# Analysis of data collected under the Waterfowl Protocol

2017-03-09

# 1. Summary of protocol:

## 1.1 Basic protocol

As taken from the protocol document:

“Establish observation points at enough sites to provide views of the entire survey area. …

Choose a date after which migrants are mostly gone. Count on 3 separate occasions during a 2-3 week period and keep highest count.”

The data collected under this protocol at each survey consists of the.

* **Species**. The species of waterfowl counted.
* **Sex**. The sex of the species.
* **Count**. The number of birds of this species and sex. Choose a date after which migrants are mostly gone

## 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 the species code being set to missing and the count field being set to 0. This needs to be done consistently over time.

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

## *1.2.2 Classifying by season*

The protocol is silent about seasonal sampling, e.g. spring vs. fall. There is some issue on how to properly pool information that is close together but crosses year boundaries. For example, surveys may be done in late December of 2013 and early January of 2014. If the maximum counts over multiple visits in the winter are collected, visits on these two months need to be “combined” even though they are in different calendar years.  
  
For this analysis we classified by month (November, December, January, February) to avoid this problem

## 1.2.3 Using the maximum count over multiple visits.

A key problem in using the maximum/minimum counts is that sample size (number of visits) has an influence on the measures, unlike measures such as the mean where sample size only affect the precision, but not the measurement. Extreme values (such as the maximum) will tend to increase with sample size. So when looking at yearly trend, an increase may reflect more measurement, rather than an increase in the mean.

Rather than using the maximum over multiple visits, a measure of central tendency (such as the median) would be preferred rather than the maximum.

## 2. Database structure

The database for this protocol is a series of Excel workbooks with multiple sheets in each workbook. The *Sample Station Information* sheet contains the information on the stations available for this year. If a station is not visited, this is indicated by a missing record 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 birds were seen, then the species code is set to missing with a count of 0.

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

* *Sample Location Label.*
* *Date*. The date the data was collected. The *Year* is extracted from this date.
* *Species*. What species were seen
* *Count*. Count of the number of birds of each species.

# 3. Sample Analyses.

A sample analysis is presented on the *Yellow Point Bog Ecological* study area. Data is available only for 2013 and 2014. Data for 2015 were simulated by drawing 30 records from the 2013/2014 data and changing the year to 2015.

This design has multiple stations that are repeated measured over time with multiple measurements at each station. Please refer to the *Fitting Trends with Complex Study Designs* document in the *CommonFile* directory for information on fitting trends with complex study designs.

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 Maximum counts.

The data is first summarized to the date level by summing the count over multiple records (the different species) for each date-sample station combination. This reduces the data to one measurement per date per site/year.

