Distribution fitting for Database Manuscript 1

Maureen Kennedy

March 20, 2018

## Distribution-fitting Methods

After an initial visual inspection of empirical distributions for each fuel loading type we decided on 4 candidate distributions: normal, lognormal, gamma, and Weibull (although, with hurdle modeling we migth stick to just lognormal and gamma--normal is almost never chosem, weibull is very similar to the other two, and this avoids issues with truncation for the normal and Weibull). The lognormal, gamma, and Weibull distributions were chosen because many of the fuel loading types for many of the EVT groups exhibited right-skewedness. Each of these distributions have two parameters that can be estimated using maximum likelihood (Table). The normal and weibull distributions can accommodate zeroes, whereas the lognormal and gamma distributions have domain > 0.

Distributions (see table)

normal pdf:

E(x) =

V(x) =

lognormal pdf:

E(x) =

V(x) =

gamma pdf:

E(X) =

V(X) =

Weibull pdf:

E(X) =

V(X) =

where:

for x>0

We chose a minimum number of entries allowed for distribution fitting (n ), and fit each of the candidate distributions to all fuel type and EVT group combinations that achieved the minimum number of entries (using fitdistrplus package in R). The distribution achieving the maximum liklihood criterion for each EVT group and fuel type was chosen as the final distribution.

citation("fitdistrplus")

##   
## To cite fitdistrplus in publications use:  
##   
## Marie Laure Delignette-Muller, Christophe Dutang (2015).  
## fitdistrplus: An R Package for Fitting Distributions. Journal of  
## Statistical Software, 64(4), 1-34. URL  
## http://www.jstatsoft.org/v64/i04/.  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {{fitdistrplus}: An {R} Package for Fitting Distributions},  
## author = {Marie Laure Delignette-Muller and Christophe Dutang},  
## journal = {Journal of Statistical Software},  
## year = {2015},  
## volume = {64},  
## number = {4},  
## pages = {1--34},  
## url = {http://www.jstatsoft.org/v64/i04/},  
## }  
##   
## Please cite both the package and R when using them for data  
## analysis. See also 'citation()' for citing R.

# Zeroes

For many of the EVT groups there were entries of zero loading for many fuel types. We evaluated two options for all distributions:

1. estimate the distribution including all values, with a value of 0.1 added for all fuel types so the lognormal and gamma distributions could be estimated
2. estimate a hurdle-type model (ref), where the distribution function (f(x)) is represented as:

if x=0

h(x) if x>0

0 otherwise,

h(x) is one of the candidate distributions (normal, lognormal, gamma, Weibull; ***note: normal is pretty much never chosen, and Weibull isn't necessarily appropriate for hurdle fitting, so might just stick with gamma and lognormal--seems pretty common with biomass modeling***), and is the estimated probability of zero loading for that fuel type in a given EVT group. Hurdle models are common tools for zero-inflated data, where the zero values are represented in greater proportion than can be modeled using the density function alone (usually with discrete count models, but have also been applied to continuous distributions with many zeroes). It has been shown (ref) that the two components of the distribution function are indepdendent , with the consequence that the proportion of zeroes can be estimated simply as the proportion of entries that have a value of zero, and then the distribution can be estimated for the remaining non-zero entries. All distributions were estimated with both approaches, and the fit of each approach compared (see below). (note, for normal and weibull this implies a truncated distribution--stick with lognormal and gamma only for these?)

See Lecomte et al. (2013) for biomass estimation with zero-inflated data, using a gamma distribution ("Delta-gamma approach"). Note that they show some issues with variable sampling effort (volume), which we will have to assume is constant here.

# Assessing fits: equivalence testing

The Kolmogorov-Smirnov test is a common test used for the equality of distributions (ref), where the null hypothesis is that the empirical distribution follows the named theoretical density function. As with most statistical tests, failure to reject the null hypothesis is not equivalent to accepting the null hypothesis, especially in cases of low power (small sample size). Therefore failure to reject the null hypothesis does not necessarily indicate an adequate match between empirical and theoretical distribution. Furthermore, in cases of high power (large sample sizes), statistical significance does not indicate large effect size (or an inadequate fit), such that the chosen distribution may be adequate even when the null hypothesis is rejected.

