# Analysis of data collected under the Alpine Protocol

2017-03-09

# 1. Summary of protocol:

## 1.1 Basic protocol

As taken from the protocol document:

“Permanent transects are established from just below present treeline to just above the transition from vegetation to rock (or ridgeline, whichever comes first). A 50cm X 50cm quadrat is sampled at regular intervals along the transect.

The % foliar cover by species is recorded in each quadrat.”

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

* **Species**. The species of vegetation in the plot.
* **% foliar cover.** What fraction of the plot is covered by this species.

## 1.2 Cautions about the protocol.

#### 1.2.1 Missing value indicates 0

If a species was not seen on a plot during a visit, this is indicated by the lack of a record for that species code. Unfortunately, it is now impossible to know if the species were actually looked for and not seen, or the species could not be identified and so was not recorded. This issue could be problematic when looking at the impact of climate change where new species are introduced to the province. How it known that scientists from 20 years ago were able t recognize this plant?

## 1.2.2. Replacing transects/plots

The protocol is silent on when and how to replace transect and/or plots. For example, is a plot is “damaged”, it should be simply dropped from further measuring for all subsequent years. However, if more than ½ of the plots are lost on a transect, then the entire transect should like be replaced. New transects should have a unique transect number (i.e. do not reuse transect labels). The old and new transect should be run in parallel for at least one season to calibrate the new transect with the existing transects.

## 1.2.3 Measuring % cover.

The protocol requires the actual % cover to be measured for each species, rather than simply classifying it into a category (e.g. between 20% and 30% cover). This is the preferred method. However, very rare species are assigned a categorical cover value (.5, .2 or .1 %). Given the rareity of these species, this should be fine.

### 1.2.4 Species recorded at Genus or Family or higher level.

This may be a bit problematic when species richness is recorded as then a species may be double counted. Some editing of the data may be required before analyzing mean (plot) species richness. Presumably, the similarity matrix (*Z*) will indicate a high functional similarity between these general reading and distinct species so that the effective number of species diversity measures will be unaffected.

## 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 transects available for this year. If a transect not visited, this is indicated by missing records in the *General Survey* worksheet.

The *General Survey* sheet contains the information collected at the plot level. There are multiple lines per plot. If a species was not seen in a plot (% cover = 0), the records is not present in the database.

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

* *Transect Label.* The transect measured*.*
* *Plot.* The plot within the transect.
* *Date*. The date the data was collected. The *Year* is extracted from this date.
* *Species*. What species were seen
* *Foliar Cover*. The percent cover of each species in the plot.

# 3. Sample Analyses.

A sample analysis is presented on the *Babine Mountain Park* study area. Data is available only for 2013 because data are collected at 4-year intervals. Data for 2017 and 2021 were simulated based on the 2013 data. The results below are not to be taken as actual results.

This design has 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. 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 Total Cover.

The data is first summarized to the plot level by summing the % cover, excluding non-plant coverage (e.g. rocks). Then the average total % cover for each transect is computed by averaging the total % cover over the plots within each transect. This makes an implicit assumption that all transects will have the same set of plots measured over time. If some plots have to be discarded (e.g. due to damage), the analysis will only be approximate, but unless the number of discarded plots is very large (e.g. more than a 1/3), the approximation should be adequate. This makes the unit of the analysis, the transect within each year.

A summary plot of the total % cover on each plot is shown in Figure 1.

