The Quantitative Study of Terrorist Events: Challenges and Opportunities

Jonathan Grossman and Ami Pedahzur

1. Extracting CPOST and GTD Data

We downloaded both databases, the Chicago Project on Security and Threats Suicide Attack Database (CPOST-SAD, henceforth CPOST) and the Global Terrorism Database (GTD-START), in April 2018. While GTD allowed downloading the full dataset, including all the variables, as a comma-separated values (CSV) file, CPOST exported into such a file only the country name, campaign name, attack date, and number of people killed and injured in every event. Other information, including the event's whereabouts and any known details about the target, the perpetrators, and the attack, could only be viewed online. Downloading this data was particularly difficult because CPOST did not allocate a unique URL address to every entry. Eventually, we

were able to extract this information using the proprietary web scraping tool *Import.io*. At the time of writing (May 2019), CPOST's data is no longer available at the project's website and it is unknown when it will become available, if at all.

2. Creating a Baseline

To test the accuracy of data in CPOST and GTD, we created an independent suicide terrorism dataset for Israel and Palestine that served as our baseline.[1] As a first step, we compiled a list of successful, abortive, and foiled suicide attacks in Israel, the West Bank, and the Gaza Strip between the years 1989–2016. To make sure that we documented as many events as possible, we triangulated between three different sources: first, a document entitled "Suicide Terrorism During the Years of the Israeli-Palestinian Conflict (September 2000 – December 2005)," published by the Israeli Intelligence Heritage and Commemoration Center.[2] While this source only covers five years, this period, known as the second *intifada*, was the most eventful one in Israeli history in terms of suicide terrorism. Second, a global suicide terrorism event dataset that one of the authors created at the University of Haifa in the previous decade. Third, we used the Wikipedia page "List of Palestinian suicide attacks."[3] We did not

[1] Another possible method to test these datasets could have been replicating their original coding process while using, for verification, only the sources cited in the databases (Eck 2012, 131. See list of references in the article). However, this method presupposes that English language sources are generally trustworthy and accurate, while our research shows that the sources that event databases cite may often be much less accurate than local sources. In addition, as we show in the article, in many cases, the sources cited by GTD cannot be retrieved, while in other cases no sources are cited at all.
[2] https://www.terrorism-info.org.il/he/18891, accessed 28 May 2019.
[3] https://en.wikipedia.org/wiki/List_of_Palestinian_suicide_attacks. Accessed 10 May 2018.

consider the accuracy of any of these sources (although we did regard the first source as an authoritative one when verifying the details of attacks) but merely used them to create a preliminary list of candidate suicide terrorism incidents.

The second step was to manually verify the details of each and every incident on this list using publicly-available sources, as we describe below. Every event that we were able to verify by at least two independent sources, and which met the inclusion criteria of CPOST, GTD, or both, was included in the baseline. After this meticulous verification process, we were left with 190 verified events whose dates and locations we could establish. In the third step, we applied the same verification protocol to the suicide attack entries in CPOST and GTD.

The verification process went as follows. First, we searched Google for the event's supposed date and location and the Hebrew words "פיגוע" (terrorist attack)," "פיגוע התאבדות" (suicide terrorist attack), or "מחבל" (terrorist); or the English expression "suicide attack." If no relevant results were returned, or if there were too many results, we included further details in the search string, such as the vehicle that was supposedly used for the attack ("car," "truck," "boat," "donkey," etc.) or the supposed names of perpetrators or victims. We only used well-known Israeli and international news organizations for the verification. We did not include blogs, social media posts, NGO websites, forums, etc., but we did follow links to news sources that were referenced by Wikipedia pages.

We also searched the ProQuest Historical Newspapers index for items containing the words “Israel” and “attack” or “Israel” and “bomb” that were published on the event date and the following days. If we could still not find any evidence, we performed a similar search on the LexisNexis online news index. For events occurring since 2000, we also consulted the online archives of the Army and Security section of Israel’s leading news site, Ynet.[4] In addition, we used several Israeli official information sources on suicide terrorism. These included the Israeli Ministry of Foreign Affair’s list “Suicide and Other Bombing Attacks in Israel Since the Declaration of Principles (Sept 1993),”[5] the national commemoration sites for soldiers[6] and civilian terror victims,[7] and the above mentioned document entitled “Suicide Terrorism During the Years of the Israeli-Palestinian Conflict (September 2000 – December 2005)”by the Israeli Intelligence Heritage and Commemoration Center.

The details that we attempted to verify with respect to each event are the following: Did the event really happen? Was it a suicide attack in accordance with the inclusion criteria of CPOST, GTD, or both? On what date did the event occur? At what location did it take place?

[4] https://www.ynet.co.il/home/0,7340,L-4269-141-344,00.html.
[5] http://www.mfa.gov.il/mfa/foreignpolicy/terrorism/palestinian/pages/suicide%20and%20other%20bombing%20attacks%20in%20israel%20since.aspx, accessed 28 May 2019.
[6] http://www.izkor.gov.il/.
[7] http://laad.btl.gov.il.

