USING IMPACT SIMULATIONS TO EVALUATE THE POWER OF SKEENA REFERENCE CONDITION APPROACH STREAM BIOASSESSMENTS

By

AARON J. DOWNIE

B.Sc., University of British Columbia, 2001

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE
in
ENVIRONMENT AND MANAGEMENT

We accept this thesis as conforming
to the required standard

.....
Dr. John L. Bailey, Thesis Supervisor

.....
Dr. Anthony Boydell, Director
School of Environment and Sustainability

ROYAL ROADS UNIVERSITY

April 2011

Acknowledgements

First and foremost, I would like to thank my supervisor John for his willingness to take me on as a graduate student without actually meeting me in person until a couple of months before this final submission. His ongoing guidance and advice was instrumental in shaping my study, and this resulting thesis. Thanks also to the B.C. Ministry of Environment, and in particular my manager Ian Sharpe, for supporting me throughout the MEM program. The raw data I used for my analysis was also provided by the Ministry.

On many occasions I relied on Shauna Bennett of Bio Logic Consulting in Terrace for ideas and inspiration. Without her I would probably still be trying to import my data into PRIMER and Systat. On the home stretch, it was Shauna who reminded me that "the only good thesis is a done thesis"; and that was part of my drive to finish.

I also owe a big thank you to Jeremy Roscoe (Ministry of Environment) for producing the site map, and assisting with edits on my draft manuscript along with Greg Tamblyn (also Ministry of Environment) and Shauna Bennett. And of course I'm grateful for my family, friends and classmates who were always there to offer support as needed.

Abstract

Effective use of bioassessments requires an understanding of their performance. This study evaluated the performance of Skeena Reference Condition Approach (RCA) bioassessments by calculating Type I and Type II error rates and power using a data set of artificially-impacted test sites.

Results from this study demonstrated that there are trade-offs between the two error types. Type I error rates – the chance of concluding that a site is impacted when it isn't – were higher than expected based on decision points set in the assessments. Type II error rates – the chance of concluding that a site is not impacted when it is – were often even greater.

To achieve sustainable development, resource managers who use Skeena bioassessments must carefully consider the risks associated with making errors, and may wish to set decision points that result in more Type I errors in order to reduce the likelihood of making costly Type II errors.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Freshwater aquatic ecosystems provide essential functions for life on earth. Healthy lakes, rivers and streams not only act as important ecosystems themselves, but they also sustain terrestrial and marine ecosystems. They treat wastes, cycle nutrients, and store and move water from highlands to lowlands and eventually to the sea (Millennium Ecosystem Assessment, 2005). They are also important for the conservation of biodiversity, through the habitats they provide and the species they support (World Conservation Monitoring Center, 1998). From an anthropogenic perspective, freshwater aquatic ecosystems provide an array of social and economic values in addition to environmental values. Lakes, rivers and streams provide food sources and are often used for drinking water supplies, recreation, irrigation, power generation and industrial use. Humans have relied on freshwater aquatic ecosystems for millennia, and in many places these systems remain an important part of local cultures, lifestyles and economies, in addition to supporting basic human needs.

Aquatic ecosystems are abundant in the northwestern quadrant of British Columbia (B.C.) and are central features on a landscape that is sparsely developed in comparison to the southern portion of the province (Ministry of Environment (MOE), 2007). Low population densities and the absence of large metropolitan areas have allowed many aquatic ecosystems to remain relatively unimpacted by human activities and able to provide the important environmental values described above (MOE, 2007). Over the past decade however, the British Columbia Environmental Assessment Office website (EAO, 2010a) shows that there has been an increasing number of resource development projects proposed for northwestern B.C. This

suggests a growing realization of the economic opportunities afforded by the abundant natural resources in the region, but may also signal an increasing risk to stream ecosystems.

Of all the existing and proposed land and water use activities in northwestern B.C., perhaps the most significant emerging stressor on the landscape is mining and mineral exploration. Provincially in 2009, there were more than 350 exploration projects underway (including over 80 significant exploration projects), 36 new mine development proposals under review, and a number of operating mines including 8 metal, 9 coal and 30 other industrial minerals (Fredericks, Grieve, Lefebure, Madu, Northcote & Wojdak, 2009). The northwest is quickly becoming a hotbed of mining activity with nearly half of the exploration expenditures in B.C. occurring in this region (Wojdak, 2010). Mining activities are now an important contributor to the provincial and national economy, and they provide many societal benefits ranging from the products they produce to the local jobs they create (Mining Association of B.C., 2010). To ensure environmental sustainability, resource managers must carefully manage these developments so they do not compromise environmental quality for future generations.

Metal mining activities can significantly affect aquatic environments by altering the hydrological regimes of watersheds and by impacting the chemistry of the water and sediments directly through managed and unmanaged discharges (Maret, Cain, MacCoy & Short, 2003; Marques, Martinez-Conde & Rovira, 2003). The most profound chemical changes often come from metal leaching and acid rock drainage, which is characterized by low (acidic) pH and increased concentrations of metals such as zinc, copper, lead, cadmium, cobalt, nickel and others (International Network for Acid Prevention, 2009). To protect against these and other impacts, mining projects (and other proposed developments) are required to undergo comprehensive environmental assessments at both the provincial and federal levels. These assessments examine

the project at a strategic level to ensure that it is able to meet sustainable development objectives (EAO, 2010b). Once a project completes the environmental assessment process, it typically proceeds through a permitting process where it obtains authorizations for various activities, including waste management. Once a mine is operational, environmental protection is achieved through an adaptive management process that utilizes environmental monitoring information to inform ongoing regulatory decision-making.

In British Columbia, environmental monitoring of the mining industry usually occurs through implementation of a program called Environmental Effects Monitoring (EEM). EEM is a scientific, cyclical monitoring program designed to help determine effects in aquatic ecosystems caused by industrial effluent (Environment Canada, 2002). According to Environment Canada (2002), EEM results are combined with other social and economic information and used to evaluate the effectiveness and appropriateness of pollution prevention and control technologies, practices and programs, and indicate where there is a need for enhanced environmental protection or other adaptive management approaches. EEM typically involves conducting effluent and receiving environment water quality monitoring studies, as well as biological studies of fish, benthic macroinvertebrates (BMI) and, in some cases, other ecosystem components.

The requirement to implement EEM is consistent with a growing recognition that conducting environmental monitoring and impact assessment by measuring the physical and chemical properties of water alone is not sufficient to understand the health of aquatic ecosystems. To truly measure the health of an ecosystem, one must measure the biota, rather than just the factors that affect them (Reynoldson, Bailey, Day & Norris, 1995). For bioassessment, BMI are commonly recognized as useful indicators of aquatic ecosystem health since they are

sedentary, ubiquitous, relatively easy to sample, and have been shown to respond to environmental stresses (Rosenberg & Resh, 1993). They act as continuous monitors of the water they inhabit and they integrate a broad range of stresses including variable and multiple contaminant concentrations which may have antagonistic or synergistic effects, and changes in water quantity and habitat degradation (Rosenberg & Resh, 1993).

