***Summary of EDDMapS infestation extent data exploration***

***Author: Mitchell O’Neill***

***Date: August 4, 2017***

The following summarizes exploratory data analysis of infested area and gross area data in the EDDMapS database. 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. “Infested area” and “gross area” hereafter refer to the “infestedAreaInAcres” and “grossAreaInAcres” fields, respectively. In the EDDMapS Invasive Plant Mapping handbook and database definitions, these fields are shown to be closely related, so I explore these two fields together.

***1. First pass through infested and gross area data.***

Out of the 82,836 records (49.8% of entire database) that have infested area and/or gross area data:

* 106,546 (12.9%) records with gross area, but not infested area
* 484,505 (58.4%) records with infested area, but not gross area
* 227,311 (28.6%) records with both infested and gross area

INFESTED AREA (721,816 records; 43.4% of database)

* 22 negatives, 6,191 zeroes (0.9%): equivalent to NULL?

GROSS AREA (343,857 records 21.0% of database)

* 26 negatives, 29,594 zeroes (8.6%): equivalent to NULL?

**Quality assurance 1**: exclude records where infested or gross area are equal to or less than zero.

* 715,603 Infested area records pass this filter
* 314,237 Gross area records pass this filter

**2. *Looking deeper into positive values of infested and gross area data.***

Table 1: Percentiles of positive infested and gross area data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentile | 0% | 25% | 50% | 75% | 100% |
| Infested Area | 2.31711e-09 | 1.00000e-02 | 8.00000e-02 | 4.00000e-01 | 2.33120e+06 |
| Gross Area | 7.723687e-08 | 5.00000e-02 | 1.50000e-01 | 1.00000e+00 | 4.62500e+05 |

* Huge range in values. Before worrying about these, explore how infested and gross area compare to their definitions.
* Infested area should not be greater than gross area according to data definitions (<https://www.eddmaps.org/tools/datadictionary.pdf>)
* For records with both infested and gross area, how do these values compare?

207,767 records with positive value for both gross area and infested area

* 134,241 (64.6%) records with gross > infested
* 65,325 (31.4%) records with gross = infested
* 8,201 (3.9%) records with gross < infested (VIOLATE DEFINITION)

**Quality assurance 2**: Use only records where infested area is less than or equal to gross area. These are the only records where we can confirm that the metadata definitions of infested and gross area are satisfied.

* 199,566 points pass this filter

***3. Exploring the large range in vales in infested and gross area data***

Table 2: Percentiles of infested and gross area data after quality assurance 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentile | 0% | 25% | 50% | 75% | 100% |
| Infested Area | 2.31711e-09 | 6.182803e-03 | 5.00000e-02 | 2.00000e-01 | 1.80000e+05 |
| Gross Area | 7.72369e-08 | 5.00000e-02 | 1.00000e-01 | 1.00000e+00 | 4.62500e+05 |

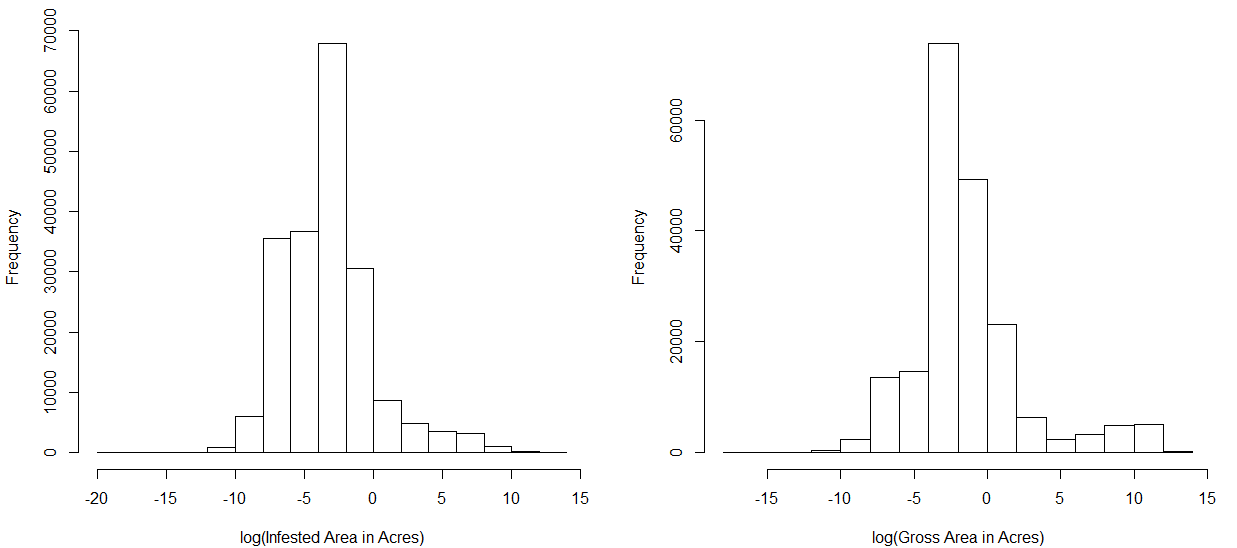


Figure 1: Histograms of the natural logarithm of infested and gross area after quality assurance 1 and 2. Note the positive skew.

