# Analysis of data collected under the Amphibian Protocol

2017-03-09

# 1. Summary of protocol:

## 1.1 Basic protocol

There are two protocols – calls and visual surveys.

As taken from the protocol document for call surveys:

“Define a survey transect along or through your wetland. … Surveyors visit the monitoring site(s) in spring and listen for calling males, recording the species and approximate number of each. Repeat surveys increase the probability that species will be detected.”

As taken form the protocol document for visual surveys:

“Surveyors monitor breeding site(s) during the active season (spring to fall), walking the shoreline of a wetland recording all species and life stages encountered. These include egg masses, tadpoles and adults. Repeat surveys increase the probability that species will be detected.”

The data collected under this protocol at each survey consists of:

* **Species**. The species of amphibian counted.
* **Life Stage**. The life stage of the species.
* **Count**. The number of amphibians at this lifestage.

## 1.2 Cautions about the protocol.

#### 1.2.1 Missing value indicates 0 or no visit.

If no species were seen during a visit, this is indicated by lack of a record for that visit. It will be necessary to assume that all species will have been searched for at each visit and that lack of count for a species implies it was searched for and not seen, rather than the observed did not know how to identify the species.

It is unclear how NO amphibians of any species will be recorded in the database?

If no visit was made to a transect, then there are no records in the *General Survey* for that transect. One must infer that if there are no records for transect at a date that it was not visited. The *Transect Information* worksheet only has information on the transect label and not when they were visited. It is preferable to include in this sheet the visit dates of each transect in that year explicitly rather than trying to infer this information from the *General Survey* sheet.

## 2. Database structure

The database for this protocol is a series of Excel workbooks with multiple sheets in each workbook. The *Transect Information* sheet contains the information on the transect available for this year. If a transect is not visited, this is indicated by a no records in the *General Survey* workseeht. The *General Survey* sheet contains the information collected. There are multiple lines per visit. If a visit was done but no amphibians were detected, then the species code is set to missing with a count of 0.

The relevant fields on the *General Survey* worksheet are:

* *Transect Label.*
* *SurveyType.* Was this an visual or calling survey?
* *Date*. The date the data was collected. The *Year* is extracted from this date.
* *Species*. What species were seen
* *LifeStage*. What life stage was counted
* *Count*. Count of the number of birds of each species.

# 3. Sample Analyses.

A sample analysis is presented on the *Alice Lake* study area.

This design could have multiple transects that are repeated measured over time with multiple plots measured on each transect that are also repeated measured over time. Please refer to the *Fitting Trends with Complex Study Designs* document in the *CommonFile* directory for information on fitting trends with complex study designs. For the Alice Lake Example, there is only one transect which simplifies the analysis considerably.[[1]](#footnote-1)

All analyses were done using the *R* (R Core Team, 2016) analysis system. An HTML document showing the results of the analysis is available. All plots are also saved as separate \*png files for inclusion into reports.

## 3.1 Analysis of Calling Data

According to the protocol, the approximate number of calls heard during the visit should be recorded. This is not available in the data base, only presence data is recorded. Given that only 1 station was visited in Alice Lake with 2 visits in 2013 and 1 visit in 2014, no analysis on presence/absence is possible at this time. A logisitic regression analysis may be possible if multiple stations are visited in each year and presence/absence of sounds is recorded.

The usual protocol for aural surveys is an occupancy study to deal with the problem of false negatives (Mackenzie et al 2005). In this protocol, multiple stations are visited multiple times and the pattern of detection at each station is used to infer the probability of detection and hence to adjust for false negatives.

## 3.2 Analysis of Visual Data – Number of Egg Masses.

This analysis will look at trend in the total number of **detected** egg masses (over all species). An analysis at the individual species-level could also be done, but the current data is extremely sparse at the species level and likely to be uninformative.