Over the years, a range of aquatic bioassessment techniques have emerged. Biologists in Australia, the United Kingdom, United States and Canada have focused on developing, testing and applying a bioassessment technique called the Reference Condition Approach (Bailey, Kennedy, Dervish & Taylor, 1998; Reynoldson et al., 1995; Wright, 1995). The Reference Condition Approach (RCA) has been extensively applied using BMI to evaluate stream ecosystem health (Bailey, Norris & Reynoldson, 2004). Its application involves quantifying the relationship between BMI communities and their environments at a large number of stream sites that have not been exposed to human-caused stressors ('references sites'). A model can then be developed to predict the BMI community expected at a site if it is in reference condition (i.e. not impacted). Evaluation of the condition of a 'test' stream site that has been exposed to human-caused stressors can then be performed by comparing its observed BMI community to that which would be expected if it was in reference condition (Reynoldson, Norris, Resh, Day & Rosenberg, 1997).

A RCA predictive model (called the "Skeena Model") was created for northwestern British Columbia in 2007, and was re-built in 2009, using the BEAST (BEnthic Assessment of SedimenT) method (Bennett, 2009; Perrin et al., 2007). This method was originally developed by Reynoldson et al. (1995) in the Great Lakes region of Canada, and similar methods have been applied in other regions of Canada including British Columbia and the Yukon (Bailey,

Reynoldson & Bailey, 2007; Reynoldson et al., 1997). As described in Reynoldson et al. (1995), application of BEAST starts by using cluster analysis to classify reference sites into fixed groups with similar BMI communities. Discriminant Function Analysis is then used to create a predictive model which quantifies the relationship between environmental variables and the BMI community at a site, identifying those environmental attributes that explain the groupings. When conducting bioassessments, the model determines which group of reference sites a particular test site should be compared to, based on its environmental attributes. The BMI community at the test site is then compared to those of the reference sites in its group and the degree to which the test site's community deviates from the expected reference condition community provides an indication of the severity of disturbance. When the BEAST method was applied in northwestern British Columbia, 129 reference sites clustered into four groups (Bennett, 2009). Details of the entire model building and validation process are included in Perrin et al. (2007) and Bennett (2009 and 2010).

While the Skeena Model was originally developed to provide a bioassessment tool to evaluate streams exposed to forestry activity, RCA bioassessments can be used for assessing effects of a range of stressors, including mining. Once a predictive model has been created, RCA is fairly simple to apply compared to other study designs such as "before-after/control-impact" (BACI). BACI studies require careful planning and intensive sampling efforts, making them costly and inappropriate for broad-scale application or for screening for impacts from ongoing activities or unforeseen events (Perrin et al., 2007). Unlike multi-metric indices which typically require expensive local calibrations, RCA predictive models can also be developed for larger regional scales, making the approach applicable for landscape-level analysis in addition to site-specific impact assessment (Bailey et al., 2004; Perrin et al., 2007). RCA bioassessments are

additionally attractive in Canada because Environment Canada has developed an online database management and analytical system called the Canadian Aquatic Biomonitoring Network (CABIN) which performs the statistical analyses necessary to conduct RCA bioassessments (Environment Canada, 2010). The CABIN programme also defines standardized protocols for data collection and analysis, and provides a repository to store and share data among users.

To most effectively use a RCA bioassessment to make sound environmental management decisions, decision-makers need to understand how that bioassessment performs. As is the case with any statistical hypothesis-testing approach, bioassessments that are based on statistical analysis of data are prone to Type I and Type II errors. A Type I error is often referred to as a false positive, and it occurs when the null hypothesis is rejected when it shouldn't be. A Type II error (also known as a false negative) occurs when the null hypothesis isn't rejected when it should have been (Samuels & Witmer, 2003). Type I error rates for statistical tests are set by identifying a desired level of protection against a false positive conclusion. Many researchers regard 5% as an acceptable risk and they set the Type I error probability (α) equal to 0.05 (Samuels & Witmer, 2003). Once α is defined, further analysis can be performed to determine the resulting Type II error rate (β).

Environmental biologists have historically focussed much more on α than on β, but in recent years there has been an increased recognition of the importance of power in biological assessments (di Stephano, 2003; Quinn & Keough, 2002). Power is the probability that a statistical test will detect an effect when one actually exists in the population, and it is related to β according to the formula $Power = 1 - \beta$ (Quinn & Keough, 2002). Power is normally managed by defining an effect size which identifies a significant difference between an observation and the null hypothesis, and then conducting power analysis calculations to determine the sample

size required to achieve a targeted power for the acceptable Type I error rate and the defined effect size. It has become common practice to simply target a power of 80% (which equates to a Type II error rate of 0.20), without any consideration to the relative size of the two error types and the consequences of making them (Bailey et al., 2004; di Stephano, 2003; Mapstone, 1995).

In RCA bioassessments a Type I error is made when a site that is actually in reference condition is mistakenly judged to be impacted. Conversely, a Type II error results when the bioassessment has failed to detect an impact when it exists, and a site is deemed to be in the reference condition when in fact it is not (Bailey et al., 2004). Both errors are undesirable since resource managers neither want to unnecessarily restrict an activity that is not actually causing an impact (because of a Type I error), nor to allow an activity to continue if it is causing serious environmental impacts (because of a Type II error) (Quinn & Keough, 2002). From an environmental perspective, the risks associated with making a Type II error can be greater than those associated with a Type I error, since Type II errors can lead to widespread impacts that may be irreversible or very costly to remediate (Bailey, Reynoldson & Bailey, 2008; Mapstone, 1995; Quinn & Keough, 2002).

When applying RCA bioassessments, sample size adjustments cannot be used to set the power and corresponding β. The probability of making a Type I error is fixed by drawing an ellipse around a group of reference sites to create a decision point in the assessment. Most BEAST-based approaches have set $\alpha = 0.10$ by drawing a confidence ellipse which encompasses 90% of the sites in a reference group, and then comparing the locations of test sites to that ellipse (Bailey et al., 2004). This choice of where to draw the ellipse not only affects the Type I error rate, but it also affects β, and in an opposite manner. If α is decreased from 0.10 to 0.01, the ellipse becomes a 99% confidence ellipse rather than a 90% confidence ellipse, and its size

becomes larger. With this decrease in α , the likelihood of an impacted test site falling inside the ellipse has increased, thus resulting in a greater likelihood of arriving at a false negative conclusion (i.e. a Type II error).

It is obvious that the best bioassessments would have low α and β. However, errors in RCA bioassessments are a reality and those developing these tools must evaluate and communicate information about their error rates and performance to decision-makers, who can then balance uncertainty with the risk of making an incorrect judgement (Bailey et al., 2008). Rudimentary testing of Skeena RCA bioassessment accuracy (its ability to determine that replicates of reference sites are not impaired) and precision (its ability to give the same assessment result to replicates of a single site) was performed at the time the model was built and it appeared to perform well (Bennett, 2009). However, potential error rates and power of the bioassessments in terms of their ability to detect impacts from activities such as mining should be examined in order to support their use for decision-making.

