Review of the AJAX Project Environmental Assessment Certificate Application

Carl James Schwarz
Statistics and Actuarial Science
Simon Fraser University
cschwarz@stat.sfu.ca

2016-04-05

This is a review of the AJAX Project Environment Assessment Certificate Application. The Application documents were download from the BC Government website on 2016-01-18.

This review will focus on the statistical aspects of the document – in particular, the adequacy of the plan to detect important effects and the statistical methods both used and proposed as part of the plan.

1. Purpose of monitoring plans

Monitoring plans need to respond to two objectives, somewhat at odds with each other. First, the program needs to be able to detect large, unforeseen, short-term consequences ("disaster" detection). Second, the monitoring program must also be designed to detect long-term smaller, cumulative effects ("trend" detection). For example, a program may wish to detect a large change in a water quality (WQ) parameter (e.g. an increase of 20 percentage points in the mean) in any year when impacted sites are compared to reference sites (disaster detection). The program may also wish to detect a cumulative 2 percentage points/year in the difference in WQ parameter between reference and control sites (trend detection).

The amount of sampling in each year and the number of years of sampling before and after project startup will depend upon the method used to detect effects (e.g. ANOVA or regression), the noise in the data (e.g. natural variability in water quality both within and among years in the absence of an effect), and on the biological effects of interest. The firs two are relatively easy to know or to estimate from pre-impact data. However, without specification of the biological effects, it is impossible to know if a monitoring program will be effective. **This is a key deficiency of this proposed plan.**

For example, it impossible to know if five years of WQ data collected using monthly sampling and 5-in-30 days samples are sufficient to detect the short-term effects without knowing the size of the short-term effects that is biologically important.

Conversely, in the absence of knowledge of the biological effects, the size of the biological effect that can be detected should be determined to see if the detectable effect provides some assurance against disaster and long-term trends. For example, how much of a difference can be detected with a monitoring plan with 5 years of monthly sampling paired in reference and control sites in a BACI (before-after-control-impact)? Is this detectable difference adequate for "disaster" or "trend" protection?

The proposal is generally weak in these aspects. There is little discussion of what changes are biologically important and if the proposed levels of sampling are adequate to detect such changes. The proponent needs to spend considerably more time in delineating these values prior to sampling – it will be too late if an analysis of the data reveals that the power to detect changes was inadequate. Conversely, the proposed plan needs to estimate what size of effect can be detected so that stakeholders can assess if this is adequate.

There are cases where this issue is obliquely referred to. For example, Table 11.2-2 shows trigger values for changes in WQ parameters that are important to detect. Under the table it states

> *"To determine if guideline requirements are met for short-term (acute) exposures, hourly samples taken over a 24-hour period will be used to monitor changes in turbidity. For long-term (chronic) exposures, daily samples taken over a 30-day period will be used (BC MOE 2015).*

So, hourly samples are specified in a bid to determine if acute exposures are being below guidelines but there is no indication if, in fact, these hourly samples are adequate or merely wishful thinking.

In Chapter 11 (p.338) it states:

> *5. Continue to monitor vegetation metal uptake at established locations (reclamation, baseline, and control sites) every five years. Ensure enough samples are taken that statistically significant conclusions can be made. "*

So how many samples will be needed? No information is presented here if the sampling every five years is sufficient, nor the number of samples needed. Does the report intend to determine this number in the future? The phrase *"... statistically significant conclusions can be made*" is also scientifically meaningless.

In some cases, the effect sizes appear to be far too large. For example in Appendix 6-7-B (p. 33)

> "*The Metal Mining Technical Guidance for Environmental Effects Monitoring (Environment Canada 2012) recommends that the benthic invertebrate community survey should have sufficient statistical power to detect a critical effect size of plus or minus two standard deviations, which in most cases results in a minimum sample size of five."*

A two standard deviation shift in the mean is HUGE – it essentially says that there is NO overlap in the data between before and after the project starts. That is the reason why so few samples are needed. The required sample size for a 10% shift in the mean which may be biologically meaningful is MUCH larger!

