Building Datasets with TABARI Output: An Analysis of Varying Temporal and Typological Aggregations

ICEWS Working Group
Pennsylvania State University
06/07/2010

Document Overview:

The central aim of this briefing is to illustrate the multiple ways of aggregating TABARI output into useable datasets. In general, Tabari generates four key variables based on the content of politically relevant electronic news stories:

1. **Date**—When the event occurred, presented in YYMMDD format
2. **Source**—The initiator of the action
3. **Target** –The object of the action
4. **CAMEO** code—a two-four digit code that reflects the action committed

This document proceeds by explaining various aggregation options for each of the four variables while briefly addressing relevant data management concerns in STATA. To facilitate explanation, explanation begins with CAMEO and concludes with Date aggregations. Additionally, it provides a case-study of the Israel-Palestine dyad to illustrate how different aggregation strategies generate varying empirical relationships and provide a brief discussion of relationship between variables.

Section 1: Aggregations

CAMEO Aggregations:

CAMEO is a coding typology that assigns a two-four digit numbers to all politically relevant events. In general, two primary methods exist for transforming raw CAMEO codes into usable data: 1) Goldstein-driven scores; 2) Event Counts.

Goldstein-driven scores:

The Goldstein Scale[1] is a conflict-cooperation continuum that provides a numerical score for all CAMEO codes ranging from -10 (most conflictual) to +10 (most cooperative). Three main strategies exist to further aggregate Goldstein scores into useable data.

1. Mean—This reflects the mathematical mean of the Goldstein scores for all events a chosen temporal domain. An "Average" variable, in theory, reflects that average intensity of a conflict during a given time frame. It is important to stress that "Averages" can be misleading. For example, a country with 15 violent attacks and 25 press conferences promoting peace talks during a month could have a Goldstein average of close to 0, because the conflict and cooperation events may largely negate each others' influence. However, a country that experiences no politically relevant events will also generate a Goldstein average of 0. For obvious reasons, treating these two countries equally on a conflict-cooperation continuum is problematic.
2. Sums—This reflects the total Goldstein scores of all events added together across a chosen temporal domain. Like averages, the "Sum" variable can suffer from negative and positive events canceling each other out. However, unlike Averages, if 20 events occur that receive a "5" on the Goldstein scale during a month, the Sum variable will reflect the number of events, generating a score of 100, while the "Average" variable would generate the same score of 5 regardless of the number of events that occurred.
    a. Negative Sums—The total of all "conflictual" events that receive negative Goldstein scores
    b. Positive Sums—The total of all "cooperation" events that receive positive Goldstein scores

[1] Give full citation

3. Total Event Counts—This reflects the cumulative number of recorded events during a specified time frame. The "Total Events" variable is not dependent on Goldstein scores and can also be generated through the "Event Count" variables. However, it is an important variable often used in conjunction with Goldstein scores to proxy for the level of media coverage.

Event Counts:

Based on raw CAMEO codes, it is possible to generate different categories of conceptually unique events and record the number of events that fall into each category during a certain time frame through the use of binary variables. Duvall and Thompson created the most commonly used category-based typology for generating event counts, which is comprised of the following four categories.

1. Verbal Conflict—Events that are spoken through statements and speeches. These may either be fully non-tangible, such as offering condolences of apologies, or may be statements about future actions that have yet to occur, such as threats of attack or aid embargos.
2. Material Conflict—Events that reflect the actual, observable use of attacks and other forms of violence.
3. Verbal Cooperation—Events that reflect either actual dialogue between leaders during meetings and negotiations or the promise of future opportunities for dialogues or the provision of beneficial services like aid or humanitarian support.
4. Material Cooperation—Events that reflect the actual transfer of beneficial resources, such as economic aid, or the implementation of policies including releasing hostages or ending sanctions.

In general, "verbal" events are statements about past/future actions or meeting/negotiations, which "material" events are the actual, tangible occurrence of actions.