Evaluating Type II error rates normally requires a set of observations which are known to deviate from the null hypothesis by a certain effect size (Bailey et al., 2004). Unfortunately, in the case of RCA bioassessments it is not possible to use actual field data because the degree of impairment in natural systems is not known, but is only inferred based on measures of physical and chemical attributes (Cao & Hawkins, 2005). To overcome this challenge, biologists have created artificial data sets by simulating impacts on reference site data, and then testing them to see if the impact is detected by the bioassessment. Research in this area is ongoing, and common protocols for simulating impacts have not yet been agreed upon by the scientific community (J. Bailey, personal communication, 2010).

In the United States, Cao and Hawkins (2005) explored the use of simulation models to evaluate how well different biotic indicators (including non-metric multi-dimensional scaling (NMDS) and taxon richness) could measure ecosystem impairment. They recognized that simulations must incorporate the different responses of different BMI taxa (based on their sensitivities to disturbance) and they developed a procedure for simulating biological impairment of stream BMI communities. Using a predictive model and data from a large number of reference and test sites, they derived tolerance values for different taxa and then used a model to determine each taxon's abundance at a hypothetical "impaired" site based on its initial abundance, tolerance score and a coefficient that controls the level of stress that has occurred. Simulated data sets were similar in many ways to data sets from streams that were actually impaired, and Cao and Hawkins' research concluded that simulations can provide a useful tool to test bioassessment accuracy.

Biologists have simulated impacts on Canadian data sets using two different methods. In both, artificial BMI community data were created by applying a set of impact rules to reference site data sets. The rules describe how various BMI taxa respond to stress based on their tolerance score. In the Fraser River Basin, Mazor, Reynoldson, Rosenberg & Resh (2006) used a Microsoft Excel macro program called "*Impairator*" which simulates a generic pollution event that reduces biomass. "*Impairator*" determines the probability of death from an impact of a certain severity for each individual in a sample based on its pollution tolerance score (Linke, Norris & Robinson, 2004, as cited in Mazor et al., 2006). In the Yukon River Basin, Bailey et al. (2007 and 2008) simulated different levels of impact by establishing sets of rules for how various taxa (classified as sensitive, insensitive or tolerant) would respond to a stress based on a tolerance value. Bailey et al. experimented with three levels of impact in 2008. For example, at a moderate impact the

following rule was applied: all sensitive taxa decrease in abundance by 75% with a loss of 50% of the taxa (randomly selected), insensitive taxa decrease by 50% with a 20% loss, and tolerant taxa decrease by 25%, with 10% loss. They found that power and error varied depending on original structure of the reference community, but they demonstrated that "*simpacted"* data (artificial data sets created by impact simulations) can be a valuable tool for assessing bioassessment performance.

Approaches to create artificial data sets have all relied on the use of tolerance values. Tolerance is the degree to which an organism can withstand a specific pollutant or other environmental factor (Blockstrom & Winters, 2006). The concept of tolerance has been well studied for BMI, and it is widely recognized that each BMI taxon has an identifiable level of tolerance when it is subjected to a disturbance or stress (Huggins & Moffett, 1988). Some BMI taxa are relatively pollution tolerant while others are highly sensitive to pollution. Because of differing levels of tolerance, communities of BMI will respond to changes in their environment through changes to their population as well as community structure (Reece & Richardson, 2000).

Comprehensive lists of tolerance values for common BMI taxa have been compiled for many different areas. In North America, the most commonly referenced (and utilized) list was developed and refined by Hilsenhoff in the late 1970's and 1980's for Wisconsin (Hilsenhoff, 1988). Hilsenhoff's values represent tolerance to nutrient pollution. His tolerance list served as the basis of the Hilsenhoff Biotic Index, which is incorporated into the United States Environmental Protection Agency's Rapid Bioassessment Protocols. These protocols, published by Barbour et al. (1999), now include lists of taxa tolerance values for five different regions in the United States including the Northwest (Idaho), Midwest (Ohio), Southeast (North Carolina), Upper Midwest (Wisconsin) and Mid Atlantic Coast.

Many other researchers and resource managers have developed their own lists of tolerance values for use in bioassessment programs. A list of tolerance values has not been developed for use in British Columbia, so biologists and resource managers in B.C. have relied on tolerance values developed for other areas when they are needed. Creating artificial data sets using the methods already developed by Bailey et al. (2007 and 2008) and others requires this as well.

The purpose of this research project is to analyze and summarize the performance of Skeena RCA bioassessments, focusing on their use and applicability in the mining sector. The findings of this project provide important information to resource managers about how to use Skeena RCA bioassessments in an effective manner to make sound environmental management decisions. This project also provides valuable information to the research community's ongoing debate about how impact simulations can be used to evaluate the sensitivity of bioassessments. The purpose is achieved by answering the following research question: **What are the error rates and power of Skeena Reference Condition Approach bioassessments for detecting impacts from metal mining activities on stream environments in northwestern British Columbia?**

Chapter 2 - Research Methodology

In this research project I relied on data collected from streams in north-central and northwestern British Columbia. These streams are the 129 reference sites used in the Skeena Model. They cover a geographical area from Tumbler Ridge in the east to the Pacific Ocean in the west, and from Tweedsmuir Park in the south to the Yukon border in the north, and they fall within the Coast and Mountains, Central Interior, Sub-Boreal Interior and Northern Boreal Mountains Ecoprovinces (Perrin et al., 2007). The sites range from high elevation steep and turbulent streams, to wide and meandering valley-bottom rivers. Most were entirely wadeable, except a few of the largest rivers where only the margins could be sampled. Figure 1 shows the distribution of sites across the study area, including their group assignment in the Skeena predictive model.

My research methodology consisted of five steps. The first step was to assemble reference site data from streams in northwestern B.C. to use as the starting point for impact simulations. In the second step, detailed methods were defined for simulating disturbances of three different degrees (low, moderate and high), based on knowledge of how BMI respond to mining-related effects. The third step involved generating a data set of artificially-impacted test sites to represent BMI communities in aquatic ecosystems impacted by mining-related activities. The fourth step was to conduct Skeena RCA bioassessments on the artificially-impacted test sites, and the final step was to calculate and evaluate Type I and Type II error rates for each of the three degrees of impact.