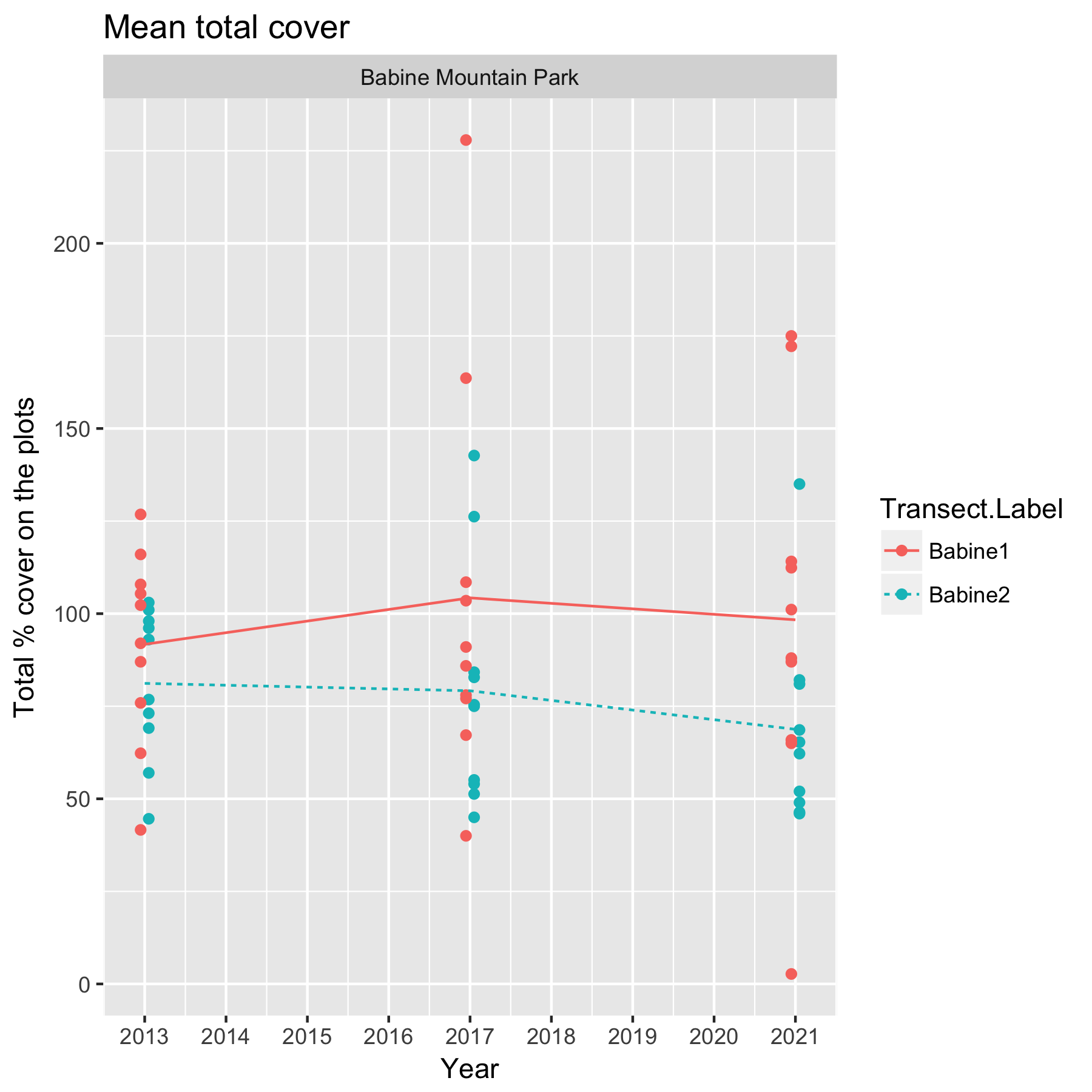


Figure 1. Summary plot of the total % cover. Each dot represents a plot on the transect.

Notice that there is evidence of a transect effect, where, for example, the total % cover on transect *Babine 1* is higher than at the other transect because of local transect-specific conditions (e.g. better soil).

A linear mixed model regression can be used to look for changes over time using the model



where *TotalCover* is the mean (over plots) total % cover for a transect in a year; *TransectF* represents the (random) transect effect; *YearF* represents the year-specific effects (process error); and *Year* represents the calendar year trend over time. The *TransectF* term allows for the fact that transect-specific conditions may tend to affect the total % cover consistently over time. The *YearF* term allows for the consistent effect of year-specific factor. It is expected that the year-specific effects will be small because of the long-lived nature of the plants under observation.

The above model can be fit using the *lmer()* 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=0.77) of a trend with an estimated slope of -0.36 (SE 0.77) /year in the mean total % cover.

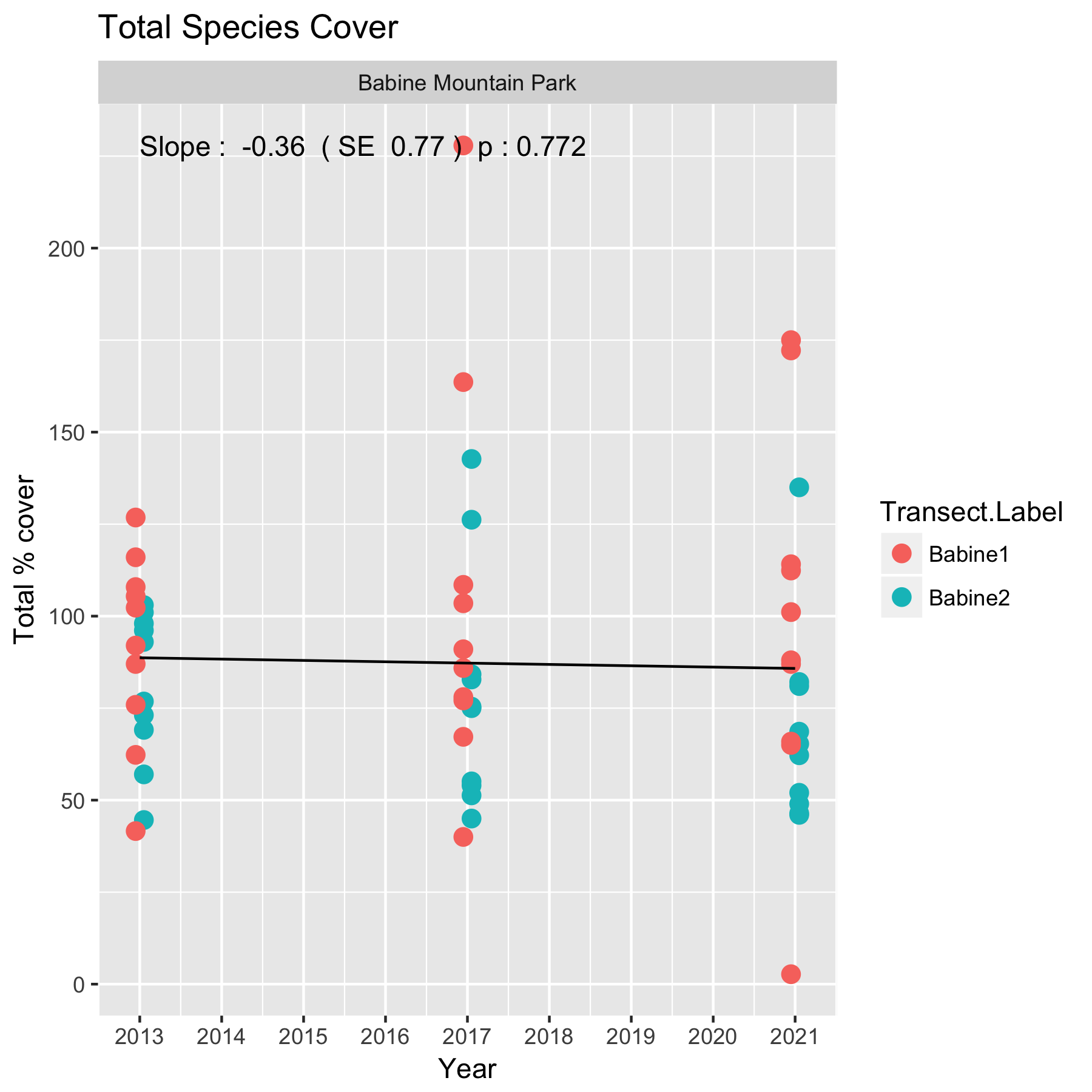


Figure 2. Summary plot of the trend in total % cover at *Babine Mountain Park*.