Robinson and Froese (i think! (Robinson and Froese 2004) ref) recommend an equivalence test to compare empirical data to model predictions using a two-one-sided t-test (TOST). In equivalence testing a maximum allowable error (or error tolerance) is defined, and the null hypothesis is that the observed distribution is outside of the error tolerance relative to the theoretical distribution. If the observed distribution is seen to be within the maximum error (or error tolerance), then the null hypothesis is rejected and the observed data is judged to be "equivalent" to the theoretical distribution. Here we use TOST to assess adequate matching between our observed empirical cumulative distribution of fuel type and the theoretical cdf associated with each candidate distribution. Let be the ith quantile of the empirical data distribution, and be the ith quantile of the theoretical distribution. Then the difference between the observed and theoretical () is:

We then calculate as the mean distance between observed and theoretical, and use TOST to determine statistically if the observed and theoretical distributions differ by more than a specified error tolerance (). This requires a standard error tolerance to be specified, which for our application would be a relatively arbitrarily defined threshold.

Prichard et al (; (Prichard et al. 2014)) use a similar equivalence procedure to evaluate the uncertainty of the fits of observed fuel consumption relative to those predicted by empirical consumption equations. For their analysis, rather than choosing a single arbitrary error threshold, they repeated the equivalence test with increasing until the first at which the equivalence test null hypothesis was rejected. This then defined the bound of uncertainty for that fuel type. We adapt their approach here, repeating the equivalence test for increasing error thresholds between observed and theoretical distributions for distributions estimated both with zeroes (and an offset), and distributions estimated for only values > 0. We then compare the minimum that rejects the null hypothesis to assess both the uncertainty in the distribution estimates and the relative fits between the two approaches (since these aren't nested, can't compare likelihoods directly--except maybe we can, need to think more about calculating a likelihood for a full hurdle model).

We like this approach because it assesses adequacy of empirical fit to theoretical, rather than statistical differences. We will use the equivalence package in R.

citation("equivalence")

##   
## To cite package 'equivalence' in publications use:  
##   
## Andrew Robinson (2016). equivalence: Provides Tests and Graphics  
## for Assessing Tests of Equivalence. R package version 0.7.2.  
## https://CRAN.R-project.org/package=equivalence  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {equivalence: Provides Tests and Graphics for Assessing Tests of Equivalence},  
## author = {Andrew Robinson},  
## year = {2016},  
## note = {R package version 0.7.2},  
## url = {https://CRAN.R-project.org/package=equivalence},  
## }  
##   
## ATTENTION: This citation information has been auto-generated from  
## the package DESCRIPTION file and may need manual editing, see  
## 'help("citation")'.

# OR!

We can just use bootstraps of the parameter estimates for each distribution, or look at MC null KS tests (following (Lilliefors 1967)) and see if we get significance.

Method: calculate KS statistic for observed distribution relative to "theoretical" distribution at estimated parameter values. Then, for 5000 MC replicates, take n (n=number of observed values in original distribution fit) random draws from "theoretical" distribution at estimated parameter values. For each of these, estimate the same theoretical distribution, then perform KS test of random to theoretical distribution at estimated parameter values. This will generate 1000 KS values when the null hypothesis is true, thus a "null" distribution. The p-value is then calculated as:

where is the number of simulated statistics in the null distribution, is the ith simulated statisic, and is the observed statistic. I is an indicator function that takes a value of 1 if the observed statistic is less than the simulated, 0 otherwise. The sum tallies the number of simulated statistics are smaller than the observed statistic. Note that we divide by +1 because we have +1 total statistics (including ). We can then evaluate, against say = 0.10, which if any distributions are "fail to reject" (ftr) The question, as above, is whether this is a good idea with respect to power and interpretation. (We'll see how many are ftr.)