A key problem with this protocol is the emphasis in the previous paragraph on **detected.** It is not possible to count all of the egg masses laid in the study area; it is also unlikely that all eggmasses along the (fixed) transects will be detected in each year. Consequently, it is necessary to make the VERY STRONG assumption that detectability is constant over time for each transect. This may be violated, for example, when different numbers of visits are made across years on the same transect (e.g. two visits were made in 2013, while only one visit was made in 2014 and 2015). Presumably, if more visits are made to a transect, then more egg masses may be detected.

We also need to make the strong assumption that egg masses are not double counted across multiple visits to the same transect. There is no information stored on the database on the exact location where an egg mass was located so it is difficult to verify this information.

Even if the above assumptions are satisfied, the number of observed egg masses is only an INDEX to the population number of egg masses.

The data is first summarized to the year level for each transect counting the number of records that identify an eggmass. This reduces the data to one measurement per transect per site/year.

A summary plot of the (reduced) data is shown in Figure 1.

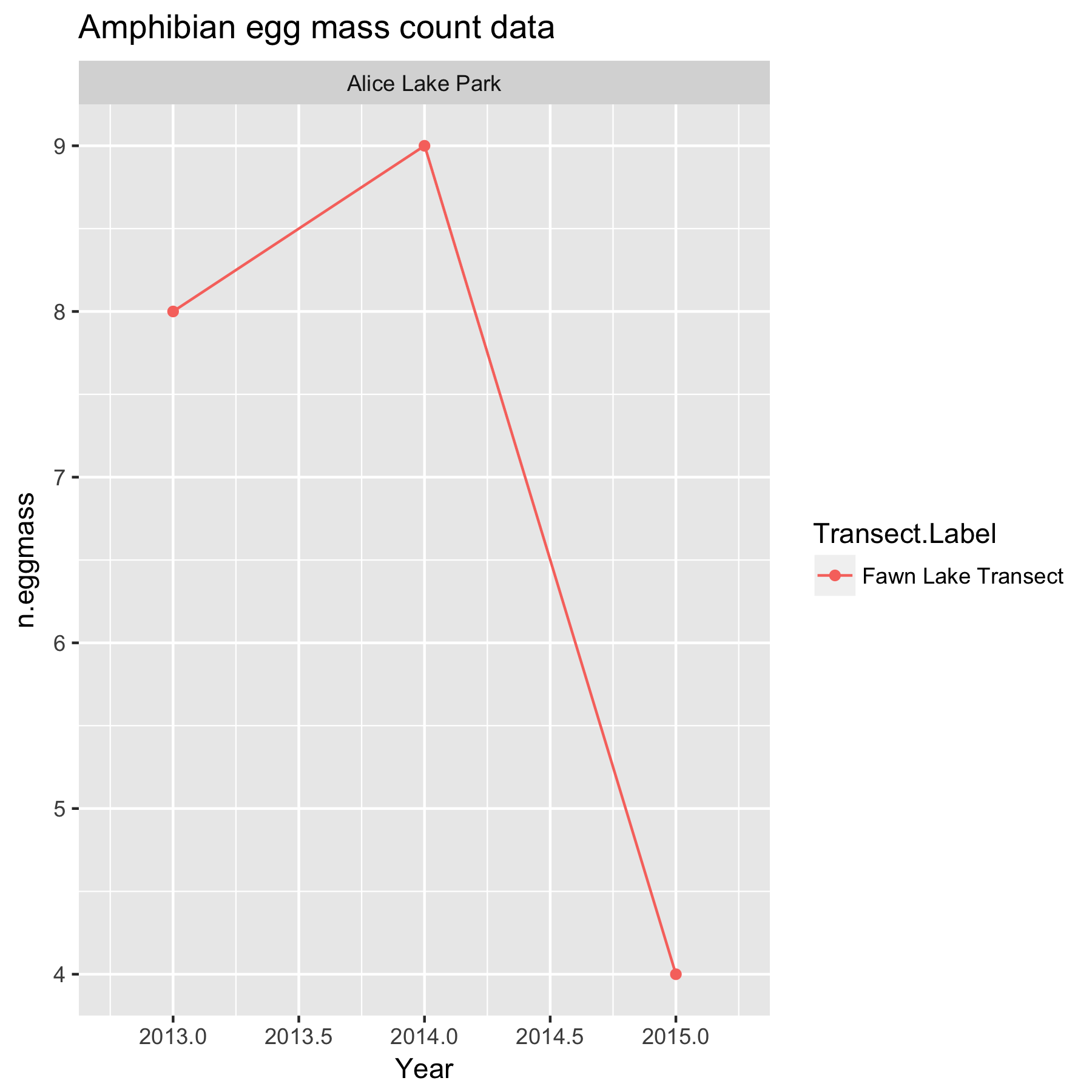


Figure 1. Summary plot of the data.

