

Shuffling Methodology for Sanitizing TCAPF Microdata

Dustin Burke, dburke@milcord.com
March 1, 2010

The following describes the methodology used for preparing Tactical Conflict Assessment Planning Framework (TCAPF) microdata using data sanitization techniques for the purpose of publishing analysis of the dataset while preserving the operational security (OpSec) for the military collecting units (United States Military and coalition partners) as well as the confidentiality of the survey responders (Afghanistan citizens).

The following table describes the columns of the TCAPF dataset.

Table 1 - TCAPF Fields

Field	Description
Unit	Collecting unit for TCAPF questionnaire
Village	Afghanistan village where TCAPF questionnaire was conducted
Date	Date record was collected
Occupation	Occupation of survey responder
Question 1	“Has population changed within last 12 months?”
Question 1-Why?	“Why change in population?”
Question 2	“What is the major problem in your village?”
Question 2-Why?	“Why is it a major problem?”
Question 3	“Who do you believe can solve problems?”
Question 4	“What should be done first?”

The sensitive fields are shown in yellow whereas the green fields are the columns we would like to preserve for their analytic value.

Assumptions and Preventative Measures

We assume that the survey responses themselves and the survey responder’s occupation are not sensitive information. These fields are recorded by mapping the user’s response to a taxonomy of acceptable responses. For instance, the responses to Question 2 are selected among:

Table 2 - Taxonomy of Acceptable Values for Question 2

Major Problem?	
FOOD	ELECTRICITY

POTABLE WATER	AGRICULTURAL
PAID WORK OPPORTUNITIES	FUEL
GOVERNMENT	CRIMINAL VIOLENCE
SHELTER	JUSTICE
HEALTHCARE	CORRUPTION
EDUCATION	INFRASTRUCTURE
PHYSICAL SECURITY	LAND / WATER DISPUTE
SANITATION	TRIBAL

This mapping ensures that unique and personally identifiable (through inference or correlation with outside sources) responses are removed. In the case of the Occupation field, this is especially important since if unique responses were allowed (not mapped) and a survey responder held a unique position (such as “President of the United States”) then this would result in a disclosure of sensitive information about the survey responder.

We further assume that the list of acceptable values for these fields is not sensitive as it could readily be enumerated from the published microdata. Without the disclosure of this information, the analytical value of the unclassified release of this dataset would be essentially zero.

It should be noted that there is a chance (although very low likelihood) that these information fields could be used to infer the village and date when the survey responses were collected if correlated with external datasets. For instance, the distribution of occupations could be correlated with information about Afghanistan’s labor markets at provincial/district/village levels to infer likely areas where the survey was conducted. For instance, if responses from village X (sanitized by redaction) were 90% for “farmer” occupations then we could use a census dataset to find the most likely village or district based on this distribution (areas known to have roughly 90% farmers). We prevent against this type of inference by randomly shuffling the Village field of responses so that the distribution of occupations for a village is randomized and distribution fitting techniques will no longer work.

Additionally, this data shuffling technique prevents against attempts at fitting a time series based on the temporal dynamics of the survey responses for a village. Suppose the survey responses followed a certain pattern where at time T_0 the Major Problem was “Potable Water”, at T_1 was “Physical Security” and at T_2 was “Corruption”. We could use news reports and other similar open sources of content to find a village that fit this pattern (where the temporal distance between $T_0 - T_1$ and $T_1 - T_2$ is known if the sanitized dataset preserved it). From a time-series based pattern inference a malicious person could potentially recover the place (village) and time of the collected survey responses. To prevent this type of inference we randomly shuffled the village and time fields, thereby destroying the temporal distances between events.

We further assume that the likelihood of more sophisticated (or unanticipated) techniques at recovering or approximating the sensitive dataset are remote due to the unique nature of the survey response dataset.

Methodology

As discussed above, our methodology involved a combination of data substitution, redaction (removed columns weren't even mentioned above), and data shuffling. We preserved the following fields: occupation, question 1, question 1-why, question 2, question 2-why, question 3 and question 4.