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

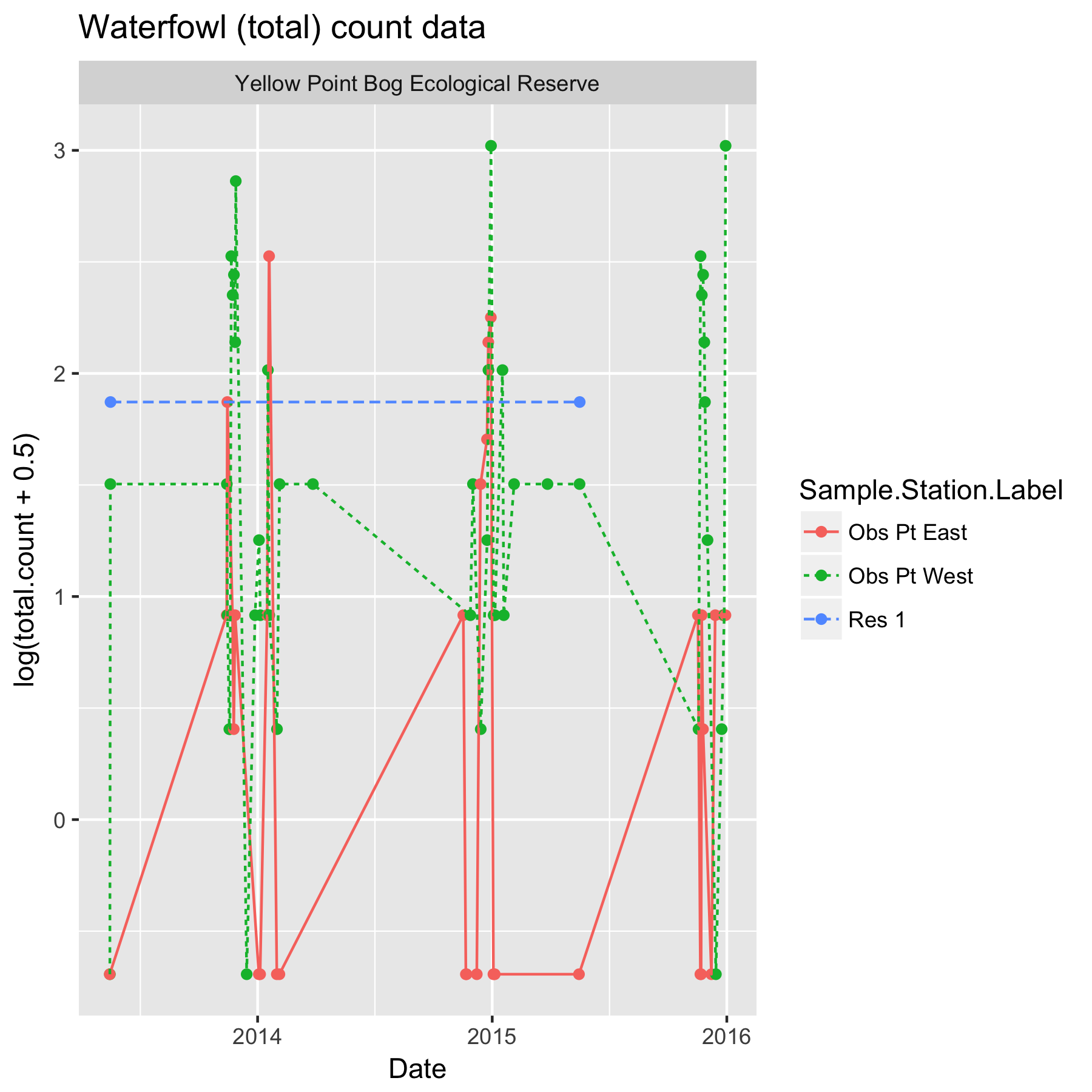


Figure 1. Summary plot of the data. Data is plotted on the logarithmic scale because of the wide range of values. Because of potential zeros in the data, 0.5 was added to all points only for plotting purposes.

Notice that there is a definite station effect, where, for example, the number of birds at station *Obs Pt West* generally is higher than at the other stations because of local station-specific conditions (e.g. better habitat).

Next the maximum count in each month in each station (but only November, December, January, and February) is found and plotted in Figure 2. As noted earlier, a better measure may be the median count rather than the mean count. The analysis would proceed in a similar fashion.