There is only one transect measured at Alice Lake Park, so there are no concerns about transeect-to-transect variation. The year-specific factors have been automatically included in the variation of the single transect over time and so it is not necessary to explicitlymodel these as well.

Because only a single transect was measured, A simple linear Poisson regression can be used to look for changes over time using the model



where *NEggMasses* is the count for that transect in a year; and *Year* represents the calendar year trend over time.

This model assume that effects are multiplicative over time, so that the actual fit is done on the logarithmic scale. For example, a trend may assume that there is constant 5% change over time rather than a fixed 1 unit change per year. An approximate analysis could be done using regular linear regression if you analyze the *log(NEggMass)*, but problems can arise if some of the count values are 0 (log(0) is not defined).

The Poisson model for a single transect can be fit using the g*lm()* function in *R.* Figure 2 shows asummary plot, along with estimates of the slope, its standard error, and the p-value of the hypothesis of no trend. There is no evidence (p=0.49) of a trend with an estimated slope of 0.00 (SE -.29) /year in the logarithm of the mean count response. This corresponds to an approximate exp(-.29)= 0.75 multiplicative change/year, i.e. the mean count in 2014 is about 0.75x the mean count in 2013, and the mean count in 2015 ix 0.75x the mean count in 2014. Because the analysis is done on the logarithmic scale, the fitted trend line will look non-linear on the original (non-transformed) scale.