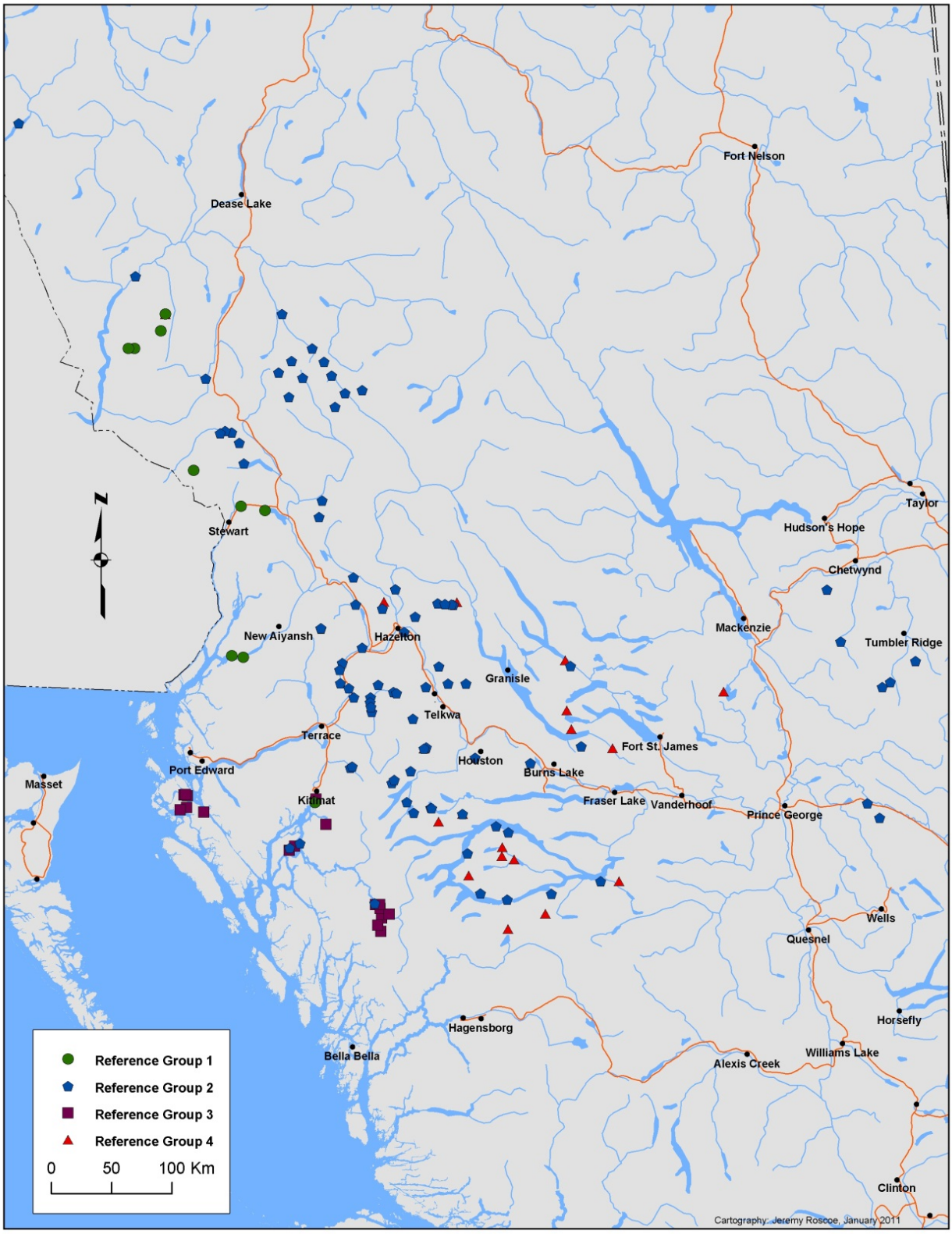

Figure 1. Distribution of reference sites across northwestern British Columbia.

I began the reference site data assembly by exporting family-level BMI taxonomic data for the 129 Skeena reference sites from Environment Canada's CABIN database into a Microsoft Excel (2007) spreadsheet. Family-level data were used since that is the taxonomic resolution used for Skeena RCA bioassessments. All sites had been sampled by the B.C. Ministry of Environment between 2004 and 2008, and samples were collected and analyzed by trained personnel using standard methods described in Bennett (2009) and Perrin et al. (2007). Raw data had been previously reviewed for QA/QC in the fall of 2008 by B.C. Ministry of Environment and Bio Logic Consulting staff. Each Skeena reference site was already assigned to one of the four groups by Bennett (2009) and, because the environmental attributes at the sites were not changed as a result of the disturbance simulations, the sites remained in their assigned groups.

I chose to use the general methodology developed by Bailey et al. (2008) because it could be easily applied with the current Skeena data set. To apply the methodology, I needed to develop a set of rules to describe how the reference site data would be manipulated in a way that changed the presence and abundance of BMI families based on their tolerance values.

A comprehensive list of BMI tolerance values for metal pollution is not available in the peer-reviewed literature (W. Clements, personal communication, 2010) so I assigned tolerance values using the Barbour et al. (1999) organic pollution tolerance list for Idaho as a starting point. I chose this list because there is no list available for British Columbia, and Barbour et al.'s lists are published and widely cited in the bioassessment literature. Compared to other areas listed in Barbour et al. and elsewhere, Idaho is geographically closest to British Columbia and is expected to have more similar environmental attributes including landscape, topography and climate, compared to other areas. I found a list of metal pollution tolerance values for Kansas

(Huggins & Moffett, 1988), but did not use it because it did not appear to be published in peer-reviewed literature and consequently, I could not confirm its scientific validity.

The 129 reference sites in northwestern B.C. contained 67 different BMI families. To simplify the reference site data set, 28 rare families that occurred at less than 5% of sites were not considered for manipulation since I did not expect them to strongly influence the definition of overall community structure in the Skeena Model reference groups. Of the remaining 39 families, 33 were assigned the Idaho tolerance value from Barbour et al. and five needed to be assigned a value from elsewhere since no value is published for Idaho. In most cases, I used the California tolerance list (Aquatic Bioassessment Lab, 2003) as a backup because of its comprehensiveness and geographic proximity to British Columbia. Table 1 summarizes the tolerance value assigned to each family, and Table 2 lists the rare families that were not assigned a tolerance value.

Table 1

Family-level tolerance values assigned to Skeena BMI.

Order	Family	Tolerance
Diptera	Blephariceridae	0
Ephemeroptera	Ameletidae	0
Plecoptera	Leuctridae	0
Trichoptera	Glossosomatidae	0
Trichoptera	Rhyacophilidae	0
Trichoptera	Uenoidae	0
Ephemeroptera	Ephemerellidae	1
Plecoptera	Capniidae	1
Plecoptera	Chloroperlidae	1
Plecoptera	Perlidae	1
Trichoptera	Apataniidae	1*
Trichoptera	Brachycentridae	1
Tricladida	Planariidae	1
Ephemeroptera	Leptophlebiidae	2
Plecoptera	Nemouridae	2
Plecoptera	Perlodidae	2
Plecoptera	Taeniopterygidae	2
Diptera	Tipulidae	3
Coleoptera	Elmidae	4
Ephemeroptera	Baetidae	4
Ephemeroptera	Heptageniidae	4
Trichoptera	Hydropsychidae	4
Trichoptera	Limnephilidae	4
Haplotaxida	Naididae	5**
Oribatei	Hydrozetidae	5**
Prostigmata	Hydryphantidae	5**
Prostigmata	Torrenticolidae	5**
Diptera	Ceratopogonidae	6
Diptera	Chironomidae	6
Diptera	Empididae	6
Diptera	Simuliidae	6
Lumbriculida	Lumbriculidae	8
Prostigmata	Hygrobatidae	8
Prostigmata	Lebertiidae	8
Prostigmata	Sperchonidae	8
Veneroida	Sphaeriidae	8
Diptera	Psychodidae	10
Haplotaxida	Enchytraeidae	10
Haplotaxida	Tubificidae	10

Note. Increasing numbers correspond to increasing levels of tolerance. The *denotes a genus rather than family level value. All values are from the Idaho list in Barbour et al. (1999) except those marked with ** which are from the Aquatic Bioassessment Lab (2003).