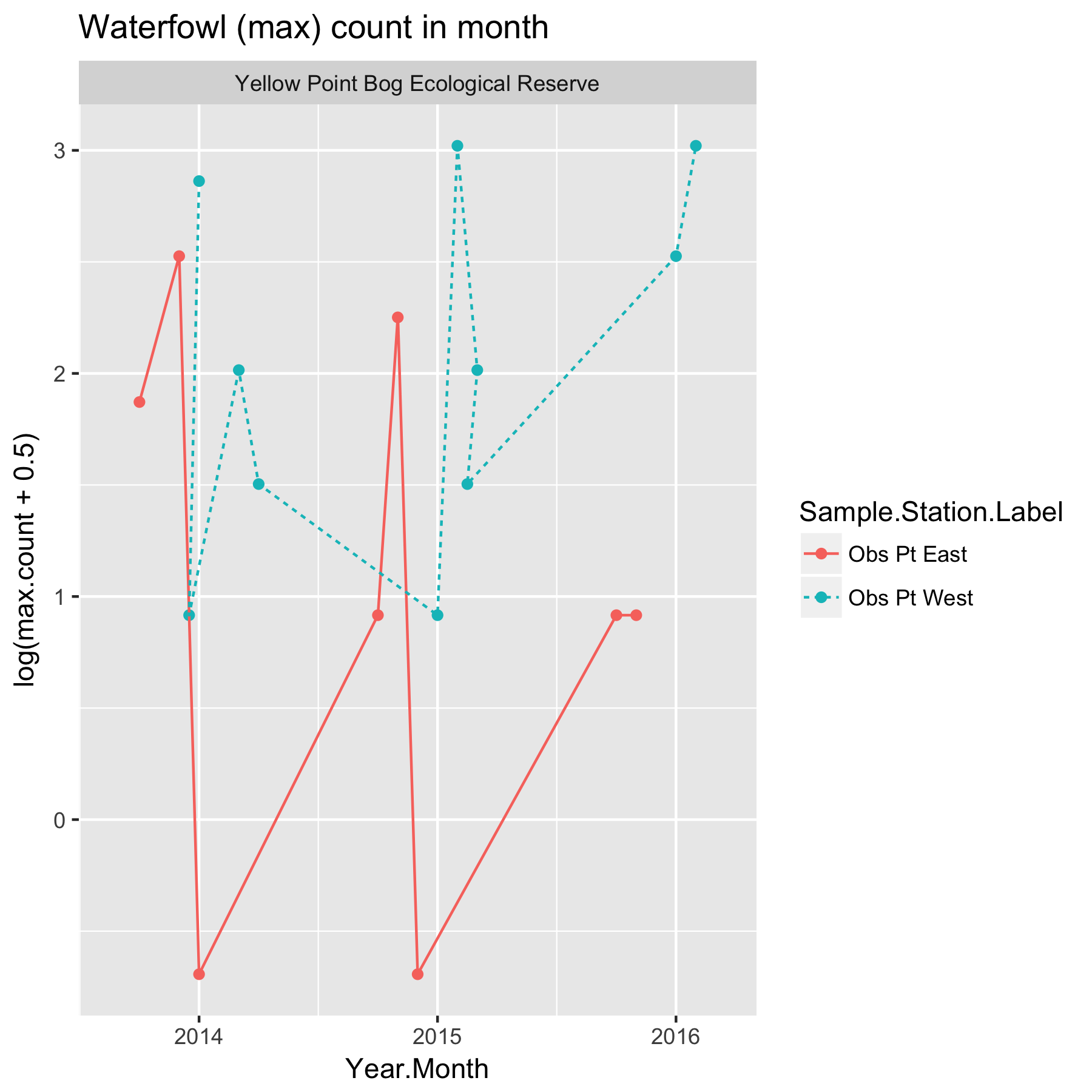


Figure 2. Plot of the maximum count in each station in November through February. Data is plotted on the logarithmic scale because of the wide range of values. Because of potential zeros in the data, 0.5 was added to all points only for plotting purposes.

Again there is evidence of a station effect. A month effect is less clear.

Because this is count data, so a linear mixed model is fit to the logarithm of the maximum call in each month at each transect in each year. The model is:



where *log(MaxCount)* is logarithm of the maximum count at that station in that month in that year; *StationF* represents the statioon effect; *YearF* represents the year-specific effects (process error), and *Year* represents the calendar year trend over time. The *StationF* term allows for the fact that station-specific conditions may tend to affect the counts on this station consistently over time. The *YearF* term represent the year-specific effects (process error) caused by environmental factors (e.g. a warmer than normal year may elicit more visits from waterfowl).

This model implicitly assumes that the trend in Nov, Dec, Jan and February are the same but with a different intercept (i.e. parallel lines on the logarithmic scale).

Model fit on the logarithmic scale assume that effects are multiplicative over time, so that the when the actual fit is done on the logarithmic scale, the trends are linear. For example, a trend may assume that there is constant 5% change over time rather than a fixed 1 unit change per year. Some caution is needed if any of the values are 0 as log(0) is not defined. In these cases, a small constant (typically ½ of the smallest positive value in the dataset) is added to all values before the analysis proceeds.

The model was fit using the *lmer()* function in *R.* Figure 3 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=0.76) of a trend with an estimated slope of -0.11 (SE 0.76) /year in the logarithm of the mean maximum count response. This corresponds to an approximate exp(-0.11)=0.89x multiplicative change/year, i.e. the mean maximum count in 2014 is about 0.89x the mean maximum count in 2013, and the mean count in 2015 is 0.89x 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.