Determining the proper amount of sampling is not a trivial task. In many designs there are multiple levels of sampling and "sample size" is a complex combination of the amount of sampling at each of the levels. Determining the biologically important effect size is an enormous, non-statistical problem that needs to be resolved so that an appropriate plan can be determined.

2. Comments about specific plans.

The document has many monitoring plans. In the following sections, I've highlighted some areas of the plans where additional work may be needed.

2.1 Invasive Plant Monitoring.

The Invasive Plant Monitoring is described in Section 11.17.4:

> "In general, monitoring will be on an observational basis with different parts of the mine covered each year, with the entire mine covered over two years. A trained observer will walk high-risk areas based on an "intuitive" sampling method focused on areas of greater risk of harboring invasive species"

In general, non-probabilistic designs should be avoided. The random selection (i.e. based on a probabilistic selection) is what "guarantees" that estimates are unbiased. Sample size does not control representation, but only control the precision of the results. This is a non-probabilistic monitoring design, but the proponents claim that all of the mine area will be covered every two years. However, "high risk areas" can easily be incorporated into a probabilistic plan with different probability of selection depending on the risk, i.e. higher risk areas have a higher probability of being sampled.

2.2 Surface/Ground Water Monitoring Plans

The proponents indicate (Section 11.23.4.1):

> "General surface water quality monitoring will be implemented with a focus on assessing for changes in surface water quality compared to baseline and predicted concentrations, generic water quality guidelines, and site-specific water quality benchmarks. "

However, no information is presented on the size of the effects that will be able to detected. Are these effects are expected to vary seasonally? Nor is information presented on the type of effects that will be examined, i.e. changes in the mean, changes in the frequency of high values, or changes in the 95th percentile, etc. Each of the endpoints require a different level of sampling effort and possibly different sampling plans.

The proponents indicate that (Section 11.23.4.2):

> "The monitoring program will include appropriate reference sites to control for natural variation and regional trends. The monitoring program will also include on-site water quality monitoring in the water management ponds and the TSF.
>
> The monitoring program will rely on baseline data presented in the Application/EIS and any additional data collected prior to the onset of Construction activities for the established reference sites and the downstream monitoring sites. These data will be used for comparison to receiving water data collected over the course of the mine development. Some of the parameters will require thresholds to be met, above which additional monitoring and/or mitigation strategies may be triggered."

No information is presented on the size of the effects that will be able to detected with the current planned level of sampling nor on how the data will be analyzed. The sampling plan has two levels of sampling – sites and samples within sites in different seasons – and simple summary measures (e.g. 95th percentile) and simple statistical tests (e.g. t-tests) are unlikely to be appropriate.

Similar concerns arise about the ground water sampling plan (Section 11.24)

Appendix 6.3B outlines the development of site specific sulphate guidelines for water quality using standard toxicity testing on *C. dubia*, *P. subcapitata*, and embryo *O. mykiss*. However, no precision was presented on the final results (e.g. a confidence interval on the LC/EC20) and given the small sample sizes, the uncertainty in the results may be quite large. Consequently, the usefulness of the reported guideline may be questionable. All estimates should be accompanied by measures of uncertainty such as standard errors and confidence intervals.

Appendix 6.3D outlines estimates of protective water quality benchmarks (WQBs). There was insufficient data to meet the minimum toxicity datasets required for Type A WQBs; and so a threshold based on the lowest observed toxicity modified by a safety factor was used. No measure of precision is presented (nor is it clear how such a measure would be computed).

2.3 Fisheries and Aquatic Life Monitoring Plan

Generally few details are provided. However, the proponent states (Section 11.25.4):

> "The AEM program will utilize a before-after-control-impact (BACI) comparison of environmental indicators (i.e., water quality, fish population, benthic invertebrate community, and fish tissue), consistent with the Metal Mining Technical Guidance for Environmental Effects Monitoring (Environment Canada 2012b). The BACI analyses will aim to incorporate the required statistical assumptions and provide transparent and reproducible results. Statistical hypothesis testing will be validated through a posteriori power analysis."