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

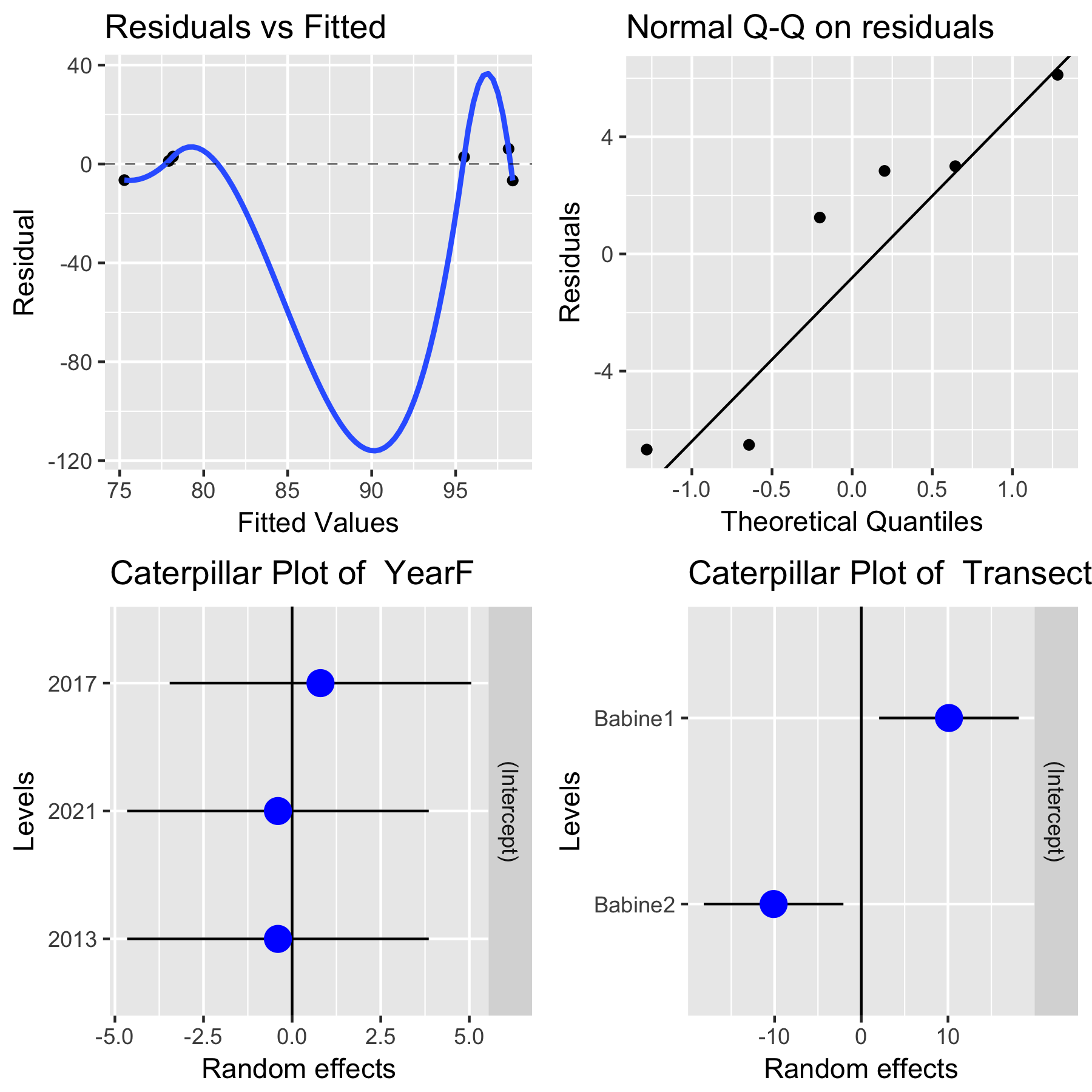


Figure 3. A sample diagnostic plot for the analysis of total % cover at *Babine Mountain Park.*

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 examines the distribution of the year-specific effects. As expected the effects are small (all of the dots are close to 0). The bottom right plot examine looks at the distribution of the random effect of transect. With only two transects the plot is not very informative, but the fact that the 95% confidence interval for each random effect does not cover 0 indicates evidence of a transect effect (this is not surprising).

It will also be possible to covariates such as soil condition in a transect 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 over time (not shown).

The analysis of the % cover for an individual species would similar steps as shown above. However, for species with smaller % cover, it is advisable to transform the observed % cover to an empirical logit(% cover)



to avoid the regression line from going below 0 or above 100%, following the recommendation of Warton and Hui (2011). You will also have to impute 0 values if the species is not listed in a plot. The *R* code is easy to modify.

## 3.2 Species Richness at the plot level.

A similar analysis can be applied to the average species richness at the plot level as in the previous section. The data is first summarized to the plot level by counting the number of plant species, excluding non-plant coverage (e.g. rocks). Some case is needed when plants are only identified to the Genus or Family level – these may have to be removed to avoid double counting. Then the average species richness (at the plot level) for each transect is computed by averaging the species richness over the plots within each transect. This makes an implicit assumption that all transects will have the same set of plots measured over time so that the transect mean richness has roughly the same precision in all transects. If some plots have to be discarded (e.g. due to damage), the analysis will only be approximate, but unless the number of discarded plots is very large (e.g. more than a 1/3), the approximation should be adequate. This makes the unit of the analysis, the transect within each year.

Here is a summary plot of the (reduced) data is shown in Figure 4.