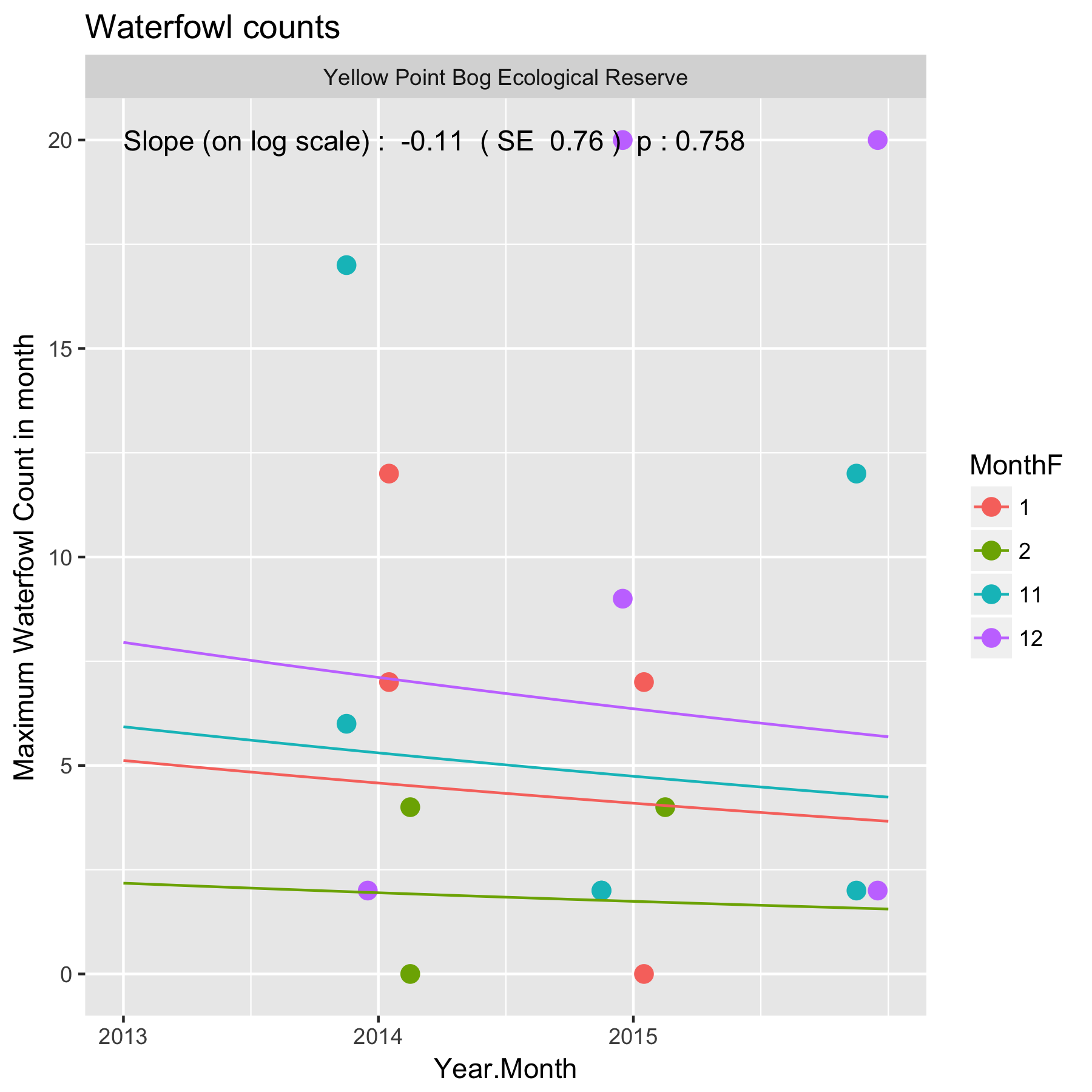


Figure 3. Summary plot of the trend in maximum birds counts at *Yellow Point Bog*. Because the model was fit on the logarithmic scale, the fitted trend line is not a straight line but curved.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 4.