We have been interpreting infested area and gross area as visual estimates that reporters are making on the ground. It is unlikely that a reporter can assess an area greater than about one acre in this way. Areas up to about 20 acres may be calculated from surveys or aggregated from finer-scale estimates. However, 18869 (9.5%) records exceed 20 acres. Our best guess is that these are from aerial surveys or calculated from classified images, methods that are very different from what is described in the EDDMapS Training Handbook (<https://www.eddmaps.org/training/EDDmapS.pdf>). Another possible explanation is the entry of incorrect units (i.e. 1000 square feet entered as 1000 acres). Most of these observations have “Acres” listed in the infested area units field. Regardless, the extremely large values are likely incompatible with the rest of the data, and as such data cleaning is necessary.

* 96 % of the records where gross area > 100 acres are contributed by one reporter:
* First name: US Army Corps of Engineers, Last name: Ombil Database
* US Army Corps consistently reports large values, with a minimum value much larger than most of observations made by other reporters. The regularity of the infested area values is unusual. Perhaps the reporters are estimating coverage within large natural areas? Nonetheless, the data given by this reporter cannot be considered the same kind of data given by other reporters.

Table 3: Percentiles of infested and gross area data for US Army Corps observations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentile | 0% | 25% | 50% | 75% | 100% |
| Infested Area | 1 | 10 | 100 | 650 | 180000 |
| Gross Area | 1 | 2795 | 13602 | 33643 | 462500 |

**Quality assurance 3**: Remove the records reported by the US Army Corps of Engineers, because the values of infested area and gross area seem consistently and unreasonably high.

* 185, 843 records pass this filter.

***4. Exploring the high values remaining after the exclusion of US Army Corps***

63% (81 + 61 /236) of remaining high values are from two reporters: United States Forest Service Intermountain Region and Mark Twain National Forest

* These reporters do appear to report higher values than most of the other observations. However, values less than one are reported and most observations appear to be reasonable ( <20). Therefore, I’m not sure if I am justified in removing these reporters. I do not know how to address the remaining high values other than removing outliers.

Table 4: Percentiles of infested and gross area data for USFS Intermountain Region and Mark Twain National Forest

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentile | 0% | 25% | 50% | 75% | 100% |
| Infested Area (USFS-IMR) | 0.0002 | 0.0043 | 0.0233 | 0.0988 | 4518.2736 |
| Gross Area (USFS-IMR) | 0.0002 | 0.0169 | 0.0913 | 0.2104 | 8405.8113 |
| Infested Area (MTNF) | 0.0016 | 0.3604 | 1.8927 | 11.4239 | 1188.0371 |
| Gross Area (MTNF) | 0.0016 | 2.5968 | 8.9682 | 36.5158 | 3207.6078 |

***5. Double-checking thrown out data***

With quality assurance 2, I decided to only include records where gross area is greater than or equal to infested area. This ends up excluding the ~500,000 records where infested area is given but gross area is not. Since I am interested in using the infested area field, throwing out data should be avoided when possible. While I don’t necessarily know that these records have unreliable infested area data, I can’t confirm that this field was not confused with gross area data like I can with records that include both gross and infested area (quality assurance 2). There 587,407 records with unconfirmed infested area data— where infested area is positive and the reporter is not US Army Corps, but gross area is NULL or non-positive). Here I compare the distributions of the unconfirmed infested area with the confirmed infested area and confirmed gross area (records that pass quality assurance 1-3).

Table 6: Percentiles of confirmed infested and gross area data (pass all quality assurance filters) as well as unconfirmed infested area data (where infested area is positive and the reporter is not US Army Corps, but gross area is NULL or non-positive).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentile | 0% | 25% | 50% | 75% | 100% |
| Confirmed Infested Area | 2.317106e-09 | 4.591360e-03 | 5.000000e-02 | 1.000000e-01 | 6.400000e+03 |
| Confirmed Gross Area | 7.723687e-08 | 5.000000e-02 | 1.000000e-01 | 5.000000e-01 | 3.000000e+05 |
| Unconfirmed Infested Area | 2.295680e-07 | 1.499079e-02 | 1.000000e-01 | 4.000000e-01 | 2.331204e+06 |

It appears that the unconfirmed infested area data have generally higher values than the confirmed infested area, with a distribution somewhat similar to the confirmed gross area. It seems that when reporters enter infested area only, they may be entering gross area data in the infested area field. Since I have no way of teasing out the true infested area from the gross area in this column, I will only use data that pass quality assurance 1-3.