The use of a BACI analysis is appropriate, but no information is presented on the ability of the design to detect biologically important differences. The proponents plan to do a retrospective power analysis. This is not recommended. The proper place for a power analysis is PRIOR to the project starting (Gerard et al, 1998). Furthermore retrospective power analyses have serious flaws and provide little useful information (Gerard et al. 1998). After the data are collected it is too late to fix the plan if the power analysis shows that insufficient data has been collected. The proper place for a power analysis is before the study begins. I could find no evidence that such an analysis took place.

Appendix 6.7B presents some details on the fish and aquatic baseline sampling. The proponents propose to (Section 2.2, Appendix 6.7B)

> "The study design recommended for sediment and tissue residue monitoring is termed the Spatial Variance Program; the Reference Condition Approach is recommend for the benthic invertebrate sampling program, while the "before, after, control, impact (BACI) experimental design is recommended for the fisheries program …"

In the Spatial Variance Program, replicate sediment samples (typically 5 composite samples) are taken from each site once per year (usually in late summer or fall during the low flow period) (Section 2.31, Appendix 6.7B). Parameters such as metal concentrations are measured on each sample. It is rather curious that values below detection limits are

excluded (Section 2.3.3.3) as this exclusion will drive up the mean and make it harder to detect effects. There are good statistical methods for dealing with below detection limit values (see later in text). The sediment values are then used to compare means among sites and across years but no details on the analysis are presented. Much more thought is needed on how this data will be used to assess potential environmental impacts of the mine. Will changes in the means over time (a regression approach) be used? Will simple before/after analyses for changes in the means be used? How big of an impact can be detected with the current sampling plan? All of these details are lacking.

There is some discrepancy in the type of plan envisioned for the benthic invertebrate sampling program. In section 2.2 (Appendix 6.7B), it was indicated that the Reference Condition Approach will be used, but in Section 2.4 (Appendix 6.7B), it is indicated that

> "A BACI study design was selected for each aquatic life parameters"

Two samples (surface and bottom) are taken each year at each site for phytoplankton. Five replicate samples are taken for periphyton at each site where each sample is a composite. Three sample for zooplankton are selected from each site in each year. Five samples are taken for benthic invertebrate sampling.

These sample sizes appear to be determined Appendix 6.7B (page 33) using:

> "*The Metal Mining Technical Guidance for Environmental Effects Monitoring (Environment Canada 2012) recommends that the benthic invertebrate community survey should have sufficient statistical power to detect a critical effect size of plus or minus two standard deviations, which in most cases results in a minimum sample size of five."*

A two standard deviation shift in the mean is HUGE – it essentially says that there is NO overlap in the data between before and after the project starts. That is the reason why so few samples are needed. The required sample size for a 10% shift in the mean is MUCH larger! Much more justification is needed on why only such a large change is to be detected – this likely corresponds to a disaster detection scenario as noted earlier. Much more sampling (both in the number of years and samples/year) will be needed for detecting smaller trends over time.

Regardless of the Environment Canada 2012 guidelines, the existing data should be used in a prospective power analysis to determine the effect size that can be detected with the current sampling plan and if this effect size is biologically reasonable.

Tissue sampling of Rainbow Trout will follow the Spatial Variance Program with multiple samples per year, but no rationale is given on the sample sizes chosen.

Standard diversity measures applied to community structures are used in the baseline report, but there are serious problems in using the standard diversity measures in simple ways (see later in this document).

Fish condition factors (Section 2.5.3, Appendix 6.7B) are computed for each site. It is unclear how these will be used to assess potential impact of streams (ANCOVA?). What size of effect can be detected with the current plan? Is this adequate?

2.4 Wildlife and Vegetation Monitoring Plan.

The proponents identify key objectives for this plan (Section 11.27.1) including:

> "…ensure monitoring efforts are able to detect natural and Project-related changes to the environment and wildlife;"

but there is no discussion of how big of an effect can be detected with the current plan, nor how big of an effect is biologically important (see earlier comments). The proponents indicated that (Section 11.27.1.1)

> "Finalization of the monitoring program will occur through consultation with Federal and Provincial government agencies, Aboriginal groups, the public, and other stakeholders."