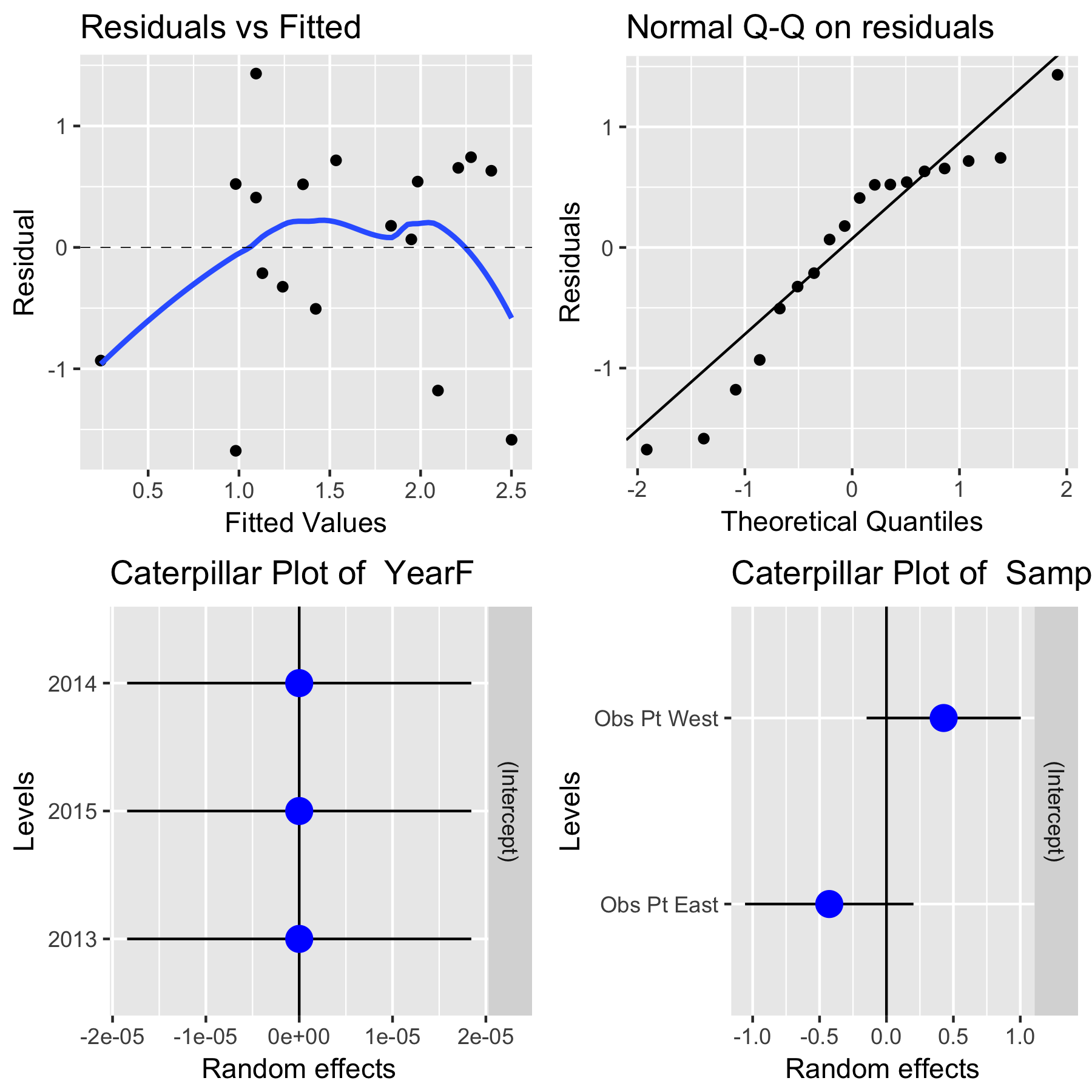


Figure 4. A sample diagnostic plot for the analysis of maximum bird counts at *Yellow Point Bog.*

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 shows that the year specific effects are very small (the dots are all close to 0) The bottom right plot shows that the station effects with *Obs Pt. West* tending to have higher counts than *Obs Pt. East*.

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. No evidence of autocorrelation over time was found.

The analysis of the median count per month would follow the same steps as shown above and the *R* code is easy to modify. Similarly, this analysis was conducted at the total count level (over all species) but could be done for individual species. 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.

Trends over time could also occur in the diversity of the birds. 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.

# 3. Summary

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

References

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 Yellow Point Bog study area data.

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

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

(c) The *General Survey* datasheet also breaks out the total count by sex and lifestage, so the sex field is likely not needed.

(d) The protocol is silent on the time of the year when surveys should be done. The current data occurs mostly in November to February but a few measurements are taken in the summer as well. I have discarded the summer measurements.

(e) 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).