# Decisions to be made

By what criteria(on) will we deem a distribution to be adequate for a given fuel load type and EVT group?

1. Set a maximum equivalence threshold and only accept those within that threshold. Drawbacks: thresholds are arbitrary unless there is a real accepted design specification!
2. Only accept those that ftr the MC KS test. Drawbacks: power!
3. Accept the best-fitting distribution, and define uncertainty around that using either: equivalence UA, or bootstrap CI. Drawback: Still feel like there needs to be some kind of poor fit beyond which we cannot accept a distribution. Also, how to incoporate that into random draws? Or don't incorporate, just use as part of visualization of uncertainty. Bootstrap standard errors as a CV--error/estimate
4. Use bootstrap to create an envelope of values in the ecdf, and see if the observed ecdf leaves the envelope. Presumably similar results to the MC KS test. Drawbacks: new! Different?

Also, ways to quantify uncertainty: equivalence threshold for "distribution error", bootstrap sd on the estimated parameter values (relative to estimate, so a bootstrap CV for the estimate)

## Results

Here we give some example results from all of the options described above. In general we're finding that the hurdle fits are either better than, or indistinguishable from, non-hurdle fits. I think we should just use that for all fuel types, understanding that for most of the fuel types there are no zeroes in the database (such that = 0).

Can create a master table with all EVTs, the best-fitting distribution, parameter estimates, bootstrap standard deviations and CV, equivalence values, MC KS p-values, and proportion of zero entries in the database.

Note, all results are generated by loading the following RData workspace (generated in the FitHurdleUnivariateDistributions.R script): HurdleFits.RData, BootstrapHurdleFitsList.RData, EquivalenceResults.RData, MasterTable.RData

Here is an example of the master summary table, for EVT 697 (Loblolly Pine Forest and Woodland)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fuelType | distr | prop0 | param1.est | param1.sd | param1.cv | param2.est | param2.sd | param2.cv | KS.pVal | equivalence.val |
| X100hr\_loading\_Mgha | gamma | 0.416 | 1.772 | 0.085 | 0.048 | 0.342 | 0.019 | 0.056 | NA | NA |
| X10hr\_loading\_Mgha | logNormal | 0.178 | -0.017 | 0.026 | -1.568 | 0.854 | 0.018 | 0.022 | NA | NA |
| X10Khr\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| X10KhrR\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| X10KhrS\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| X1hr\_loading\_Mgha | logNormal | 0.056 | -2.440 | 0.027 | -0.011 | 0.952 | 0.019 | 0.020 | NA | NA |
| X1Khr\_loading\_Mgha | logNormal | 0.170 | 2.289 | 0.162 | 0.071 | 1.031 | 0.114 | 0.111 | NA | NA |
| X1KhrR\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| X1KhrS\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| cwd\_loading\_Mgha | logNormal | 0.584 | 1.262 | 0.055 | 0.044 | 1.269 | 0.039 | 0.031 | NA | NA |
| cwd\_rotten\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| cwd\_sound\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| duff\_depth\_ | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| duff\_loading\_Mgha | logNormal | 0.112 | 1.779 | 0.030 | 0.017 | 0.982 | 0.021 | 0.021 | NA | NA |
| fwd\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| GT10KhrR\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| GT10KhrS\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| herb\_loading\_Mgha | gamma | 0.249 | 0.703 | 0.027 | 0.039 | 1.791 | 0.099 | 0.055 | NA | NA |
| herb\_percentlive\_ | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| lichen\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| litter\_depth\_ | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| litter\_loading\_Mgha | weibull | 0.027 | 1.477 | 0.033 | 0.022 | 12.885 | 0.260 | 0.020 | NA | NA |
| moss\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| ms\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| os\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| shrub\_loading\_Mgha | logNormal | 0.122 | 0.697 | 0.023 | 0.034 | 0.798 | 0.017 | 0.021 | NA | NA |
| snag\_loading\_Mgha | logNormal | 0.537 | 0.508 | 0.052 | 0.102 | 1.242 | 0.036 | 0.029 | NA | NA |
| tree\_crown\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| tree\_loading\_Mgha | weibull | 0.008 | 1.400 | 0.031 | 0.022 | 84.948 | 1.801 | 0.021 | NA | NA |
| us\_loading\_Mgha | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