Source and Target Aggregations:

Source and Target codes are presented as three, six, or nine-letter abbreviations based on the amount of available information for each actor. The first set of three letters (XXX******) reflect the country of origin. The second set of three letters, when presented, (***XXX***) reflect a sub-country identification, such as GOV for government or REB for rebel. The third set of three letters, when presented (******XXX) provide additional information, such as the branch of government. For example, USAGOVPRE reflects an actor associated with the Presidency branch of the Government of the United States. By selecting different source and target string lengths in STATA, databases can be tailored to reflect either inter-state or intra-state events.

Date Aggregations

Tabari provides six-digit date codes for every event in YYMMDD format. The unit of analysis is directed-dyad events, meaning that more than one event between a directed-dyad can occur on the same day. STATA contains an imbedded calendar that allows for four different temporal aggregations (Daily, Weekly, Monthly, Quarterly), which determine the temporal domain at which CAMEO variables (i.e. Goldstein-driven variables and Event counts) are collapsed. Across all temporal domains, "count" variables (Goldstein Sum, Event Count, verbal conflict, material conflict, verbal cooperation, and material cooperation,) are summed while Goldstein Mean is averaged. Dates between dashes reflect the treatment of an event that occurred on April 19, 2009 by the various aggregations.

1. Daily—19apr1999-Daily is the minimum level of temporal aggregation, which collapses CAMEO variables by directed-dyad day.
2. Weekly—1999w18—Weekly collapses CAMEO variables according to STATA's imbedded Saturday to Friday calendar.
3. Monthly—1999m4—Monthly collapses variables based on standard, Gregorian months
4. Quarterly—1999q2—Quarterly collapses variables according to quarters that run for in standard three-month intervals.

In addition to choosing between various temporal aggregations, one must also choose treatment of non-observations. In order to perform time-series or lagged OLS analyses, the data must be set (tsset) and filled (tsfill) in STATA. For each directed dyad, many days, weeks, months, or even quarters may have no observations. These periods with non-observations can either be treated as "missing" and left omitted from analyses (which will occur if the non-observations are left as ".") or they can be replaced with 0. For analyses utilizing the "count" variables, non-observations should be replaced with 0's to reflect that no event occurred. However, treatment of non-observations among analyses using the Goldstein Mean requires additional theoretical considerations because on the Goldstein Scale, "0" reflects a natural event rather than a non-observation. As such, replacing non-observations with 0's will affect Goldstein Means by increasing the number of observations (i.e. the denominator) while not affecting the numerator. The following section provides an illustration of the empirical effects of treatment of missing data as well as various CAMEO and Date aggregations.

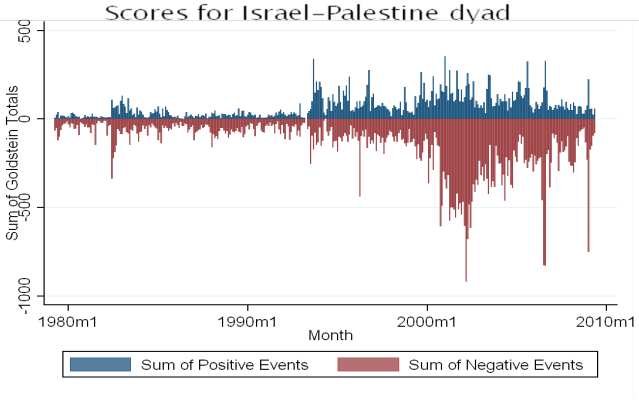
Section 2: Empirical Demonstration

To provide an empirical demonstration of the aforementioned variables and aggregation options, this study analyses events between the Israel-Palestine and Palestine-Israel directed-dyads from the 223,693-event, TABARI-coded Levant dataset from 1979 to 2009. The dependent variable for all regression analysis is Material Conflict.