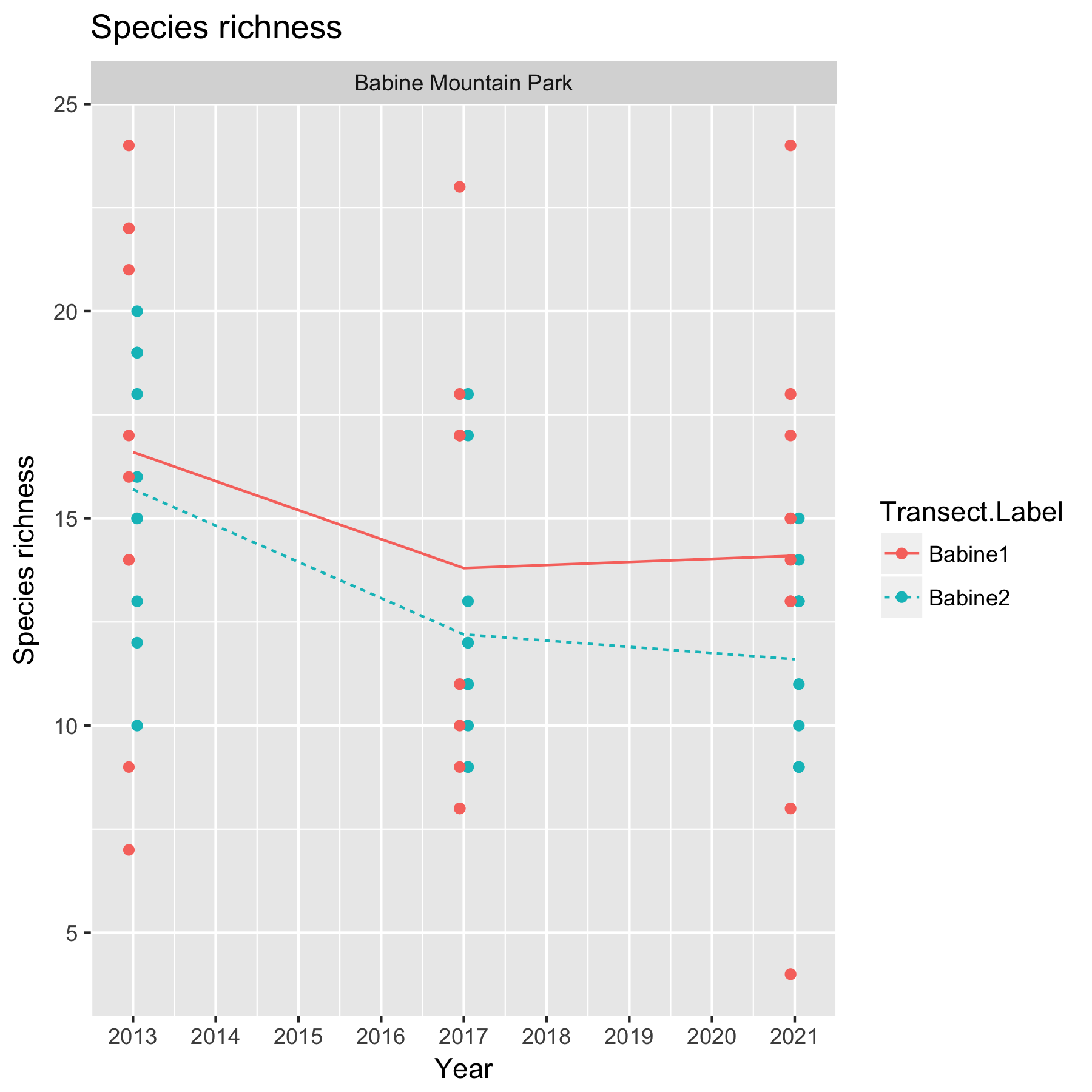


Figure 4. Summary plot of the data for mean species richness at the plot level. Each dot represents a plot on the transects.

Notice that there is a definite transect effect, where, for example, the plot-level species richness on transect *Babine 1* is higher than at the other transect because of local transect-specific conditions (e.g. better soil).

A linear mixed model regression can be used to look for changes over time using the model



where *Richness* is the mean (over plots) richness for a transect in a year; *Transect* represents the (random) transect effect; and *Year* represents the calendar year trend over time. The *TransectF* term allows for the fact that transect-specific conditions may tend to affect the species richness consistently over time. The *YearF* term allows for the effect of process error (year-specific effects) on mean species richness. The year-specific effects are expected to be small.

The above model can be fit using the *lmer()* function in *R.* Figure 5 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=0.26) of a trend with an estimated slope of -0.41 (SE 0.26) /year in the mean richness.