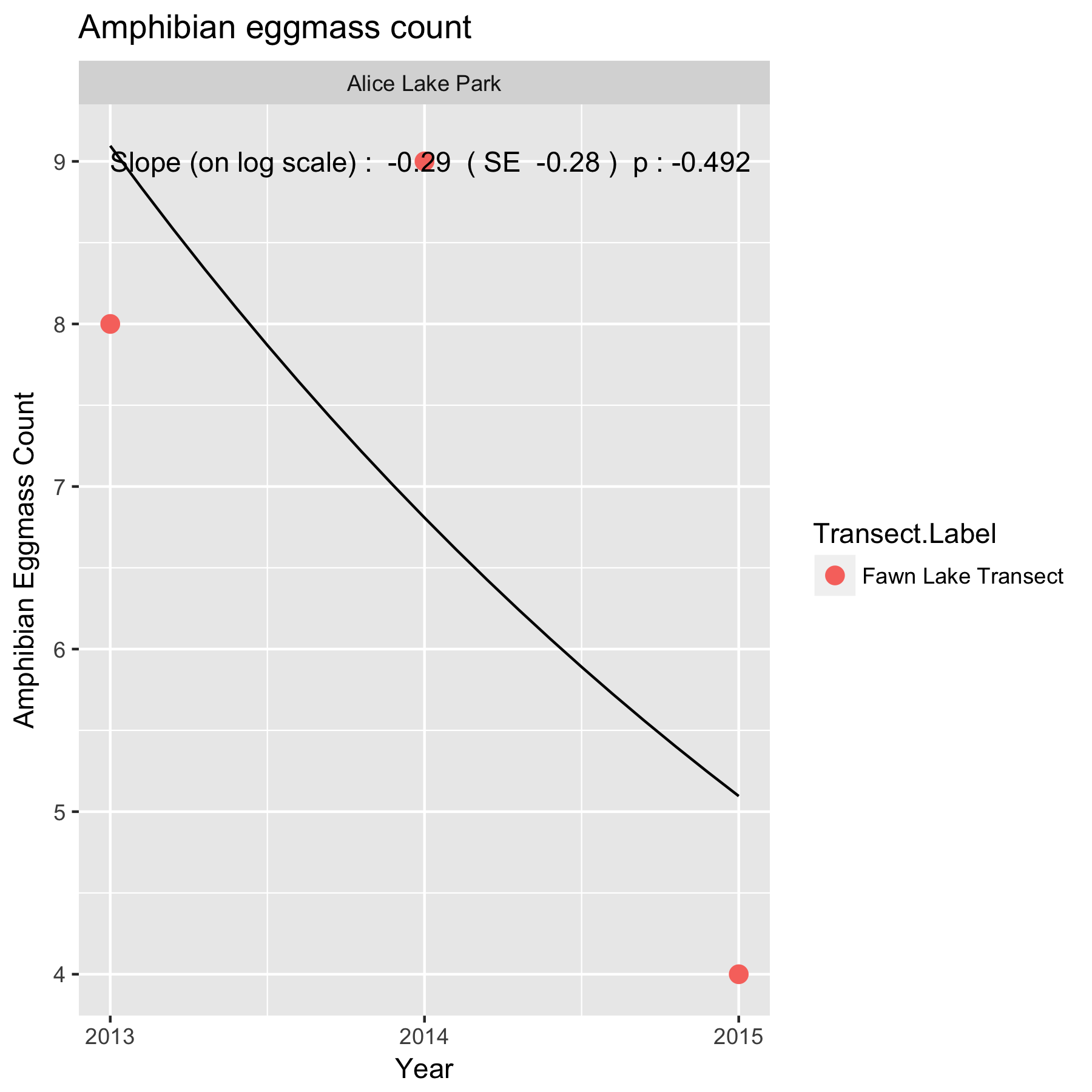


Figure 2. Summary plot of the trend in number of egg masses at *Alice Lake*. Because the analysis was done on the logarithmic scale, the fitted trend line is not a straight line but curved.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 3.