Table 2

Rare BMI families and the number of sites where they were present.

Order	Family	# sites
Amphipoda	Crangonyctidae	1
Basommatophora	Planorbidae	2
Coleoptera	Amphizoidae	1
Coleoptera	Dytiscidae	3
Coleoptera	Hydrophilidae	1
Coleoptera	Staphylinidae	3
Collembola	Isotomidae	3
Collembola	Sminthuridae	5
Diptera	Athericidae	2
Diptera	Deuterophlebiidae	5
Diptera	Ephydridae	4
Diptera	Oreoleptidae	1
Diptera	Phoridae	1
Diptera	Sarcophagidae	2
Diptera	Stratiomyidae	1
Heterostropha	Valvatidae	2
Odonata	Gomphidae	1
Oribatei	Halacaridae	1
Oribatei	Oribatidae	2
Plecoptera	Peltoperlidae	1
Plecoptera	Pteronarcyidae	1
Prostigmata	Aturidae	2
Prostigmata	Mideopsidae	2
Prostigmata	Stygothrombidiidae	3
Trichoptera	Hydroptilidae	5
Trichoptera	Lepidostomatidae	3
Trichoptera	Leptoceridae	2
Trichoptera	Philopotamidae	3

In this project, I examined three levels of disturbance: low, moderate and high, and I classified the 39 BMI families into three tolerance classes for the impact simulations. At the low level of impact, sensitive families were reduced in richness and abundance, moderately tolerant families were unchanged and tolerant families increased in abundance. At moderate impact, richness and abundance of sensitive families were further reduced, the moderately tolerant families' abundance and richness were reduced and the tolerant families' abundance further increased. At the high impact level, all sensitive families were eliminated, moderately tolerant families further reduced in abundance and richness and tolerant families also reduced in abundance and richness. These simulated impacts are detailed in Table 3.

Table 3

Impact simulation rules used for manipulating reference site communities

	Low Impact		Moderate Impact		High Impact	
Tolerance	Abundance	Richness	Abundance	Richness	Abundance	Richness
0-2	-25	-10	-50	-50	-100	-100
3-6	n/c	n/c	-25	-10	-50	-50
7-10	+25	n/c	+50	n/c	-25	-10

Note. Changes are indicated as percentages; n/c = no change.

The impacted BMI data set was created by manipulating the reference site data in Microsoft Excel. The abundance changes in the disturbance simulations were simply a percentage increase or decrease in individuals at a site, depending on their original abundance. Formulas were written in Microsoft Excel to scale the abundances up or down according to the impact rules in Table 3. The richness manipulations were more complex. Since each original reference site had variable family richness as a starting point, the number of families to remove

from each site was calculated based on the average richness of the sites in its group. For example, Group 1 sites had an average richness of 5.4 sensitive families, therefore a 50% decrease in sensitive family richness was achieved by removing 2.7 (=3) families. Group 2 sites, however, had an average richness of 9.4 therefore the 50% reduction was achieved by removing 4.7 (=5) families. To determine which families to remove, I selected taxa randomly using random numbers generated in Microsoft Excel. Different random numbers were generated for each site, so the families removed from one site were not necessarily the same as the families removed from another.

The result of the Microsoft Excel manipulations was spreadsheets with 3 data sets (low, moderate and high impact) of 129 test sites each. To help quantify the extent of the simulated impacts, and to ensure that the test sites appeared to be impacted by mining, I calculated total family and Ephemeroptera, Plecoptera and Trichoptera (EPT) family richness and abundance for the sites in each group at each simulated impact level.

Test site assessments were performed using methods described in Bennett (2009). PRIMER (Version 6) software was used to create a Bray-Curtis similarity matrix for the family-level BMI data for each single test site and the reference sites in its predicted group. A non-metric multidimensional scaling (MDS) 3-dimensional ordination plot was created and the spatial position of the test and reference sites was recorded into a Microsoft Excel spreadsheet. In the ordinations, the position of the test site relative to the reference sites indicates the degree of similarity between the community structure of the reference and test sites. If the test site is near the reference sites, the community structure is more similar while if the test site is far away from the reference sites, the community structure is less similar.

Systat (Version 13) software was then used to plot the coordinates of each single test site along with the reference sites of its corresponding group. Probability ellipses were drawn at the 75%, 90%, 99% and 99.9% levels in Systat, with the test site and its corresponding original reference site highlighted. From the ordination plots, the location of each test site and the reference site it was derived from were recorded. If the impacted test site fell within a particular confidence ellipse, a Type II error was recorded. If it fell outside the ellipse, an error had not been made. In addition, the location of the corresponding reference site was also recorded, and if the reference site fell outside a particular ellipse, a Type I error was recorded. No Type I error occurred if it fell within the ellipse (see Table 4 for criteria, and Figure 2 for example plots of the types of errors).

Table 4

Criteria for occurrence of Type I and Type II errors for a given probability ellipse.

Site location	Reference Site	Test Site
Inside Ellipse	No Error	Type II Error
Outside Ellipse	Type I Error	No Error

To display the location of the sites in 3-dimensional space, three 2-dimensional ordination plots were created (Axis 1 vs. Axis 2, Axis 1 vs. Axis 3, and Axis 2 vs. Axis 3). If any of the three plots for a given site displayed an error, then an error was recorded because a site that falls outside an ellipse in any one 2-dimensional plot would fall outside the 3-dimensional ellipse if it could be drawn on paper.