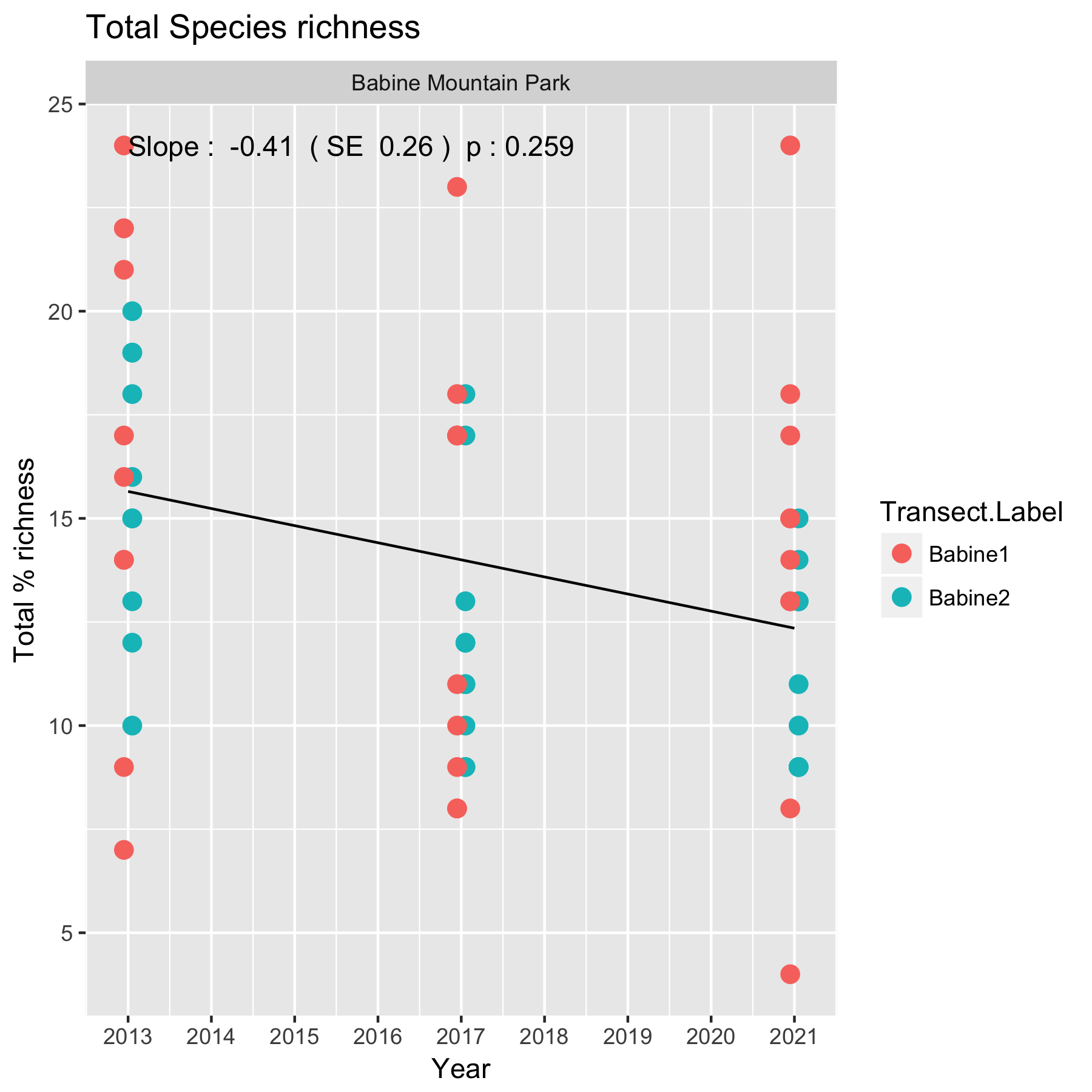


Figure 5. Summary plot of the trend in species richness at *Babine Mountain Park*.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 6.

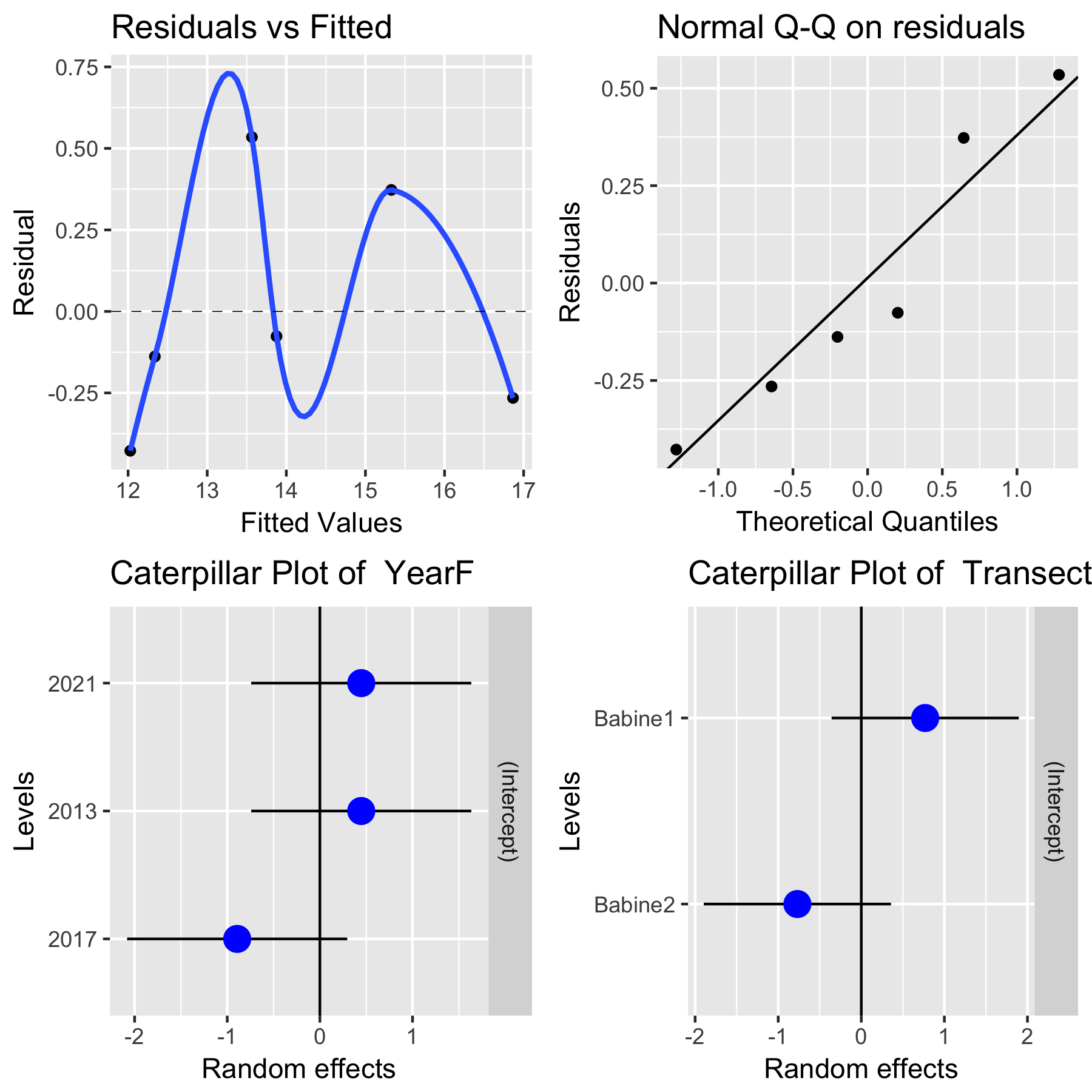


Figure 6. A sample diagnostic plot for the analysis of species richness at *Babine Mountain Park.*

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 look at the distribution of the year-specific effects. The bottom right plot examine looks at the distribution of the random effect of transect. With only two transects the plot is not very informative, but the fact that the 95% confidence interval for each random effect does not cover 0 indicates evidence of a transect effect (this is not surprising).

It will also be possible to covariates such as soil condition in a transect 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.

In this example, species richness in individual plots was over 10 and no transformation will be needed. However, if the species richness values are very small (on the order of 5 or less with many zeros in plots), Poisson regression should be used as shown in other protocols.

A potential weakness of this approach is that richness if computed for each plot separately. So two plots each with completely different set of species would have the same plot richness, but the richness for the entire transect would be much higher. The next section indicates ways to deal with this problem.

## 3.3 Diversity

Species richness is one measure of diversity and there are many more such as Simpson’s or Shannon’s measure of diversity. A similar analysis as for species richness can be done on these other measures of diversity, but are no longer recommended because these measures are often insensitive to big changes in diversity and because they do not account for functional similarity among species. For example, the diversity of six dramatically different species of plants is considered to be no more diverse than six related species.

Leinster and Cobbold (2012) showed how different diversity measures all belong to a continuum (called a diversity profile) and how species similarity can be used to augment the standard diversity measures. The diversity profile also has a number of advantages over the traditional measures of diversity as noted by Chao and Jost (2013) and Jost (2013).