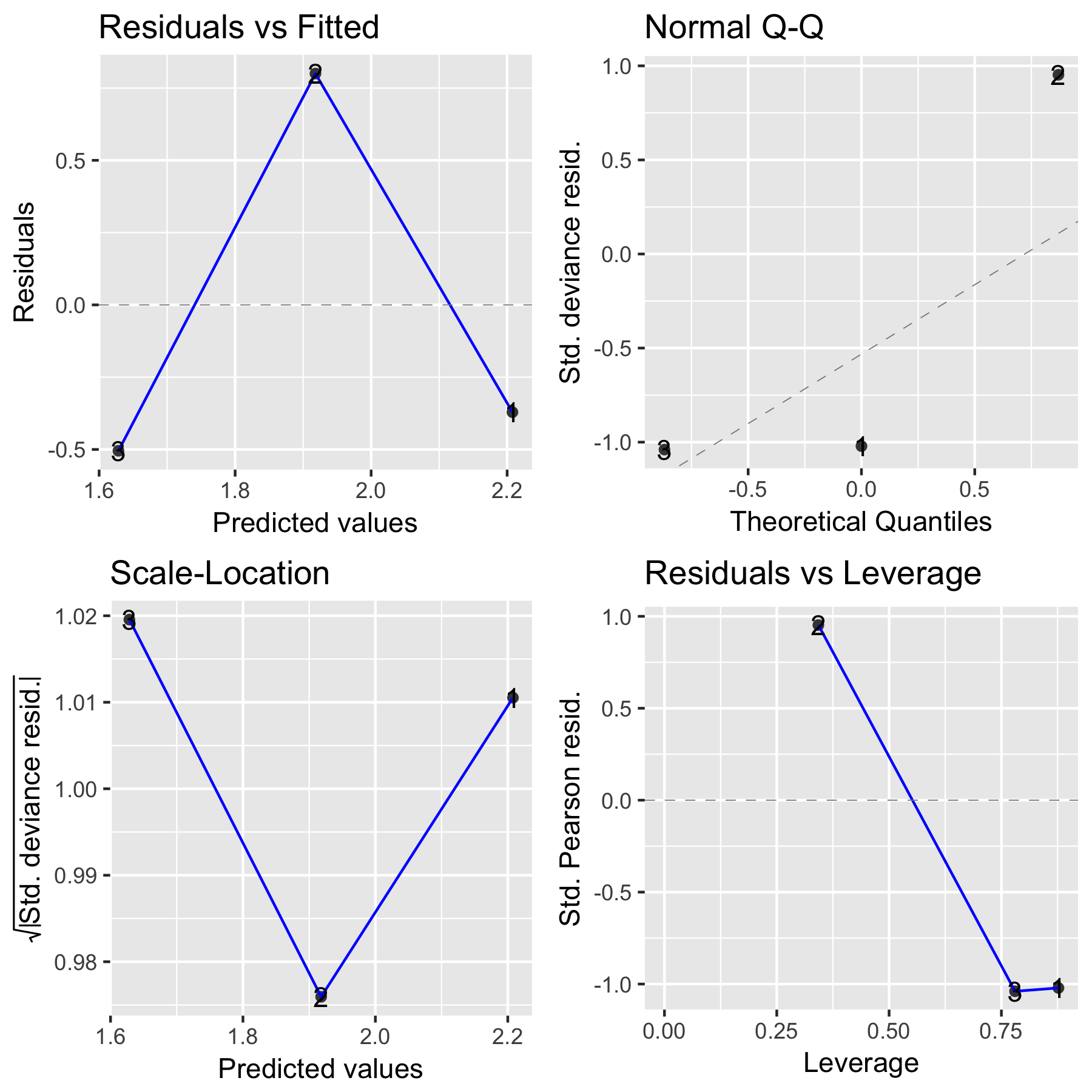


Figure 4. A sample diagnostic plot for the analysis of number of egg masses at *Alice Lake.*

With only 3 years of data, the plots are not very informative. In the upper left corner is a plot of residuals vs. the fitted values. A good plot will show a random scatter around 0. Any large deviations from 0 should be investigated as potential outliers. In the upper right is a normal probability plot of the residual. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality of the residuals (and in fact makes no assumption of normality) so only extreme departures are worrisome. The bottom left plot examine the assumption that the variation about the line is constant over the line. You would expect to see a constant band of points. Finally the bottom right plot is a leverage plot – this is not useful for this simple model and can be ignored.

It will also be possible to covariates such as mean winter temperature or degree days in the year to try and explain some of the variation over time using a multiple regression. With only three years of data available, this not sensible.

Whenever an analysis of a trend over time is conducted, the analysis should test and adjust for autocorrelation. Autocorrelation usually isn’t a problem (and likely cannot be detected) unless you have 10+ years of data. The test for autocorrelation commonly used is the Durbin-Watson test. There was no evidence of autocorrelation.

This analysis was conducted at the total count level (over all species) but could be done for individual species – but the data are likely to very sparse and not useful. One potential problem is that in some cases, species information is only recorded at the Genus or higher level. In these case, this data will have to discarded when the analysis is done at the species level, but then you are making an implicit assumption that recording at the Genus level happens at random and is unrelated to the response. If this assumption is violated (e.g. perhaps when there are larger number of birds, it is too difficult to record at the individual species level) then this is not occurring at random and some effort must be made to “split” the genus level information among the species.

In theory, the actual number of eggs in each eggmass could also be used rather simply the number of eggmasses. I did not do this analysis because (a) it makes little sense to average the size of the eggmass over the different species and (b) the number of eggs in the egg mass may not represent the total number laid because of unknown amounts of predation.

