***Summary of results yielded from exploration of the invasivecover column in the EDDmapS data***

***Author: Ceara Sweetser***

***Date: July 26, 2017***

***Cleanup Protocol***

Invasivecover was a column created by Mitch O’Neill to account for quirks in the EDDmapS provided Canopycover and Percentcover columns. The issue here was that if Canopycover is not NULL and the Percentcover is NULL then Canopycover is the area covered by the invasive species. If Percentcover is not NULL then Percentcover is the area covered by the invasive species and Canopycover (NULL or not NULL) would be the native vegetation’s cover. Three columns were created to break up the invasivecover column. The first was invasiveNUMERIC. The entries in this column were filled in by calculating as.numeric(invasivecover) and if this operation resulted in an NA then the entry in invasivecover was transferred as is to a new column called invasivecoverCHAR.

The last column created is called InvasivecoverCHAR.CLEAN. It has 6 acceptable entries: trace, low, moderate, high, very high, and NULL. The numerical ranges for these bins were set up by Professor Bethany Bradley in her exploration of the Quantitative cover column.

|  |  |
| --- | --- |
| Quantitative Cover | Qualitative Cover |
| <1% | Trace |
| 1-5% | Low |
| 5-25% | Moderate |
| 25-50% | High |
| >50% | Very High |

InvasivecoverCHAR.CLEAN was calculated by first going through the invasivecoverCHAR column to properly bin the values and statements into the 6 categories (trace, low, moderate, high, very high, NULL). If the value or range of values fell explicitly into one of those bins it was given the correct label. Values that matched the bin boundaries were placed into certain bins as described below:

1: low 5: low 25: moderate 50: high

For ranges that spanned more than one bin, that entry was labeled with whichever bin covered most of the range. For example for a value of 20-75% that range mostly falls into “high” so invasivecoverCHAR.CLEAN would take on the value “high”. For the rows with a written out range such as “trace-low” invasivecoverCHAR.CLEAN took on the lower bin label. In this case, for example, “trace”.

There were a few values that came from a contributor’s ranking system (1-5). Those values and how we designated them are listed below.

“1=<5%,n” : low “2=5-25”: moderate “3=25-5”: high “4=50-7”: very high

“5=75-1”: very high

Lastly, to fill out the rest of the invasivecoverCHAR.CLEAN column the values from the invasivecoverNUMERIC column were evaluated and binned.

The total number of entries in the invasivecoverCHAR.CLEAN column that were not NULL ended up being 442,025.

***Exploring Questionable Values***

First we wanted to look at the abnormal, ranked values (ex. 3=25-5) in invasivecoverCHAR. There was only one person who contributed values like these (David Dick) but he also provided acceptable values in the numeric and char columns that were not of this format.

Because all entries in invasivecover should be percentages we wanted to look at the values that were over 100. There were only 10 of these values provided by 3 contributors: The Appalachian National Scenic Trail, Alaska Exotic Plant Information Clearinghouse database, and John Odell. These were likely entry mistakes.

Next we wanted to look at values less than 1. We were curious if these were actually trace values or whether contributors were accidentally providing proportions versus percentages. There were 40,975 values (~9% of non-NULL values) under 1 provided by 44 contributors (including the three already mentioned above). We were interested in seeing if any reporters provided exclusively numbers under 1 because this would probably indicate that they recorded proportions. This search led us to 10 reporters: Ryan Wersal, Robin Bond, Matt Gower, Michael S Howell, Michael C Cox, Jimmy D. Peeples, Daniel Z. Reynolds, Evan O’Donnell, Lesley C. Coats, and Kristin Follmer. They did not contribute values to the character column of any kind or values in the numeric column that were greater than 1. We looked at the unique values they contributed as a group and they do indeed appear to be proportions.

1.00 0.45 0.80 0.10 0.90 0.40 0.50 0.25 0.75 0.30 0.65 0.02 0.20 0.15 0.55 0.60 0.70 0.95 0.85

We will multiply these values by 100 and reassign the bins accordingly.

Lastly, we wanted to see if there were any reporters that exclusively contributed numbers that were 1, 2, 3 ,4, or 5, thinking that this might be evidence of a ranking scheme such as the one employed by David Dick. There are 115,460 values between 1 and 5 which is a significant amount of the total non-NULL values (~26%). A brief exploration of a few of the reporters that contributed values 1, 2, 3, 4, or 5 showed that they also provided acceptable values greater than 5. For now, we can mostly likely conclude that the integers from 1 and 5 inclusive in the numeric column are not part of a bin/ranking scheme but rather actual percentages.