The biologically important effect sizes should be determined in advance of any such meetings so that the proper sized plan can be developed.

Section 11.27.3.3 has a brief description of the heavy metals monitoring plan. Basically samples will be taken from impact and baseline sites and this section also indicates:

> "5. …Ensure enough sample are taken that statistically significant conclusions can be made.
>
> "6. …Data analysis used the student's t-test to assess if differences between the individual species samples and baseline individual samples were statistically different (using a 95% confidence interval)"

The proponents recognize that some planning is needed to ensure that sampling is adequate but is confusing statistical significance with biological importance. Statistical significance says nothing if the effect size is biologically important or not; conversely, failing to detect evidence of a difference in the mean (not statistically significant) does not imply no difference. A proper power analysis/sample planning exercise is needed.

The proponents claim that a Student's t-test will be used, but this will be an inappropriate analysis due to the hierarchical nature of the sampling plan (multiple sites in impact and control; multiple samples in each site; multiple years monitoring over time). One of the key issues is that there may differences in the response among sites due to natural factors and what is important if these differences are changing over time. A variant of a BACI analysis will be needed.

Section 11.27.4 discusses sampling for wildlife and concludes that

> ."…data from monitoring programs will be analyzed using best practices and assessed for statistical power to detect changes in wildlife populations or habitat availability;"

This is a good step, but no further details are provided.

Appendix 3-H has additional details on the baseline vegetation data collection. Basically, seven transects were established on seven different soil management unit, with one transect per soil management unit. Multiple sample plots were measured along each transect were taken as described on page 6 of the appendix. Unfortunately, the proponents have confused experimental units and observations units (i.e. have collected

pseudo-replicates at each transect, see Hurlbert, 1984). The transect is the experimental unit for each soil type and unfortunately, there is only replicate transect for each soil type. Consequently, inference is limited to determining if the mean response at THAT PARTICULAR transect is similar to the mine and cannot be generalized. Similarly, the variation measures presented in the Figures in the report, are within-transect measures and not at the correct level of analysis (multiple transects in each soil type are needed).

What is needed are at least two transects in each soil type collected at two separate locations to obtain a measure of site-to-site variation which is the applicable unit of analysis. The proponents recommend that at least 5 reference sites be established (page 12, Appendix 3-H) but give no rationale for the number.

The information presented in Table 4.0 (page 5, Appendix 3-H) is of limited usefulness. Diversity measures presented in Figure 2 (page 6, Appendix 3-H) also need to be treated with extreme caution (see elsewhere in report on problem in using standard diversity measures). In particular the comparison of diversity measures (Section 4.1.2, Appendix 3-H) is extremely difficult to interpret (again refer to problems in using standard diversity measures elsewhere).

3. General Issues.

3.1 Use of 95th percentiles.

Many of the WQ parameters (and other parameters) appear to use a 95th percentile as "normal range". For example, refer to Table 6.3-3. Here a large number of statistics are report on baseline monitoring. By definition, the 95th percentile indicates that about 95% of values are less than this point and about 5% are above this point.

I have a few concerns about the use of percentiles.

- Many of the percentiles are based on a very small number of data points (e.g. PC-EF-04 has 3 data points- yet a 95th percentile is being reported corresponding to values that are to be exceeded less than 1 time in 20? How is that possible? Such values are likely to be nonsensical.
- The data is collected using a mixture of monthly samples and 5-in-30 days samples. Simple percentiles on the pooled date are not valid estimates of the percentiles because the sample values represent different number of days in the population. For example, a monthly sample represents 30 days of the month in which it was taken; a 5-in-30 sample would represent 7 days for the week it was sampled. When these are pooled over a year, it is not valid to compute a 95th percentile on these pooled values.
- There is no associated measures of precision. Confidence intervals when applied to percentiles are called tolerance intervals. These should be computed and this would allow you to way that you are "90% confident that no more than 5% of future samples will fall above this point". Alternatively, the methods of Kilgour (1988) could be used to indicate if samples are in or outside of normal range. In both bases, the actual uncertainty is so large that I suspect that many of the data series in these tables are insufficient for future monitoring except to detect

anything but disasters.