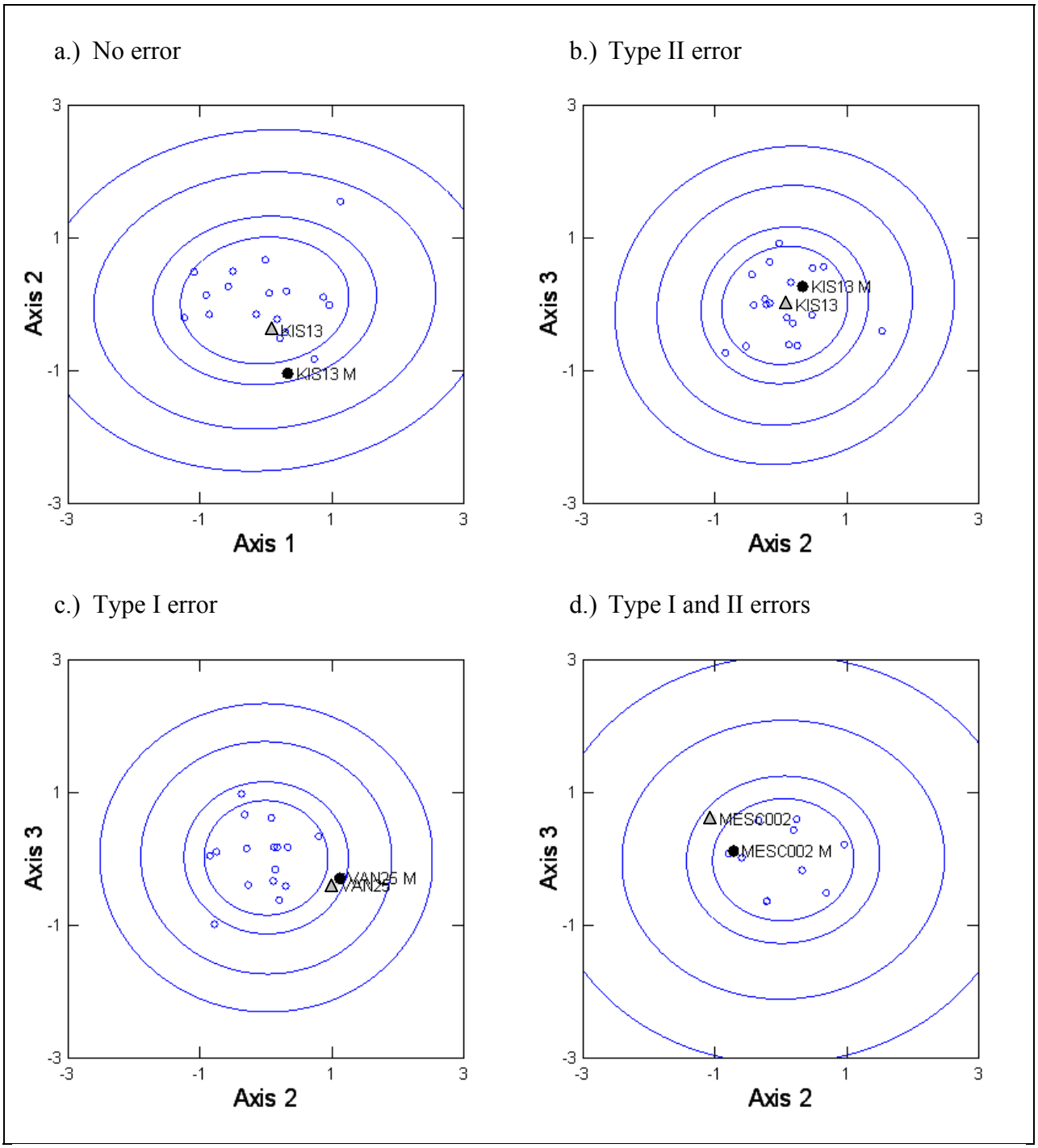

Figure 2. Example ordination plots show the location of a test site (black dot) and its corresponding original reference site (grey triangle), along with the rest of the reference sites in its sample group (open circles). Four confidence ellipses [75% (innermost ellipse), 90%, 99% and 99.9% (outermost ellipse)] are drawn in each plot. In (a) there are no errors at the 75% confidence ellipse level because the reference site is inside the 75% ellipse and the test site is outside. In (b) there is a Type II error because the test site lies inside the 75% ellipse. In (c) there is a Type I error because the reference site is outside the 75% ellipse. In (d) both error types exist because the reference site is outside the 75% ellipse while the test site is inside.

Chapter 3 – Results

3.1 Reference Site Data Sets

Benthic macroinvertebrate data were compiled for 129 reference sites that had been assigned into four groups. Figure 3 summarizes the taxonomic composition of the four reference groups, and illustrates that there are substantial differences in total and relative taxa abundances among the different groups.

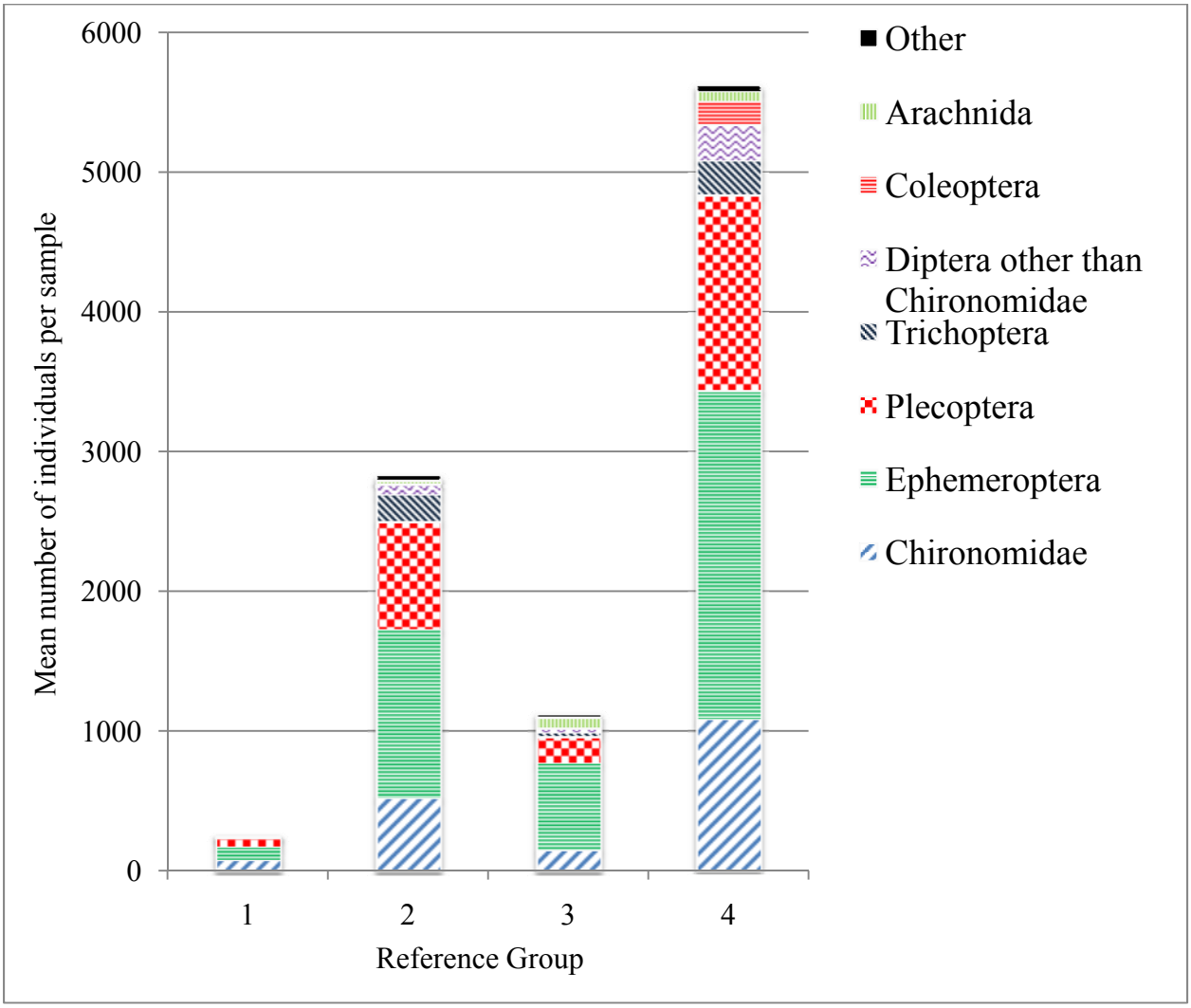

Figure 3. Taxonomic composition of the four reference groups in the Skeena Model (adapted from Bennett, 2009).