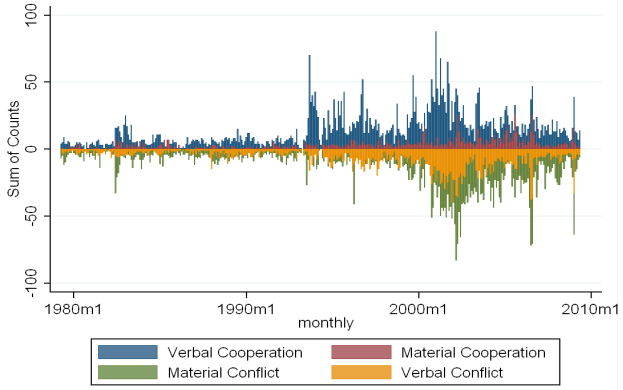
Graphical Analyses:

Although TABARI data has been successfully used in various government and academic capacities, it is helpful to provide graphical illustrations of the dynamics of the Israel-Palestine conflict through the use of monthly Goldstein scores and Event Counts.

Graph 1: Monthly Sum of Negative and Positive Goldstein Scores for Israel–Palestine dyad



Graph 2: Sum of Verbal Cooperation, Verbal Conflict, Material Cooperation, and Material Conflict Counts



The two graphs illustrate the types of events that drive Goldstein Sums. Graph 2 shows that Verbal Cooperation accounts for the majority of positive Goldstein scores, while Material Conflict generates the majority of negative Goldstein Scores. Additionally, the Graphs illustrate the tendency of cooperative and conflictual events to co-vary. This is not surprising given the nature of media coverage and foreign policy. Escalations in attacks lead to increased media coverage, meaning that other types of events are more likely to make electronic press. Additionally, events that comprise Verbal Cooperation (statements, meetings, negotiations, etc.) tend to increase in density around conflictual events.

Empirical Analyses:

This section begins with a discussion of data aggregated at the monthly level, as this is most in both academic and policy oriented analyses. All analyses utilize OLS regression.[2]

Monthly

Table 1: Effects of Count variables on Material Conflict with 1 and 2 month lags (Unfilled missing)

	Model 1 (1 M.lag)	Model 2 (1 M. lag)	Model 3 (2 M. lag)	Model 4 (2 M. lag)
Verbal Conflict	.132 (.086)	-.620*** (.111)	.169* (.096)	-.503*** (.123)
Verbal Cooperation	.070* (.036)	-.662*** (.036)	.075* (.040)	-.584*** (.057)
Material Conflict	.728*** (.036)		.659*** (.040)	
Material Cooperation	.038 (.131)	-.711*** (.150)	.087 (.144)	-.582*** (.166)
Total Event Count		.732*** (.036)		.659*** (.040)
R-Squared	.6672	.6687	.5945	.5939
Constant	1.346	1.346	1.724	1.708
N	694	694	690	690

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

[2] Although daily-level are possible to build, they are rarely used in regression-based analyses. As such, this section does not address daily aggregations.

Table 2: Effects of Count variables on Material Conflict with 1 and 2 month lags (Filled missing)

	Model 1 (1 M. lag)	Model 2 (1 M. lag)	Model 3 (2 M. lag)	Model 4 (2 M. lag)
Verbal Conflict	.129 (.085)	-.623*** (.109)	.170* (.094)	-.503*** (.120)
Verbal Cooperation	.074* (.036)	-.659*** (.050)	.078** (.040)	-.581*** (.056)
Material Conflict	.729*** (.036)		.660*** (.039)	
Material Cooperation	.043 (.128)	-.707*** (.147)	.090 (.142)	-.580*** (.163)
Total Event Count		.733*** (.036)		.660*** (.040)
R-Squared	.6730	.6745	.6017	.6012
Constant	1.242	1.221	1.616	1.601
N	722	722	720	720