Using the 95th percentile as a trigger point is also very insensitive to shifts in the mean. There could be a shift of 10% in the mean (i.e. the mean SQ parameter changes by 10%), but still only a few values would exceed the 95th percentile and so the plan would have limited power to detect this shift.

It appears that exceedances of the 95th percentiles or the MOE WQ guidelines will be one set of triggers. However, in Table 6.3-6 some of these exceedance values do not seem sensible. For example, look at Petersen Creek at location PC-EF04. There are only 4 sample points, yet the table reports that 6% of values exceed the Turbididty guidelines? How is this possible – with only 4 samples, the only possible values are 25% (1/4), 50% (2/4), 75% (3/4) or 100% (4/4). There are many other similar problems in the reported tables.

3.2 Computation and Analysis of Diversity measures

Diversity measures are computed for vegetation, plankton, and invertebrate communities. The standard diversity measures are a poor choice for monitoring and likely will NOT be useful in detecting changes. The problem is best explained by Jost (2016)[1] who provides an example of the problems with traditional diversity measures:

> *"…Suppose you want to compare the epiphyte diversity of a primary forest to the epiphyte diversity of a disturbed forest. You take big samples of each of these two communities.*
>
> *Now you want to answer the question "Are the diversities different?"*
>
> *There are two parts to the answer. We must measure the **magnitude** of the difference, which enables us to judge its biological importance, and we must measure the statistical significance of the difference, to see whether it is arose by simple sampling variability or because of a real difference in diversity. These are completely different kinds of questions. Many published studies apply some statistical test and find statistically significant results, and then go no further. But mere statistical significance is not in itself interesting. Any tiny difference can be made statistically significant if sample size is large enough. Given that the difference is statistically significant, the more interesting question is: How big is the difference?*
>
> *Biologists have to get past their fixation on statistical significance and concern themselves more with this question of the size of the difference. The use of raw diversity indices made it difficult to answer that question in the past. If the Gini-Simpson indices of the communities are .99 and .97, is that an important difference or a small one? If the Shannon entropies (Shannon-Wiener indices) of the two communities are 4.5 and 4.1, is that a big difference or a little one?*

---

1 http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/How%20to%20compare%20the%20diversities%20of%20two%20communities.htm

It is hard to say based on these numbers. At first glance the difference in Gini-Simpson indices of .99 and .97 looks small. But this index is highly nonlinear. Converting to effective number of species (which is the true diversity, as explained in the other parts of this site) proves that the difference is in fact huge: the community with a Gini-Simposn index of 0.99 has the same diversity as a community with 100 equally-common species, while the community with a Gini-Simposn index of 0.97 has the same diversity as a community with 33 equally-common species. The difference between a community with 33 equally common species and one with 100 equally common species is enormous. The same holds for the Shannon entropy values just mentioned: 4.5 converts to 90 effective species while 4.1 converts to 60 effective species. The second community is only 2/3 as diverse as the first community, according to this measure. If the diversity drops that much between undisturbed forest and disturbed forest, that is a serious and biologically significant drop. It is hard to realize that the drop is so dramatic if one looks only at the raw indices.

Suppose you calculate the true diversities (effective numbers of species) and find a big drop. That is when it is time to ask the other question, the question of statistical significance: could such a drop be due to random sampling effects? If the samples are large, the t-test suggested by Hutcheson (1970) can answer this question for diversity of order 1 (Shannon measures). Randomization tests are also available. On this website I do not deal with this side of the question because the statistical tests in the literature are mostly correct. Where the literature fails is in its treatment (or its lack of treatment) of the other (and more important) side of the question, the magnitude or biological significance of the difference.