In concept, think of the survey responders as snowflakes within a snow globe whereby we give the globe a good shake and the snowflakes (left intact) land in random places and at different times. This is essentially how we sanitized the TCAPF dataset - plucking up the survey responders, keeping their responses intact and randomly moving them to a new place and time at which their response was collected. Preserving the responses retains the full analytical value of the dataset at the aggregate level. However, at the more granular level (i.e. at the village level), the analysis isn't accurate.

We substituted the list of military Units with meaningless descriptors (e.g. "2nd Battalion, 3rd Marine" replaced with something like "Unit 1"). We randomized the order in which the ordinal identifiers were assigned so that Unit 1 doesn't correspond in any meaningful way with "1st Battalion" or the largest (or most popular as published in news reports) military unit collecting TCAPF questionnaires. The mapping is totally arbitrary. It's worth noting, however, that we preserved the same NUMBER of units. Additionally, the distribution of TCAPF records over the Units is preserved (meaning if Unit X was mapped to Unit Y and there are N_y records for Unit Y then this implies that there are N_x records for Unit X). If a malicious user possessed external knowledge about the relative distribution of TCAPF collection activities for military units in Afghanistan then they could readily recover the Unit field. Although the unit can be readily recovered in this way, it doesn't constitute a violation of the security introduced by data sanitization since the user already had access to knowledge of the military Units and their relative distribution of TCAPF related activities. It would be fair to assume that United States Military Civil Affairs companies are published publically and detailed in numerous news reports and thus doesn't constitute sensitive information. However, using this information to infer a place and time when this unit was operating and collecting TCAPF responses would constitute the sensitive nature of this data. To prevent against this inference, we randomly shuffle the Unit field so that the distribution of Units across villages, dates and survey responses is randomized and not recoverable.

The Date field was randomly shuffled as well, to eliminate the possibility of using time-series based techniques to pattern-fit and infer the temporal information about the survey responses. Additionally the range of Dates for the survey responses was translated in time by a random number of days so that the likelihood of temporal overlap with the actual collection period is uncertain.

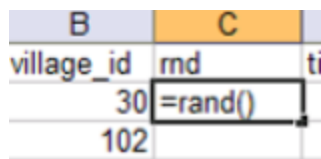
A combination of techniques were used on the Village field, including both data substitution and data shuffling. The unique nature of the TCAPF dataset is such that the list of village names themselves are sensitive since the responses are frequently collected from extremely remote

locations within Afghanistan. Simply shuffling the data values within the village column would disclose too much information about the locations of our military units and present potential operational security consequences. In a previous effort, we attempted to geocode the list of village names against an authoritative gazetteer such as the "Settlements" dataset published by Afghanistan Information Management Systems (AIMS) or the National Geospatial Agency (NGA) GeoNet Names Gazetteer to resolve the placename to a latitude and longitude geo-coordinate. We know first hand the difficulty of performing this task using either manual or automated means. Most placenames in the TCAPF dataset were not able to be resolved to a corresponding record in either gazetteer due to the coverage issues of these gazetteers in austere, remote conflict environments. Since a mapping doesn't exist, we realized the suitability of using the gazetteers as sources for data substitution on the Village field.

We assigned a random mapping from the villages to the villages within the AIMS Settlements dataset. Next, for records with missing village information, we randomly assigned villages from the gazetteer. This approach (roughly) preserves the relative distribution of TCAPF responses across villages, except that the number of villages within the sanitized dataset is larger than the source dataset due to the random assignment for NULL values. Using external knowledge about TCAPF collection activities and the population density within areas, one could potentially recover the reverse mappings for several of the villages (this line of reasoning only applies to urban areas). However, to prevent against this, we randomly shuffle the village column so that inferring the village mappings (of urban areas) would not lead to further compromise of the Unit or Time fields.

Here is the very simple method used for shuffling a column of data within Microsoft Excel.

1. To sort the "village_id" column, create a new column to the right of it.
2. Fill the values of this new column with randomly generated numbers using the rand() function as shown in the screenshot below (Figure 1)
3. Select the village_id column and the column containing the randomly generated values.
4. From the main menu, select "Data -> Sort..." You will see a Sort Dialog like the one shown in Figure 2.
5. Select to sort by the "rnd" field (containing random values) and click OK.
6. The resultant shuffled data column is shown in Figure 3.