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Table 1 and Table 2 both illustrate the relationships between the four count variable and Material Conflict with one and two month lags, though Table 2 replaces “non-observations” with 0s. The empirical results are virtually identical, which is not surprising given that the Israel-Palestine and Palestine-Israel directed-dyads are among the most active in the world. Table 1 and Table 2 also demonstrate that when controlling for the lagged dependent variable, only Verbal Cooperation and Material Conflict—the lagged dependent variable—achieve statistical significance. However, when Material Conflict is replaced by the Total Event Count variable, Verbal Conflict, Verbal Cooperation, and Material Cooperation are all significant and negative and Total Event Count is significant and positive. This suggests that as the number of recorded events increases a T-1 and T-2, we should expect more material conflict at T. However, when these events are anything other than Material Conflict, their marginal effect on Material Conflict decreases. For example, in Model 2, we would expect a one-unit increase in Verbal Conflict at T-1 to only increase the number of Material Conflict events at T by .01 (a one unit increase in Verbal Conflict would cause our expectations of Material Conflict at T to decrease by -.623, but

also increase by .733 as this event would cause a one-unit increase in the Total Event Count variable).

When controlling for the lagged-dependent variable, the effects of Verbal Cooperation in model 1 and Verbal Conflict and Verbal Cooperation in Model 3 become positive. Due to the relatively small size of coefficients, it is likely that this effect is a result of the positive relationship between Total Event Count at T-1 and T-2 and increases in Material Conflict at T.

Table 3: Monthly Correlation Matrix between Count Variables (Filled missing)

	Verbal Cooperation	Material Cooperation	Verbal Conflict	Material Conflict
Verbal Cooperation	1.00			
Material Cooperation	.3878	1.00		
Verbal Conflict	.7444	.4470	1.00	
Material Conflict	.5755	.5923	.7492	1.00

Table 3 provides a correlation matrix between the count variables and suggests that at the monthly level, Verbal Cooperation and Verbal Conflict co-vary with each other, but other variables do not. The relationship between the Verbal variables is expected given that increases in dialogue are likely to lead to increases in both positive and negative rhetoric.

Table 4: Effects of Goldstein variables on Material Conflict with 1 month lags (Filled missing)

	Model 1 (1 M. lag)	Model 2 (1 M. lag)	Model 3 (1 M. lag)	Model 4 (1 M. lag)
Goldstein Sum	-.084*** (.003)			
Goldstein Mean		-.450*** (.036)		-.616*** (.096)
Event Count			.347*** (.011)	.350*** (.010)
R-Squared	.5250	.0103	.5940	.6154
Constant	5.065	8.745	.7179	-.800
N	722	722	722	722

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Table 5: Effects of Goldstein variables on Material Conflict with 1 month lags (Unfilled missing)

	Model 1 (1 M. lag)	Model 2 (1 M. lag)	Model 3 (1 M. lag)	Model 4 (1 M. lag)
Goldstein Sum	-.083*** (.003)			
Goldstein Mean		-.459*** (.159)		-.655*** (.096)
Event Count			.346*** (.011)	.350*** (.010)
R-Squared	.5225	.0105	.5860	.6097
Constant	5.319	9.086	.7710	-.8711
N	694	694	694	694

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Table 4 and Table 5 illustrate the relationship between the Goldstein variables and Material Conflict with and without replacing non-observations with 0s. Like Table 1 and Table 2, the effects of replacement are minimal. Further, as expected, as the Goldstein Sum and Mean increase at T-1, we should expect less Material Conflict at T. However, the extremely small r-squared value of Model 2 indicates that Goldstein Means have very little explanatory power. By accounting for both the Mean and the count, the r-squared score increases to a more robust .61, meaning that the raw number of recorded events explains the majority of variation in Material Conflict.

Weekly[3]

[3] Due to the lack of variation between different treatments of non-observations, this document only presents statistical results derived from data with non-observations replaced with 0s.