*So, to compare the magnitudes of two diversities, calculate the effective numbers of species (the exponential of the Shannon entropy, for example) of the two communities so that you can compare them on a linear scale and get an intuitive feel for the difference. (Download my Excel sheet **Indices to Diversities** which does this conversion for you.)You can divide the smaller diversity by the larger one and come up with a meaningful fractional drop in diversity (something you can't do with raw diversity indices because they are nonlinear with increasing diversity). Then check to see if that drop could be due to chance, by calculating the t-test of Hutcheson (for Shannon measures) or other tests. If the difference in true diversities (effective numbers of species) is both large in magnitude and statistically significant, then you have found something important. Congratulations!*

The proponents should use more modern methods for monitoring diversity (e.g. Leinster, and Cobbold 2012). More details are available on Jost's website and references cited within.

3.3 Values below detection limit.

For some parameters, values will be below detection limits. The report is unclear how these will be handed. However, in some cases (Section 2.3.3.3 or Appendix 6.7B, p.29)

> "*Data are summarized by site and sampling date … Only values above the detection limits were included in the statistical summaries*".

Excluding values below detection limits is incorrect. For example, suppose that the baseline study collects 100 samples of which 99 are below the detection limit of .001 and the last value has the value of 5. If you exclude the values below the detection limits, then

the reported mean concentration would be 5, which is not an appropriate characterization of the parameter. There are many methods of dealing with values below detection limits (technically known as censored data) such as Helsel (2005).[2] This report needs to adopt current practices in dealing with censored data.

3.4 Return Periods

Flood forecasts and other extreme event predictions rely on 1-in-100 year return periods. These have been computed using historical data. Are these return periods relevant in the face of climate change? In some cases, Appendix 6.1-A section the proponents did try and account for this by looking to see if the maximum rainfall events are increasing in severity over time. A similar type of analysis needs to be done whenever the document uses 1-in-100 or 1-in-200 year events. I didn't see any sensitive analysis to see if these adjustments will provide sufficient protection for future events.

4. Summary and Recommendations

A monitoring plan with low power to detect biologically important changes is worse than doing no monitoring. A low-power design gives a mistaken impression that deleterious effects have not occurred, when in fact, your plan simply failed to detect them.

In order to evaluate the effectiveness of a monitoring program, it is necessary to know the size of effect that is biologically important or conversely the size of the biological effect that can be detected given a particular plan. **The current document is severely lacking in discussing either the biologically important effect sizes that need to be detected, or conversely, what size of effect can be detected with the current plan.**

The document is also extremely vague on how the data will be analyzed. In some cases, (e.g. diversity measures, percentiles, comparing WQ over time)**, inappropriate analyses are performed or proposed to be used**. Far too often, important defects in the plan are unrecognized until after the data are collected and a statistical analysis is started. Then it is too late to go back and correct the plan. A comprehensive data analysis plan, usually conducted on "fake-but-realistic" data is needed to ensure that the analysis will be feasible.

The proposed monitoring plan can be improved by:

(1) Providing biological effects sizes for both short-term (disaster protection) and long-term (cumulative effects) need to be developed.
(2) Based on the biological effect sizes, determine if the current sampling protocols have adequate power to detect these effects (if they exist). If the protocols do not have sufficient power, how much data must be collected? Conversely, what size of effects can be detected given the current monitoring program?

(3) Creating a data-analysis plan. Mock data sets can be constructed to show exactly the form of the data that will be collected. Sample analysis scripts can be run using this mock data to ensure that the output from the procedures provides information that will be useful in evaluating the results of the monitoring plan.

References.

Gerard, P. D., Smith, D. R., and Weerakkody, G. (1998), "Limits of Retrospective Power Analysis," Journal of Wildlife Management, 62, 801–807.

Helsel DR. 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley and Sons, New York.

Jost, L. (2014a). The new synthesis of diversity indices and similarity measures. Available at http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/EffectiveNumberOfSpecies.htm. Accessed 2014-09-06.

Jost, L. (2014b). Comparing the diversities of two communities. Available at http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/How%20to%20compare%20the%20diversities%20of%20two%20communities.htm. Access 2014-09-16.

Leinster, T. and Cobbold., C.A. 2012. Measuring diversity: the importance of species similarity. Ecology 93:477–489. http://dx.doi.org/10.1890/10-2402.1