In all groups, a majority of the BMI individuals fell within the two most sensitive tolerance classes, 0-2 and 3-6. Figure 4 summarizes the average proportion of individuals that fell within each tolerance class in each of four groups in the Skeena Model.

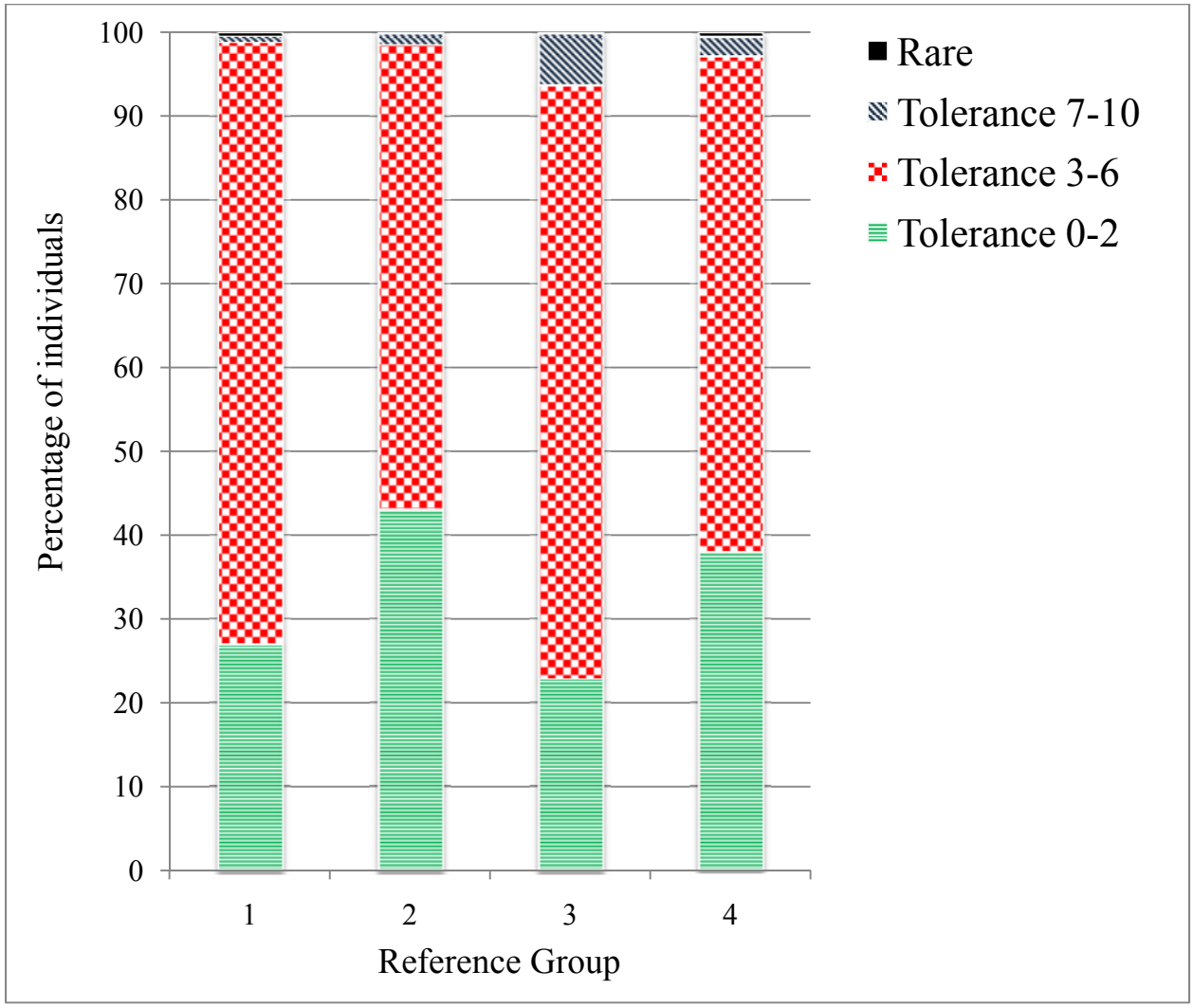

Figure 4. Average percentage of individuals in each group that were assigned to each tolerance class.

3.2 Impacted Benthic Macroinvertebrate Data Sets

The disturbance simulations created 129 test sites at each of the three levels of impact. In all cases, total abundance and total richness dropped as the level of impact increased. The low impact scenario had only a minor effect on these metrics, but much more pronounced differences occurred with the moderate and high impact scenarios. EPT family abundance and richness metrics responded more prominently than the total abundance and richness metrics. Community level responses for total abundance and richness and EPT family abundance and richness, are illustrated in Figures 5 and 6.