The key difficulty with traditional measures of diversity is that there is non-linear relationship between the diversity measure and biological implications. For this reason, it is often useful to think in terms of “effective number of species” (Jost, 2006; Jost 2013). Briefly, the effective number of species is the number of equally abundant species that give rise to a particular diversity index value.

Jost (2013) provides an example of the problems with traditional diversity measures:

“… suppose you are comparing the diversity of aquatic microorganisms before and after an oil spill. You wouldn't want to measure that diversity by species richness because even a massive toxic event is sure to leave a few vagrant individuals of each pre-spill species, and species richness doesn't distinguish between one individual of Species X or a million; the pre- and post-spill species counts might not be very different, even if the pre- and post-spill species frequencies are very different. So if you are a good traditional biologist you might use the popular Gini-Simpson diversity index, which is  [where  is a measure of relative abundance for species *i*]. Suppose that the pre-spill Gini-Simpson index is .99 and the post-spill index is .97. If you are a good traditional biologist you would figure out that this drop is statistically significant, but you would conclude that the magnitude of the drop is small. You might even say  (very wrongly) that the diversity has dropped by 2%, which sounds like a small drop, nothing to worry about.

The error which virtually all biologists make is that the Gini-Simpson index is not itself a diversity, and is highly nonlinear. The pre-spill community with a Gini-Simpson index of 0.99 has the same diversity as a community of 100 equally-common species. The post-spill community with a Gini-Simpson index of  0.97 has the same diversity as a community of 33 equally-common species. The difference between the pre-and post-spill diversities is in fact enormous. The drop in diversity is 66%, not 2%! This is not just a matter of different definitions of diversity, as some people would like to say. Rather, it is a matter of the indices being nonlinear with respect to our intuitive concept of diversity.”

Leinster and Cobbold (2012) defined the diversity profile as series of “effective numbers” dependent on an index *q* (which ranges from 0 to infinity) and a similarity matrix **Z** (whose (*i,j*) entry measures the similarity of species *i* and species *j* from 0 (not similar) to 1 (completely similar)). The diversity profile is computed as:

 (1)

where the vector **p**is the relative abundance of the species present (i.e. excluding species with 0 abundance), and **Z** is the similarity matrix among the vector of species. As shown by Leinster and Cobbold (2012), many of the common diversity indices are special cases of (1). For example, if *q* = 0*,* (1*)* reduces to species richness; if *q* = 1 and **Z=I**, then (1) is related to the Shannon Index; and if *q=2* and **Z=I**, then (1)is related to the Simpson Index of diversity.

The **Z** matrix (measure of similarity) resolves a number of problems with the common diversity measures. If two species are virtually identical (entries of **Z** close to 1), then the diversity measure (1) effectively treats them equivalently as a single species. The difficulty, is of course, defining this similarity matrix.

I’ve included a briefing note on the analysis of diversity in an environmental impact assessment case with this material as illustration of the method. For the rest of this section, the diversity measures will be computed assuming a diagonal Z matrix, i.e. every species is functionally distinct which is obviously incorrect

Figure 7 shows the diversity profiles for each transect for each year in the study.