We explain most of our coding in the codebook below, but our verification process for event locations was somewhat more complicated: when a location coded by CPOST or GTD matched the administrative unit where the event had really taken place (e.g. a city, a village, a local council, etc.), we considered the location "CORRECT." If the event happened near, in the outskirts of, or at a reasonable distance from the location recorded in the database (for example, an attack on an IDF checkpoint near Qalqilya whose location is coded "Qalqilya"[8]), we used the code "AREA," which we also regarded as accurate. GTD also has a dichotomous *vicinity* variable, which, when checked, means that an event took place near an administrative unit rather than in it. If the *vicinity* variable was checked and the location specified by GTD was really near the actual whereabouts of the attack, we considered GTD's coding accurate.

We applied the code "INCORRECT" if the incident took place in a different locality than the one recorded by the database or if the locality closest to the incident was different than the one recorded by the database. For example, both CPOST and GTD coded a 10 April 2002 bus bombing at the Yagur Interchange in Northern Israel as taking place in Haifa, whereas the actual distance between Haifa and the attack location is about 2.5 miles and Kibbutz Yagur is only one mile from this junction. The correct location for this event, then, should have been Yagur (a few other towns and villages are also closer than Haifa to the actual location). Thus, we used the code "INCORRECT" with respect to both databases.[9] In CPOST, which differentiates between Jerusalem's two parts, we

[8] Cpost Event -1335384661; GTD Event 200409140005.
[9] CPOST Events 1279559861; -2081651973 (duplicate); GTD Event 200204110002.

also used the code “JERUSALEM INCORRECT” if an attack in East Jerusalem was attributed to West Jerusalem (the opposite did not happen).

3. Replication Data

Our whole database is available for download at https://github.com/jonathan-grossman/Terrorism-Event-Data. Our raw material – the links used for verification – is available at https://github.com/jonathan-grossman/Terrorism-Event-Data/blob/master/Replication_Material_Verification_Links.pdf. Some of these links have become broken since the time of research. While we do maintain a copy of these sources for replication purposes, we cannot make them available for download for copyright issues.

4. Codebook

Event Title

A short title describing the event.

Date

The correct and verified date of the event.

Location

The correct and verified location of the event.

Nearest City/Junction

The closest locality or junction to the event. If a location recorded by CPOST or GTD matched this variable, but coded the relevant value as “AREA” and considered it accurate.

Event Happened?

The event described in the entry really happened, regardless of its nature (that is, even if the incident was not a suicide terrorism event). In our verification, we considered an event true if we could find evidence for it in at least two independent sources (that is, sources that did not cite one another or quoted a common third source, such as two news articles citing the same Reuters newswire). If we found such conclusive evidence that the event happened, we coded the *Event happened?* variable as “YES.” If we did

not find any authoritative source that verified the event, we coded it as “NO.” If we found some evidence as to its existence but less than two independent sources, we coded it “UNCLEAR.”

“YES” = we were able to verify the event with evidence from at least two independent sources

“NO” = we were able to rule out the event

“UNCLEAR” = we were unable to either verify the event or rule it out

“YES, OUTSIDE ISRAEL” = we were able to confirm the event but it did not take place in Israel or Palestine

Suicide Attack?

The event was a suicide attack (there is evidence that the perpetrators killed themselves or intended to do so).

“YES” = we were able to verify that the event was a suicide attack

“PROBABLY” = there is some evidence that suggests that the event was a suicide attack, but we were not able to determine it beyond doubt

“NO” = we were able to confirm that the event was not a suicide attack

“UNCLEAR” = we were unable to determine whether the event was a suicide attack or not

“ATTACK AVERTED” = a foiled attack. Can still count as suicide attack according to GTD’s criteria (but not according to CPOST’s)

“DUPLICATE” = event is a duplicate

In Baseline?

The event is included in our baseline. In other words, we were able to verify the event’s existence, date, location, and (with a high degree of certainty) satisfaction of the inclusion criteria of CPOST, GTD, or both.

“YES” = the event features in the baseline

“NO” = the event does not feature in the baseline

In CPOST?

The event is included in the Suicide Attack Database of the Chicago Project on Security and Threats.

“YES” = the event features in CPOST’s database of suicide attacks in Israel and Palestine

“NO” = the event is not recorded by CPOST

“DUPLICATE” = this is a duplicate of an event already coded by CPOST

“YES, OUTSIDE ISRAEL” = the event is included in CPOST but coded for a different location than Israel/Palestine (we used this code for events in South Lebanon that GTD incorrectly coded as taking place in Israel)

CPOST ID

The event’s identification number in CPOST.

CPOST Criteria

Does the event meet CPOST’s criteria for suicide attack? If we were not able to find conclusive evidence, we coded the *CPOST Criteria* variable as “UNCLEAR,” allowing CPOST the benefit of the doubt. If most sources suggested that the event met the database’s criteria yet some sources pointed to the contrary, we coded the variable as “PROBABLY.” If the event was obviously not a suicide attack, we coded it as “NO.”

