***Summary of EDDMapS zero abundance/implied absence data***

***Author: Mitchell O’Neill***

***Date: August 7, 2017***

The following summarizes exploratory data analysis of records in the EDDMapS database that are not specified as absence data, but contain entries that imply absence in abundance columns (i.e. “0 plants”). This EDA uses a data download of all invasive plant point data (1,664,224 points). These data were requested by Bethany Bradley and delivered by Rebekah Wallace on 18 Oct, 2016.

***1. Absence Data***

Reporters have the option to specify a record as a presence or absence by selecting “positive” (presence) or “negative” (absence), with positive as the default (<https://bugwoodcloud.org/CDN/eddmaps/tools/Report.Neg.Data.pdf>). With positive as the default, a record is assumed to be a presence unless specified otherwise. As presence data is much more likely than absence data to be entered into a database like this, this assumption is likely safe one for most observations. However, because a record is specified as a presence without necessarily requiring any action from the reporter, it is reasonable that some absence records could be specified as presences.

* In the bulk data download (see top of page), this information is given in a field called “negative”. The default value is zero (presence), while a value of one indicates an absence.
* 41,192 (2.5%) of the occurrence records (negative field = 0) have an entry in an abundance field that implies absence (Table 1)

**Table 1:** Lists the entries in each field that could imply an absence.

|  |  |
| --- | --- |
| Abundance Field | Absence Indicator |
| Abundance Text | "None " |
| Infested Area | 0 |
| Gross Area | 0 |
| Density | "None", "0 PLANTS", "Trace, NONE FOUND", "0%", "0" |
| Invasive cover (from percent cover and canopy cover columns) | "00%", "none", "0", "0% ", "0 " |
| Percent Plot Covered | "0", "0,0", "00%", " 0%" |
| Stem Count | "0", "0.000000" |
| Number Observed | “0” |
| Texas Abundance | “None", "0" |

* 92.6% of these have a contradiction in a different abundance field (i.e. 0 in stem count, but 0.5 in infestedAreaInAcres). Could this mean that NAs were coerced to zeroes, or that zeroes can be default values?

**Quality assurance 1:**Exclude all implied absences with contradictions.

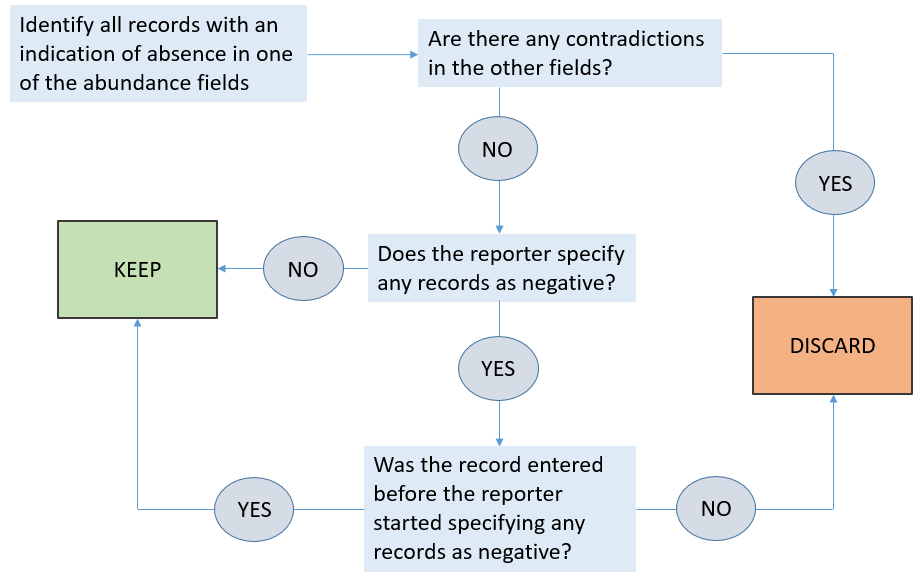
* 3,032 records pass this filter.

***2. Looking at the reporters of these implied absences***

If any of these implied absences are truly absences, it would be because the reporter neglected to specify negative versus positive. If this is the case for any reporter, we might expect that the reporter fails to specify negative vs. positive in any entries. In our dataset, this would show up as a reporter whose observations all have a 0 in the negative field. However, this could also indicate a reporter who only reports presences. As 0 is the default value, we cannot distinguish between the two possibilities. Nonetheless, if a particular reporter is shown to specify some records as absences, then it is unlikely that the reporter would also enter an absence without specifying it as an absence. Essentially, we should exclude all implied absences entered by reporters who have any records where a 1 is entered in the negative field.

**Quality assurance 2:**Exclude all implied absences entered by reporters who have any records where a one is entered in the negative field.

* 3,022 (99.7%) records pass this filter.
* It is also conceivable that a reporter fails to specify true absences as negative until they become aware of the option to so. As such, we could consider any implied absences reported before the reporter specified a record as negative for the first time. This logic could be used to filter through the records that did not pass quality assurance 2 (implied absences that have no contradictions but are entered by reporters who have specified records as negative). However, none of these records passed through the proposed filter.



**Figure 1**: Flow chart of filtering implied absence data through quality assurance 1 and 2.

***3. Are the remaining implied absences true absences?***

* The main concern with the implied absence records that pass through the quality assurance 1 and 2 filters is that zero-abundance-implying entries are really the default values used by the reporter.
* If these zero-abundance-implying entries are indeed default values for a reporter, then we would expect that “NULL” does not appear in the particular abundance field in any of the records entered by that reporter.
* Some data exploration of these reporters has shown that for most of these reporters (>95% of the 407 reporters), “NULL” is never entered or rarely entered for at least one of the abundance columns.
* Some records have entries in the Comments field that imply presence.
* This could mean that these implied absences are not true absences and are merely due to misleading default values.

***4. Potential quality assurance measures:***

* There are 3,022 implied absences entered by 407 unique reporters that pass through the previously described quality assurance filters, but there is still reasonable doubt that these are true absences.
* It could be that for each reporter certain abundance fields contain true absence information while other abundance fields simply have zero abundance as the default.
* It could also be that some zero-implying entries indicate true absences (i.e. “trace, NONE FOUND”) while others are simply default values or equivalent NULL (i.e. “0” or “none”). [Note: there are also many different entries that imply NULL, i.e. “N/A” or “unknown” or “Select Abundance”]

This would require classifying implied absences as true or false absences based on:

1. Who is reporting the observation?
   * An implication of absence may be interpreted as a potential absence for some reporters, but as an error for other reporters.
2. Which field implies abundance?
   * For each reporter, there may be some fields that are trusted to imply absence and some fields which are ignored.
3. What is the entry that implies abundance?
   * Within each column for each reporter, there may be entries that are interpreted as the default, as well as entries that are interpreted as true absence data.

\*Alternatively, implied absences may be interpreted as generally unreliable or too difficult to interpret, and values that imply absence (i.e. zeroes) could be coerced to NULLs. However, the value of absence data and the errors that could arise from absences being specified as presences should be kept in mind.