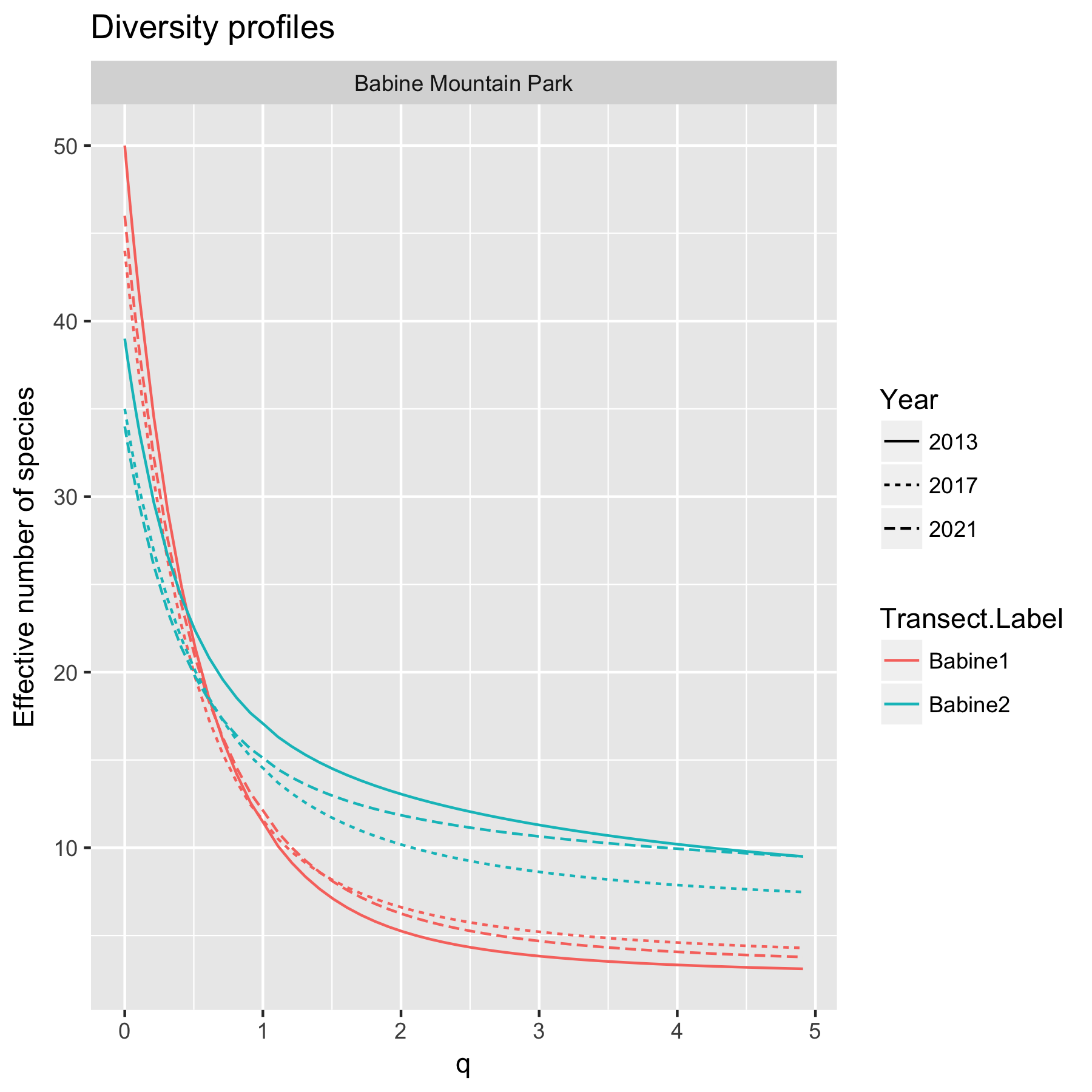


Figure 7. An illustration of the diversity profiles for each transect in each year at *Babine Mountain Park*. The effective number of species when *q = 0* is equivalent to simple species richness; when *q = 1* this is related to the Shannon Index of diversity; when *q = 2* this is related to the Simpson Index of diversity. As values of *q* increase more “weight” is given to common species. As value of *q* move towards 0, more “weight” is given to rare species.

Once the diversity profiles have been computed, the effective number of species for any value of *q* can be extracted and analyzed in the same way as the analysis for the average species richness.

For example, here are the extracted diversity values for *q = 2* which is related to the Simpson Index of diversity:

Study.Area.Name Year Transect.Label q diversity

Babine Mountain Park 2013 Babine1 2 5.3

Babine Mountain Park 2013 Babine2 2 13.1

Babine Mountain Park 2017 Babine1 2 6.6

Babine Mountain Park 2017 Babine2 2 10.2

Babine Mountain Park 2021 Babine1 2 6.3

Babine Mountain Park 2021 Babine2 2 11.9

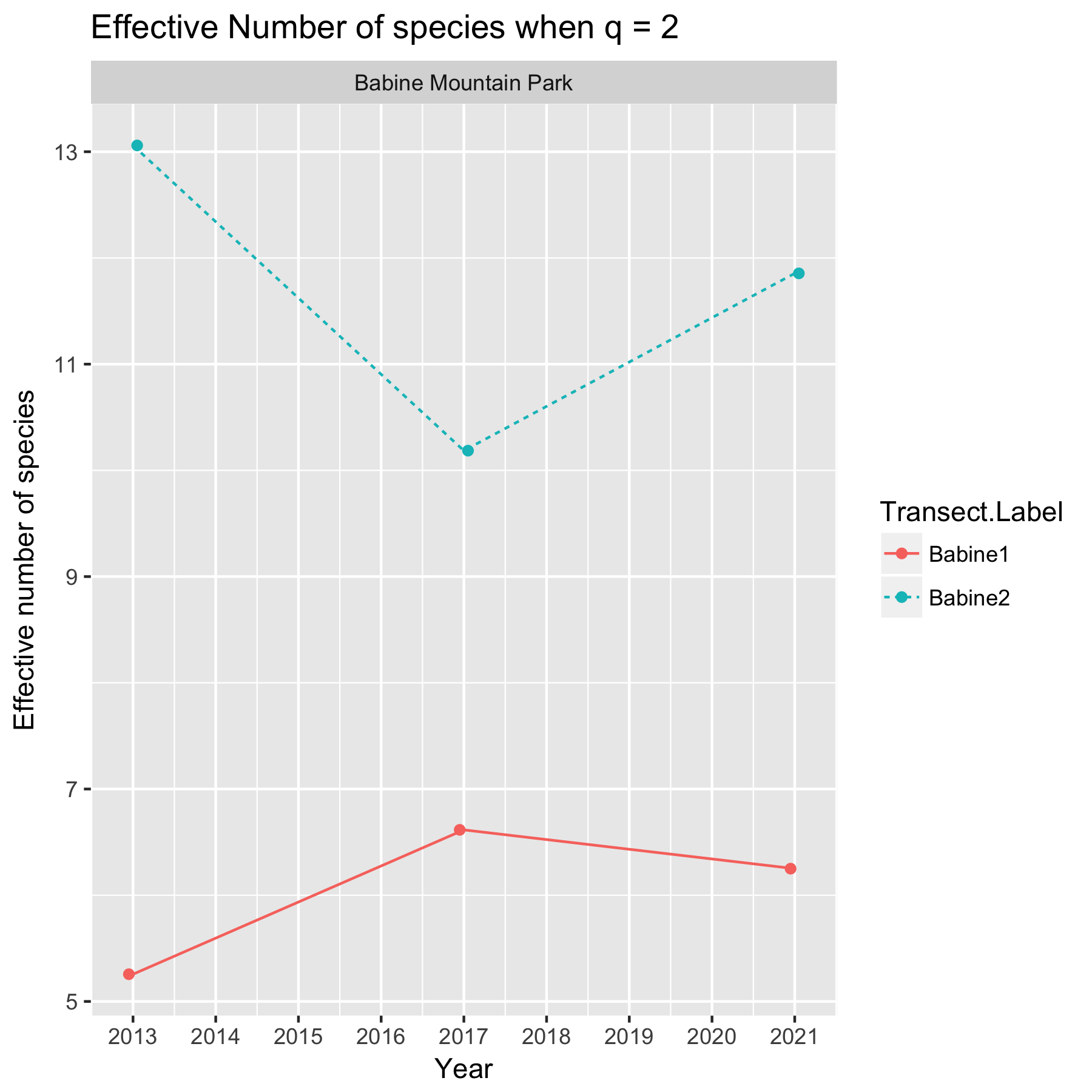
A summary plot of the effective number of species over time: is shown in Figure 8.

Figure 8. Summary plot of the effective number of species for *q = 2*.

Notice that there is a definite transect effect, where, for example, the effective number of species on transect *Babine 1* is lower than at the other transect because of local transect-specific conditions (e.g. better soil).

A linear mixed model regression can be used to look for changes over time using the model



where *Effective#Species* is the effective number of species for a transect in a year; *TransectF* represents the (random) transect effect; *YearF* represents the year-specific effects; and *Year* represents the calendar year trend over time. The *TransectF* term allows for the fact that transect-specific conditions may tend to affect the species richness consistently over time. The *YearF* term allows for the impact of year-specific effects that influence all of the points wihin a year simultaneously.