Table 6: Effects of Weekly Count variables on Material Conflict with 1 and 2 week lags (Filled missing)

	Model 1 (1 W. lag)	Model 2 (1 W. lag)	Model 3 (2 W. lag)	Model 4 (2 W. lag)
Verbal Conflict	.162*** (.085)	-.428*** (.042)	.023 (.036)	-.491*** (.046)
Verbal Cooperation	.093*** (.036)	-.493*** (.026)	.139*** (.020)	-.373*** (.028)
Material Conflict	.586*** (.036)		.514*** (.018)	
Material Cooperation	.352*** (.51)	-.242*** (.057)	.447*** (.056)	-.070*** (.064)
Event Count		.586*** (.016)		.552*** (.018)
R-Squared	.5171	.5182	.4214	.4118
Constant	.3231	.3210	.4528	.4521
N	3133	3133	3131	3131

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

The results at the weekly level of aggregation are largely consistent for Model 2 and Model 4, which replace the lagged dependent variable with a Total Event Count variable. The Event Count variable is statistically significance and negative, while Verbal Conflict, Material Cooperation, and Verbal Cooperation are significant with negative coefficients. However, unlike in Table 2, all variables (with the exception of Verbal Conflict in Model 3) achieve statistical significance in Model 1 and Model 3 of Table 6. The coefficients are positive, which is in line with the results of Table 2 and fit our theoretical expectations that increases in any type of event at T-1 should lead to increases in Material Conflict at T. The R-Squared values are consistently lower at the weekly level of aggregation, which is expected.

Table 7: Weekly Correlation Matrix between Count Variables (Filled missing)

	Verbal Cooperation	Material Cooperation	Verbal Conflict	Material Conflict
Verbal Cooperation	1.00			
Material Cooperation	.3203	1.00		
Verbal Conflict	.4865	.2925	1.00	
Material Conflict	.4310	.4150	.5622	1.00

Table 7 provides a correlation matrix of the four count variables aggregated at the weekly level.

The correlation scores are smaller than the monthly aggregations, which is expected as trends

are more difficult to recognize and "noise" increases as the temporal domain decreases. This concept is further reflected in the smaller r-squared values in Table 6 relative to Table 2.

Table 8: Effects of weekly Goldstein variables on Material Conflict with 1 week lags (Filled missing)

	Model 1 (1 W. lag)	Model 2 (1 W. lag)	Model 3 (1 W. lag)	Model 4 (1 W. lag)
Goldstein Sum	-.064*** (.002)			
Goldstein Mean		-.126*** (.013)		-.073*** (.001)
Event Count			.333*** (.006)	.328*** (.006)
R-Squared	.3438	.0283	.4586	.4682
Constant	1.104	1.525	.2173	.122
N	3133	3133	3133	3133

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Empirical results presented above in Table 8 largely mirror the equivalent regressions run using monthly aggregations presented in Table 4. Although the Goldstein Sum, Goldstein Mean, and Total Event Count variables achieve statistical significance with the same coefficient signs as in Table 4, the size of the coefficients and r-squared scores are smaller, meaning that weekly aggregations provide less explanatory power than monthly. Again, the Goldstein Mean by itself provides little explanatory power, as reflected with the minute r-squared score of .03. Model 4 suggests that the majority of variation in Material Conflict is explained by the raw number of observed events as opposed to the nature of the events. Nevertheless, Goldstein scores at T-1 still maintain predictive abilities for the level of Material Conflict at T.