“YES” = we were able to confirm that the event met CPOST’s inclusion criteria

“NO” = we were able to determine that the event did not meet CPOST’s inclusion criteria

“PROBABLY” = the event probably meets CPOST’s criteria. However, we either did not find two independent sources that unequivocally confirmed this or we found some contradicting evidence. Despite this, we allowed CPOST the benefit of the doubt and considered the event to qualify as a suicide attack

“UNCLEAR” = based on the sources we found, we could neither verify nor rule out the event’s satisfaction of CPOST’s inclusion criteria

“DUPLICATE” = this entry is a duplicate of an event already coded by CPOST

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occured

CPOST Criteria – Why?

A brief note explaining our coding of the *CPOST Criteria* variable (when needed).

CPOST Date

The date of the event according to CPOST.

Cpost Date - Discrepancy

Is the date coded by CPOST the actual date on which the event happened?

“CORRECT” = CPOST’s date is the date on which the event happened

“INCORRECT” = CPOST coded the wrong date for this event

“DUPLICATE” = CPOST coded the correct date but this record is a duplicate

“INCORRECT+DUPLICATE” = CPOST coded the wrong date. In addition, this is a duplicate

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occurred

CPOST - Location

The location of the event according to CPOST

CPOST Location - Discrepancy

Is the location coded by CPOST the actual whereabouts at which the event took place?

“CORRECT” = CPOST’s location is the location at which the event happened

“INCORRECT” = CPOST coded the wrong location for this event

“AREA” = the event occurred in the vicinity of the location coded by CPOST

“CORRECT+DUPLICATE” = the location is coded correctly but this is a duplicate event

“AREA+DUPLICATE” = the event occurred in the vicinity of the location coded by CPOST but this is a duplicate event

“INCORRECT+DUPLICATE” = CPOST coded the wrong location for this event, which is also a duplicate

“JERUSALEM INCORRECT” = the event did happen in Jerusalem, as coded by CPOST. However, it happened in East Jerusalem while CPOST coded its location as “Jerusalem (West)”

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occurred

In GTD?

The event is recorded in the Global Terrorism Database.

“YES” = the event features in GTD and is categorized as a suicide attack

“NO” = the event is not recorded by GTD

“DUPLICATE” = this entry is a duplicate of an event already recorded by GTD

“NOT SUICIDE” = the event features in GTD but is not coded as a suicide attack (this is the case with some events in CPOST and our baseline)

GTD ID

The event’s identification number in GTD.

GTD Link

A hyperlink to the event’s entry on the GTD website.

GTD Criteria

Does the event meet GTD’s criteria for suicide attack? When we were not able to find conclusive evidence, we coded the *GTD Criteria* variable as “UNCLEAR,” allowing GTD the benefit of the doubt. If most sources suggested that the event met the database’s criteria yet some sources pointed to the contrary, we coded the variable as “PROBABLY.” If the event was obviously not a suicide attack, we coded it as “NO.”

“YES” = we were able to confirm that the event met GTD’s inclusion criteria

“NO” = we were able to determine that the event did not meet GTD’s inclusion criteria

“PROBABLY” = the event probably meets GTD’s criteria. However, we either did not find two independent sources that unequivocally confirmed this or we found some contradicting evidence. Despite this, we allowed GTD the benefit of the doubt and considered the event to qualify as a suicide attack

“UNCLEAR” = based on the sources we found, we could neither verify nor rule out the event’s satisfaction of GTD’s inclusion criteria

“DUPLICATE” = this entry is a duplicate of an event already coded by GTD

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occured

GTD Criteria – Why?

A brief note explaining our coding of the *GTD Criteria* variable (when needed).

GTD Date

The date of the event according to GTD.

GTD Date - Discrepancy

Is the date coded by GTD the actual date on which the event happened?

“CORRECT” = GTD’s date is the date on which the event happened

“INCORRECT” = GTD coded the wrong date for this event

“DUPLICATE” = GTD coded the correct date but this record is a duplicate

“INCORRECT+DUPLICATE” = GTD coded the wrong date. In addition, this is a duplicate

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occurred

GTD - Location

The location of the event according to GTD

GTD Location - Vicinity

The *vicinity* variable in GTD is coded as positive. If this is the case, the event supposedly happened near the location coded under the *GTD - Location* column rather than in that location.

GTD Location - Discrepancy

Is the location coded by GTD the actual whereabouts at which the event took place?

“CORRECT” = GTD’s location is the location at which the event happened (or near this location, if the *GTD Location - Vicinity* variable is positive

“INCORRECT” = GTD coded the wrong location for this event

“AREA” = the event occurred in the vicinity of the location coded by GTD

“CORRECT+DUPLICATE” = the location is coded correctly but this is a duplicate event

“INCORRECT+DUPLICATE” = GTD coded the wrong location for this event, which is also a duplicate

“EVENT DID NOT HAPPEN” = we could not verify that this event ever occurred