## 3.3 Analysis of Visual Data – Number of Adults.

A similar analysis can be done on the total number of adult amphibians detected. In *Alice Lake* there were only a total of 7 frogs detect over the three years – the data is just too sparse to analyze at this time.

# 4. Summary

Some caution is required to ensure that all transects are visited equally often in a year. In this balanced design, it is straightforward to compute statistics over all measurements of a transect and all transects in a year have the same number of visit. It is possible to modify the analysis is only some transects are visited on a particular date with an unequal number of visits to a transect in a year. A simple way to deal with unbalance would be to delete some of the observations, but better methods are available.

Trends over time could also occur in the diversity of the amphibians. In theory, standard diversity measures could be used and tracked over time, but these have a very strong assumption that all species are equally detectable by the observer. This is unlikely to be true. Secondly, the actual counts are quite small, and diversity measures that rely on actual counts (e.g. Simpson’s diversity) will not perform well. For this reason, I do not recommend an analysis on the diversity of the observations.

I am extremely dubious that this protocol as implemented will produce any useful monitoring information at the study area level except for disaster detection (e.g. all species vanish). Even then, the absence of a detection does not imply that the species was not present – it could simply be a false negative. The current study does not provide any information to estimate detection probabilities.

As an example of how this protocol could be improved is based on large scale surveys of occupancy. The shoreline would be divided in many (at least 50) segments. On each visit, the presence/absence of amphibians would be recorded at the segment level. Then standard occupancy models could be used to estimate the proportion of segments that are occupied. Changes in the occupancy probability over time would be an indication that the population is increasing/decreasing over time.

Or this particular study area could be embedded into a much larger regional analysis where again, multiple visits are made in a year to estimate detection probabilities and changes in occupancy serve as an index to changes in the population.

References

[MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L.L. Bailey, and J. E. Hines. 2005. Occupancy Estimation and Modeling - Inferring Patterns and Dynamics of Species Occurrence.](http://www.elsevier.com/wps/find/bookdescription.cws_home/706666/description) Elsevier Publishing.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Appendix A.

Issues encountered when doing a trial analysis on the *Alice Lake* study area data.

The following issues were encountered in the databases when a trial analysis on the Alice Lake study area data was performed. The spreadsheets for the sample analysis were corrected prior to the analysis.

(a) The Excel workbook for 2014 and 2015 has two rows at the end of the *General Survey* spreadsheet where at least one column has blank (rather than being empty). This causes the programs to insert a row with missing values when reading the data.

(b) Inconsistency in Transect labeling.

The same transect appears to be used in all years, but has different labels

Year

Transect.Label 2013 2014 2015

Fawn Lake Station 2 1 0

Fawn Lake Tansect 0 11 9

Fawn Lake Transect 15 0 0

(c) Species code labeling not consistent. The worksheets use A-xxxx to refer to amphibian species. An inconsistency when recording at the genus level, i.e. “A” should be added as a prefix, e.g. *A-AMPHIBIAN*

(d) Survey Type field was missing in 2015 data. It is assume to be all VI (visual surveys),

(e) Life Stage codes not consistent across years

Life.Stage Code

Year A AD E EG J LA

2013 0 3 0 8 0 6

2014 0 3 0 9 0 0

2015 4 0 4 0 1 0

(f) Inconsistency in field names.

In 2013, the count field is labelled as *Count..for.visual.surveys* while in other years it is labelled simply as *Count*. For some reason, the field name also has an “invisible blank” at the end i.e. *Count.* where the . represents a blank.

Other inconsistency in field name are:

"Background.Noise..for.auditory.surveys."

"Calling.Noise..for.auditory.surveys."

"Preceeding.24.hour.precipitation"

"Preceeding.24.hour.precipitation..mm."

(g) Date formatting. I suggest you always use yyyy-mm-dd as the format for ALL protocols. This set of workbooks currently use *dd-mmm-yy* format (with 2 digit years).

1. The R code is general enough that if more than one transect is present, it will automatically choose the more complex linear mixed models as seen in the other protocols. [↑](#footnote-ref-1)