**Dataset Description**

The accompanying dataset has ~717,000 records representing ~230,000 unique conjunction events. A CA screening run is accomplished once per day, looking 7 days into the future; and any conjunctions that have a miss distance of less than 1 km generate a record for this database (data run from SEP 2021 to MAY 2022). If a conjunction meets those criteria for multiple days, then multiple records are generated, even though it is the same conjunction event. A single conjunction event has the same primary and secondary objects (and therefore the same primary and secondary object numbers) and the same Time of Closest Approach (TCA). Well, almost the same TCA; with orbital updates for the primary and secondary, the TCA will change slightly, although usually not more than a second or so; we usually define any events with the same primary, secondary, and TCA within plus or minus 15 minutes as the same event.

Two data files are provided: one is a binary, .mat file ready for use with MATLAB, and the other is this same information as a .csv file. There are also two conference papers provided: one treats the issue of CA event dynamics and the “dilution region” effect, which should be understood at at least a high level in order to work with CA datasets; and the second is a treatment of the “consequence” of collisions in terms of estimating the amount of debris produced. I would read the former carefully and skim the latter.

Of the 49 fields (columns) in the provided dataset, only a relatively small subset are relevant to your work:

* **48: Event Number**. Identifies all the records associated with a single event. Useful for time-series analysis or for determining the distribution of Pc values at a certain time to TCA.
* **49: Days to TCA**. The time to TCA, which is really just the difference of fields 39 and 38. Useful for the applications given in the discussion of field #48 above.
* **9: Pc Best**. The best estimate of the probability of collision, using a complicated approach to estimate both objects’ sizes from sensor radar cross-section data. I can send a whole briefing on this technique if there is interest.
* **6: Pc Nom**. If the size estimation approach described above in the treatment of field #9 is not possible, usually because RCS data are not available, a NAN will be present in field #9; in such cases, the Pc value here is field #6 should be used (this is the Pc based on a nominal value of 1.5m for both primary and secondary, 3m total.
* **12: ProbCatIfColl**. This field and the next couple to be discussed relate to the amount of debris that would be produced if a conjunction were to result in a collision. The NASA EVOLVE 4.0 breakup model produces these estimates by making a basic distinction between collision types: catastrophic collisions, in which both objects are completely fragmented; and non-catastrophic collisions, in which the smaller (lighter) object is completely fragmented but the larger object merely cratered. Field #12 gives the probability that the collision will be catastrophic. With debris-on-debris events, most collisions would be catastrophic. If the probability of a catastrophic collision is above 0.5, then I would probably assume that it will be catastrophic.
* **14: NumFragIfCatColl**. This field gives the number of fragments 5cm or greater in size that would be expected to be produced in this particular collision were catastrophic. You can use this field in conjunction with field #12 to determine the number of fragments expected: if #12 is greater than 0.5, then #14 will give the number of fragments expected.
* **15: NumFragIfNonCatColl**. If the collision is expected to be non-catastrophic, then this field gives the number of debris fragments expected.

What are some ways forward with this analysis? There are a number of potential paths; the below are things I might try:

* First, rework the dataset slightly so that you have a single column that gives a Pc for every record (i.e., invokes the Pc Nom value when there is no Pc Best)
* Next, divide the entries up into bins by time to TCA. Maybe one-day bins (0.5 to 1.5 days to TCA, 1.5 to 2.5 days to TCA, &c.). Make sure that each event contributes only one entry to a bin. Plot the PDF/CDF of the Pc values for each bin and get a sense of the situation. Maybe also determine the number of events with Pcs > 1E-04 and > 1E-05. These values would be typical contenders for warning. See how many warnings would be issued at each time to TCA.
* Now introduce the number of fragments as a second consideration. Plot a PDF/CDF of the number of fragments for each event (making the catastrophic / non-catastrophic collision decision for each event as described earlier) and get a sense of the situation. Almost anyone would say that 10 or fewer fragments is irrelevant, and many operators would extend that same sentiment up to 50 fragments. See what thresholds at different levels would mean.
* After this, one needs to turn attention to the time-series aspect. One cannot establish threshold norms against actual collision data, since there really isn’t any; so the usual proxy is the Pc at 0.5 to 1 days to TCA. If the General wanted a five-day warning time, what threshold would have to be applied at five days to TCA to warn for every conjunction that will have a Pc > 1E-04 or 1E-05 at 1 day to TCA? For any given threshold, what are the false alarm and missed detection rates? You can read the Wikipedia article on “receiver operator characteristic” curves to learn a little more about this, but plotting efficient frontiers with missed detections on one axis and false alarms on the other is a good way to display the overall trade-space. You can do this with multiple parameters in play (both Pc and debris fragments).

Well, that’s about enough for now. Feel free to e-mail me with questions at matthew.d.hejduk@aero.org.