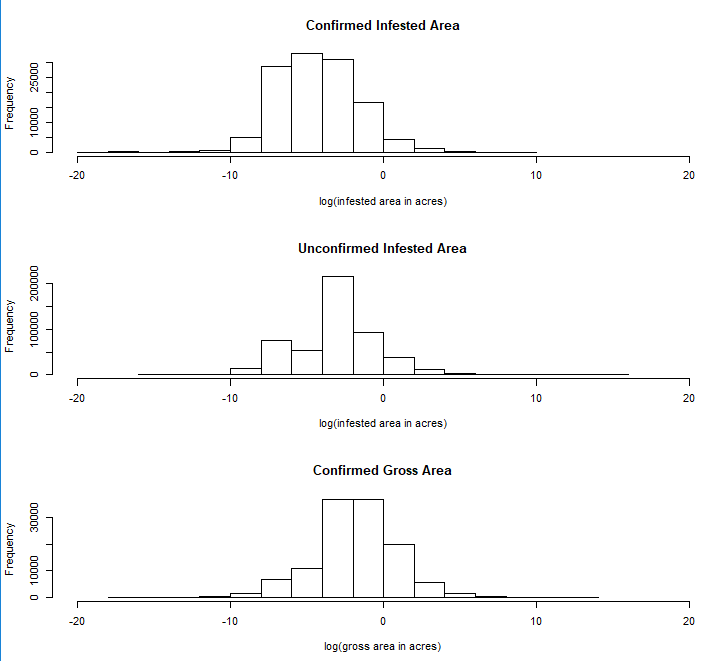


Figure 4: Comparison of distributions of confirmed infested and gross area (top and bottom; passed all quality assurance filters) and unconfirmed infested area (middle; where infested area is positive and the reporter is not US Army Corps, but gross area is NULL or non-positive).

*Review of quality assurance filters:*

* **Quality assurance 1**: exclude records where infested or gross area are equal to or less than zero.
* **Quality assurance 2**: Use only records where infested area is less than or equal to gross area. These are the only records where we can confirm that the metadata definitions of infested and gross area are satisfied.
* **Quality assurance 3**: Remove the records reported by the US Army Corps of Engineers, because the values of infested area and gross area seem consistently and unreasonably high.

***Conclusion***

There is still an issue with the large range of values in the infested and gross area columns. After all of my quality assurance measures: there are 236 infested area values that exceed 100 acres, reaching up to 6,400 acres, while 89.2% of the infested area values are 1 acre or less. For percentiles see Table 6 (previous page; confirmed infested area). I do not know how to address the remaining high values other than removing outliers.

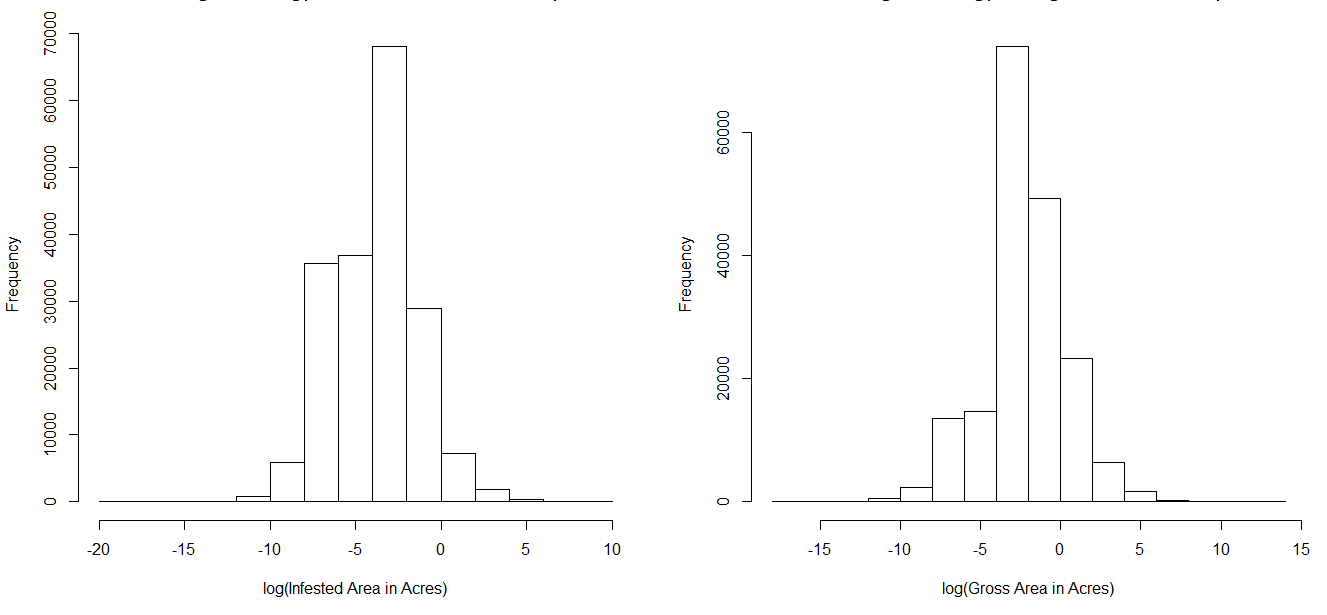


Figure 3: Histogram of infested and gross area data after quality assurance 1-3. Note that the positive skew is largely resolved, but unreasonably high numbers remain (see Table 6).