The screenshot shows a portion of an Excel spreadsheet. Column B is labeled 'village_id' and contains the values 30 and 102. Column C is labeled 'rnd' and contains the formula '=rand()'. The cell containing the formula is highlighted with a black border, indicating it is the active cell.

B	C	ti
village_id	rnd	
30	=rand()	
102		

Figure 1 - Excel rand() function

	B	C	D	E	F	G	H	I	J	K	
	village_id	rnd	time_by_d	occupation	q1_raw_id	q1_why_ra	q2_raw_id	q2_why_ra	q3_raw_id	q4_raw_id	
2	30	0.615047						1		3	1
1	102	0.132883						27	7	9	22
4	155	0.789857						14	7	9	12
5	79	0.371389						11	16	5	
2	13	0.087055						7	16	14	4
5	169	0.327246						23	2	9	15
4	163	0.586125						11		5	
2	30	0.398391						1		3	1
4	87	0.17712						7		5	4
6	87	0.715958						7		5	4
1	168	0.502063						23		5	15
2	168	0.699644						23		5	15
1	53	0.346401						23		5	15
5	226	0.759025						1		5	1
1	35	0.910374						23		5	15
2	53	0.187512						23		5	15
1	226	0.72399						23		5	15
1	35	0.529568						1		5	1
2	67	0.298744						23		5	15

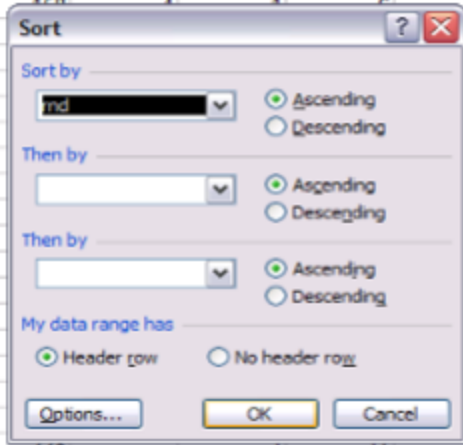


Figure 2 - Sort Dialog (before)

	B	C	D	E	F	G	H	I	J	K
	village_id	rnd	time_by_d	occupation	q1_raw_id	q1_why_ra	q2_raw_id	q2_why_ra	q3_raw_id	q4_raw_id
2	215	0.026242	158	4	1	6	1		3	1
1	30	0.274021	258	7	3		27	7	9	22
4	44	0.939094	274	9	3		14	7	9	12
5	213	0.307897	274	9	2		11	16	5	
2	217	0.788137	274	9	2	17	7	16	14	4
5		0.558593	274	9	1	17	23	2	9	15
4	227	0.42081	228		2	17	11		5	
2		0.749317	156	19	1	6	1		3	1

Figure 3 - Shuffled Data Column (after)

Future Work

The purpose of our data sanitization was to prepare a simulated dataset useful for TCAPF analysis at the aggregate-level and publish a paper showing charts and describing analysis derived from the data. In future work, we would consider additional considerations for releasing TCAPF microdata in unclassified, open source format for aggregate-level analysis and the implications for the types of analysis desired by the consumers of the published microdata.

Conclusion

We presented a novel methodology for sanitizing a TCAPF survey response dataset that preserves the analytical value of the microdata for aggregate-level analysis and the descriptive statistics of the survey responses without compromising operational security for the military units collecting these reports in conflict environments within Afghanistan or disclosing sensitive information about the survey responders (Afghanistan citizens).

References

1. Edgar, Dale. "Data Sanitization Techniques", Net 2000 Ltd. White Paper, 2004. <http://www.orafaq.com/papers/data_sanitization.pdf>

2. Muralidhar, Krish. “*A Primer on Data Masking Techniques for Numerical Data*”, Gatton College of Business & Economics, 2008.
<<http://www.cs.uky.edu/events/dmSec08.ppt>>
3. Muralidhar, Krishnamurthy and Rathindra Sarathy. “*Data Shuffling—A New Masking Approach for Numerical Data*”, *Management Science*, Vol. 52, No. 5, May 2006, pp. 658-670. DOI: 10.1287/mnsc.1050.0503
<<http://mansci.journal.informs.org/cgi/content/abstract/52/5/658>>