Quarterly:

Table 9: Effects of Quarterly Count variables on Material Conflict with 1 and 2 quarter lags (Filled missing)

	Model 1 (1 Q. lag)	Model 2 (1 Q. lag)	Model 3 (2 Q. lag)	Model 4 (2 Q. lag)
Verbal Conflict	-.147 (.158)	-.983*** (.190)	-.213 (.176)	-.853*** (.212)
Verbal Cooperation	.208***	.628***	.445***	-.199*

	(.080)	(.099)	(.088)	(.110)
Material Conflict	.831*** (.058)		.646*** (.064)	
Material Cooperation	-.146 (.211)	-1.013*** (.250)	.133 (.233)	-.527* (.277)
Total Event Count		.834*** (.058)		.643*** (.064)
R-Squared	.7574	.7584	.7084	.7072
Constant	1.346	1.287	.452	.4172
N	240	240	238	238

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Table 9 reports similar results to the previous two tables that analyzed the effects of count variables on Material Conflict. When controlling for the lagged dependent variable with a 1 and 2 quarter lag, only Material Conflict and Verbal Cooperation achieve statistical significance. The coefficients are in the expected direction, with increases in both variables at T-1 and T-2 leading to expected increases in the number of Material Conflict events at T. Model 2 and Model 4, which replace the lagged dependent variable with the Total Event Count variable, generate results that largely mirror previous regressions using monthly and weekly data. One difference is that Verbal Cooperation generates a positive coefficient in Model 2. This runs contrary to previous findings and suggests that even when controlling for the Total Event Count, increases in meetings, negotiations, and communication in general during one quarter should lead to increased Material Conflict in the following quarter. The r-squared scores are consistently larger than both monthly and weekly aggregations, suggesting that quarterly aggregations at T-1 and T-2 provide the most explanatory power regarding the number of Material Conflict events at T.

Table 10: Quarterly Correlation Matrix between Count Variables (Filled missing)

	Verbal Cooperation	Material Cooperation	Verbal Conflict	Material Conflict
Verbal Cooperation	1.00			
Material Cooperation	.5417	1.00		

Verbal Conflict	.8460	.5417	1.00	
Material Conflict	.6694	.7290	.7421	1.00

Table 10 provides a correlation matrix of the count variables at the quarterly level of temporal aggregation. Verbal Conflict and Verbal Cooperation generate the highest correlation score, which differs from previous correlation matrixes. This suggests that the dynamics of opposing leaders' statements may not be fully captured at weekly or even monthly levels. Instead, Table 10 suggests that dialogue tends to correlate across broader time frames, which may reflect the time it takes to generate cohesive policy changes. Overall, the correlation scores are highest at the quarterly level.

Table 11: Effects of Quarterly Goldstein variables on Material Conflict with 1 Quarter lag (Filled missing)

	Model 1 (1 Q. lag)	Model 2 (1 Q. lag)	Model 3 (1 Q. lag)	Model 4 (1 Q. lag)
Goldstein Sum	-.103*** (.005)			
Goldstein Mean		-.136 (.756)		-1.532*** (.415)
Event Count			.386*** (.017)	.394*** (.017)
R-Squared	.6115	-.0041	.6876	.7033
Constant	9.536	22.158	-.275	-4.650
N	240	240	240	240

Significance levels: *** <.01, **<.05, *<.10 Standard errors in () below coefficient

Table 11 presents empirical results similar to findings at the weekly and monthly levels of aggregation for Goldstein scores. Overall, the results suggest that more events at T-1 should lead to a greater number of Material Conflict events at T. Moreover, as the nature of events becomes more cooperative in nature (i.e. Goldstein Sums and Means increase) we should expect to see less Material Conflict. Additionally, Goldstein Means alone provide virtually no explanatory power, though means used in conjunction with Total Event Count generate a large r-squared score of over .7.

Conclusion:

This document has briefly outlined leading strategies that may be employed to transform raw TABARI output into a useable database. Though a large percentage of events data-driven studies utilize data aggregated by techniques presented here, this document is by no means comprehensive. Moreover, as Section 2 illustrates, different aggregation strategies yield varying empirical results. In the case presented—all coded actions between Israel and Palestine from 1979-2009—empirical findings are largely consistent across varying temporal domains and non-observation treatment. However, in directed-dyads with fewer observations, variation is greater between aggregation strategies.