in terms of variable name spellings, date of death, date of birth, and location of death. Many different comparisons are done with each field. For example, two dates could be equal; they could vary by a certain number of days; they could be the same date with the month and day fields interchanged; in this project, Benetech considered eighteen comparisons among pairs of records. Then a 'training set' was generated which included a list of both plausible pairs and records identified as non-matches by the rejection rules. A human being (a Syrian expert, in this case) reviewed the training set and classified pairs as referring to the same person (a match) or to different people (a non-match). This data was used to teach the computer how to classify all the pairs as matches or non-matches. The expert examined 8,280 pairs of records for this step in the matching process.

From the numeric comparisons on the training set, Benetech calculated a model which predicts which pairs of records refer to the same person and which refer to different people. The model is calculated from the training set and applied to all the millions of candidate pairs.

With a set of all the pairs identified by the model as matches, the matched pairs are combined into groups of records which all refer to the same person. This process is called clustering. Lastly, the records in each cluster are merged into a single record containing the most precise information available from each of the individual records.

A.2 Matching technical details

Matching databases using partial information has a long history - first formulated by Dunn (1946) and Newcombe et al. (1959), and approached theoretically by Fellegi and Sunter (1969).[8] Specifically, Benetech used the iterated procedure described in Sarawagi and Bhamidipaty (2002). As described in the previous section, the data were divided into three partitions so that different combinations of sources could be matched in each partition. The expert reviewed a total of 8,280 pairs of records drawn from the

[8] See the reviews of the problem, called variously "record linkage,", "matching," and "database deduplication" in Winkler (2006) and Herzog et al. (2007). A key method is approximate string distance, see Levenshtein (1966).

various sources (called the "training pairs"). He classified the training pairs as either referring to the same person or to different people. Using the training pairs, Benetech generated a computer model called an Alternating Decision Tree (ADT).[9] Benetech implemented a method for transliterating and comparing names written in Arabic and Latin script developed by Freeman et al. (2006). The model classified all the possible pairs of records from all seven datasets as referring to the same person (a *match*) or to different people (a *non-match*). When tested against the examples, averaged across the partitions, the model classified 92% of the training pairs accurately. In a stratified 10-fold cross-validation, the kappa statistics for the three partitions were 0.78, 0.77, and 0.79, respectively. The pairs were organized into larger groups of records that refer to the same person by a method called "clustering." Benetech used a Hierarchical Agglomerative Clustering algorithm (Manning et al., 2008). Records in each cluster were merged to preserve the most specific information in the group.

References

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health*, 36(12):1412–1416.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Freeman, A. T., Condon, S. L., , and Ackerman, C. M. (2006). Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 471–478.

Freund, Y. and Mason, L. (1999). The Alternating Decision Tree Learning Algorithm. In *Sixteenth International Conference on Machine Learning, Slovenia*.

[9] For an overview of machine learning techniques for classification and clustering, as well as a description of the software Benetech used for classification, the Weka software, version 3-7-4 (Hall et al., 2009), see Witten et al. (2011). The ADTree software is documented at `http://weka.sourceforge.net/doc/weka/classifiers/trees/ADTree.html`. The algorithm for ADT was first described by Freund and Mason (1999) and optimized by Pfahringer et al. (1996).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter*, 11:10–18.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques.* Springer.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, 130(3381):954–959.

Pfahringer, B., Holmes, G., and Kirkby, R. (1996). Optimizing the Induction of Alternating Decision Trees. In *Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.*

Sarawagi, S. and Bhamidipaty, A. (2002). Interactive Deduplication Using Active Learning. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278. ACM Press.

Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. Technical Report RRS2006/02, Statistical Research Division, U.S. Census Bureau.

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman.

About the Benetech Human Rights Program

The Benetech Human Rights Program has more than 20 years of experience applying statistical analysis to data about human rights violations. Our expertise has been sought by nine Truth and Reconciliation Commissions, by UN missions and official human rights bodies, by international and domestic criminal tribunals, and by many non-governmental human rights organizations. We have conducted projects in El Salvador, Ethiopia, Guatemala, Haiti, South Africa, Kosovo, Sierra Leone, Sri Lanka, Timor-Leste, Colombia, Perú, Liberia, and the DR Congo, among others; and provided extensive guidance on data processing and analysis methodologies to non-governmental organizations and partner groups in many countries throughout the world. With our partners, we make scientifically-defensible arguments based on rigorous evidence.[10]

This project was commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR). Additional funding was provided by the Sigrid Rausing Trust, the Oak Foundation, and a private, anonymous US-based foundation donor through their core support to the Benetech Initiative.

The materials contained herein represent the opinions of the authors and editors and should not be construed to be the view of the Benetech Initiative, any of Benetech's constituent projects, the Benetech Board of Directors, the donors to Benetech or to this project, or of OHCHR.

[10] (http://www.benetech.org, http://www.hrdag.org).