And here is how it would look if we dismiss the NAs:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fuelType | distr | prop0 | param1.est | param1.sd | param1.cv | param2.est | param2.sd | param2.cv | KS.pVal | equivalence.val |
| X100hr\_loading\_Mgha | gamma | 0.416 | 1.772 | 0.085 | 0.048 | 0.342 | 0.019 | 0.056 | NA | NA |
| X10hr\_loading\_Mgha | logNormal | 0.178 | -0.017 | 0.026 | -1.568 | 0.854 | 0.018 | 0.022 | NA | NA |
| X1hr\_loading\_Mgha | logNormal | 0.056 | -2.440 | 0.027 | -0.011 | 0.952 | 0.019 | 0.020 | NA | NA |
| X1Khr\_loading\_Mgha | logNormal | 0.170 | 2.289 | 0.162 | 0.071 | 1.031 | 0.114 | 0.111 | NA | NA |
| cwd\_loading\_Mgha | logNormal | 0.584 | 1.262 | 0.055 | 0.044 | 1.269 | 0.039 | 0.031 | NA | NA |
| duff\_loading\_Mgha | logNormal | 0.112 | 1.779 | 0.030 | 0.017 | 0.982 | 0.021 | 0.021 | NA | NA |
| herb\_loading\_Mgha | gamma | 0.249 | 0.703 | 0.027 | 0.039 | 1.791 | 0.099 | 0.055 | NA | NA |
| litter\_loading\_Mgha | weibull | 0.027 | 1.477 | 0.033 | 0.022 | 12.885 | 0.260 | 0.020 | NA | NA |
| shrub\_loading\_Mgha | logNormal | 0.122 | 0.697 | 0.023 | 0.034 | 0.798 | 0.017 | 0.021 | NA | NA |
| snag\_loading\_Mgha | logNormal | 0.537 | 0.508 | 0.052 | 0.102 | 1.242 | 0.036 | 0.029 | NA | NA |
| tree\_loading\_Mgha | weibull | 0.008 | 1.400 | 0.031 | 0.022 | 84.948 | 1.801 | 0.021 | NA | NA |
| > page\_break |  |  |  |  |  |  |  |  |  |  |

# References

Lecomte, Jean Baptiste, Hugues P. Benoît, Sophie Ancelet, Marie Pierre Etienne, Liliane Bel, and Eric Parent. 2013. “Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume.” *Methods in Ecology and Evolution* 4 (12): 1159–66. doi:[10.1111/2041-210X.12122](https://doi.org/10.1111/2041-210X.12122).

Lilliefors, H W. 1967. “On the Kolmogorov-Smirnov test for normality with mean and variance unknown.” *American Statistical Journal* June (318): 399–402.

Prichard, S J, E C Karau, R D Ottmar, M C Kennedy, J B Cronan, C S Wright, and R E Keane. 2014. “Evaluation of the CONSUME and FOFEM fuel consumption models in pine and mixed hardwood forests of the eastern United States.” *Canadian Journal of Forest Research-Revue Canadienne de Recherche Forestiere* 44 (April): 784–95. doi:[DOI 10.1139/cjfr-2013-0499](https://doi.org/DOI 10.1139/cjfr-2013-0499).

Robinson, Andrew P., and Robert E. Froese. 2004. “Model validation using equivalence tests.” *Ecological Modelling* 176 (3-4): 349–58. doi:[10.1016/j.ecolmodel.2004.01.013](https://doi.org/10.1016/j.ecolmodel.2004.01.013).