The above model can be fit using the *lmer()* function in *R.* Figure 9 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=0.94) of a trend with an estimated slope of -0.01 (SE 0.94) /year in the effective number of species.

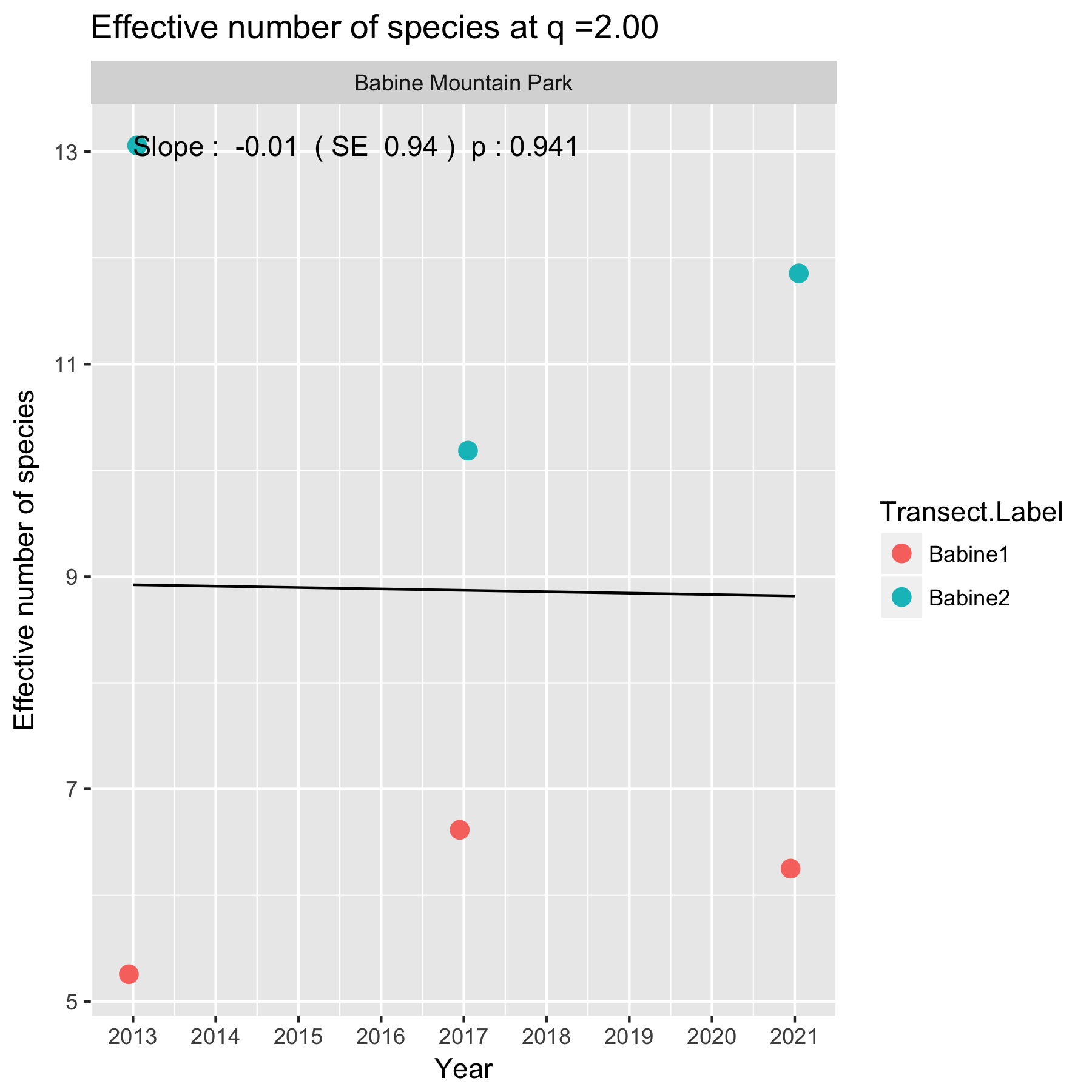


Figure 9. Summary plot of the trend in effective number of species when *q = 2* at *Babine Mountain Park*.

Following the fit, the diagnostic plots should be examined in the usual way (not shown). There was also no evidence of autocorrelation.

It will also be possible to covariates such as soil condition in a transect to try and explain some of the variation over time using a multiple regression. With only three years of data available, this not sensible.

In this example, the effect number of species we all well bound from zero so no transformations are needed. However, if the effective number of species values are very small (on the order of 5), Poisson regression should be used as shown in other protocols.

# 3. Summary

There are several measures that can be monitored over time. Among the simplest are simple averages of total % cover or plot-level species richness.

Species richness is one of many measures of diversity. Recent work has places it and other measures on a continuum where the effective number of species (also called the Hill number) is a more sensible measure of diversity. This can account for both the relative abundance of species and also their functional similarity. The latter (the *Z* matrix above) needs to be developed to effectively use these newer methods for diversity.

Beta diversity and ordination methods are usually used when community structures are measured across a gradient (e.g. environmental gradient) and the relationship between the community structure and the environmental variables are of interest. This does not appears to be of of interest here and are not pursued.

References

Chao, A. and Jost, L. (2013). Diversity Analysis. Chapman and Hall.

Jost, L. (2013). Effective number of species. Available at <http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/EffectiveNumberOfSpecies.htm>. Downloaded 2013-12-10.

Leinster, T. and Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. Ecology, 93, 477-489.

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/.

Warton, D.I. and Hui, F.K.C. (2010) The Arcsine Is Asinine: The Analysis of Proportions in Ecology. Ecology, 92, 3-10.

Appendix A.

Issues encountered when doing a trial analysis on the *Babine Mountain Park* study area data.

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

(a) Date formatting. I suggest you always use yyyy-mm-dd as the format for ALL protocols. This set of workbooks currently use *dd mmmm yy* format (with 4 character month names, 2 digit year, and blanks separating the sub-fields.).

(b) Need separate field for plot code

In the Babine worksheet, the plot code is in the comment field. In some cases, the comment field also contains other information:

Plot1; cup ends

Plot2; cover of rock 10%