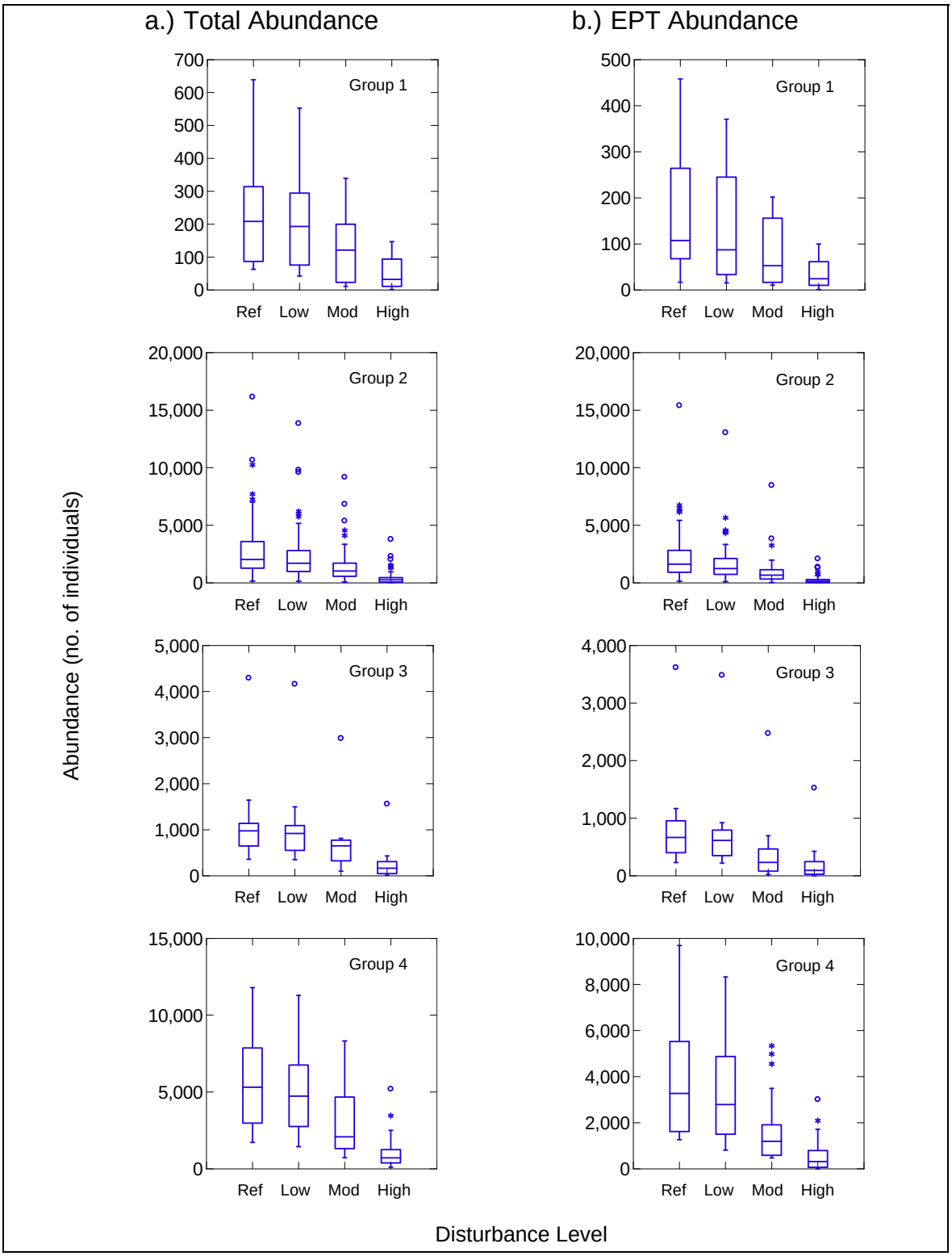

Figure 5. Community-level abundance responses to impact simulations for (a) total abundance and (b) abundance of Ephemeroptera, Plecoptera and Trichoptera (EPT) families. Box plots show the minimum, maximum, median and quartile values, with outliers identified as dots.

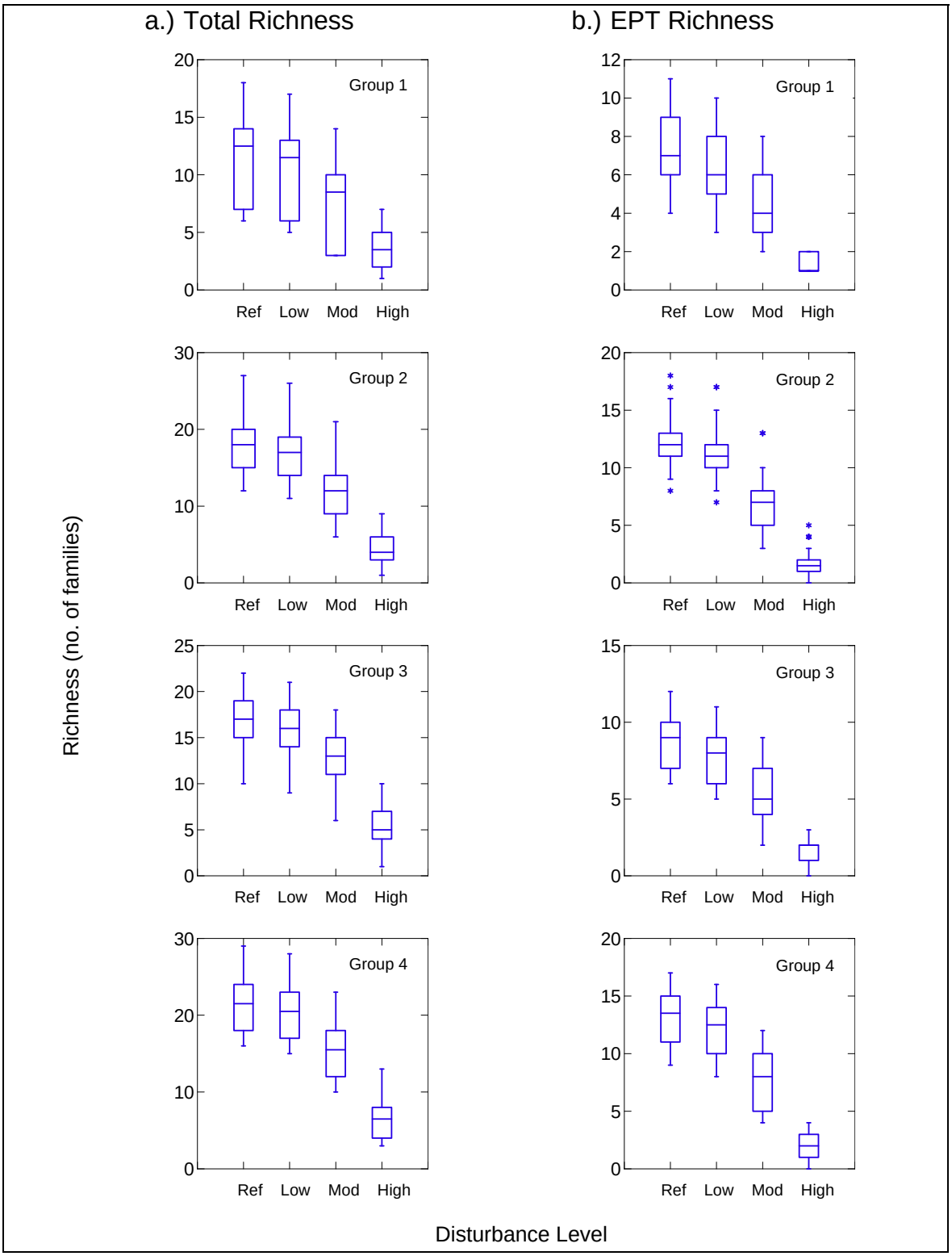

Figure 6. Community-level richness responses to impact simulations for (a) total richness and (b) richness of Ephemeroptera, Plecoptera and Trichoptera (EPT) families. Box plots show the minimum, maximum, median and quartile values.

3.3 RCA Bioassessments of Simulated Impact Data Sets

Assessments were performed on a total of 387 sites - 129 sites at each of the three levels of simulated impact. In each case, the test site and its corresponding reference site were plotted on three different axis combinations (Axis 1 vs. Axis 2, Axis 1 vs. Axis 3, and Axis 2 vs. Axis 3). Figure 7 provides an example of the three different plots for a test site (PRG02H) and its corresponding reference site (PRG02). If any of the plots for a given site displayed an error, then an error was recorded.

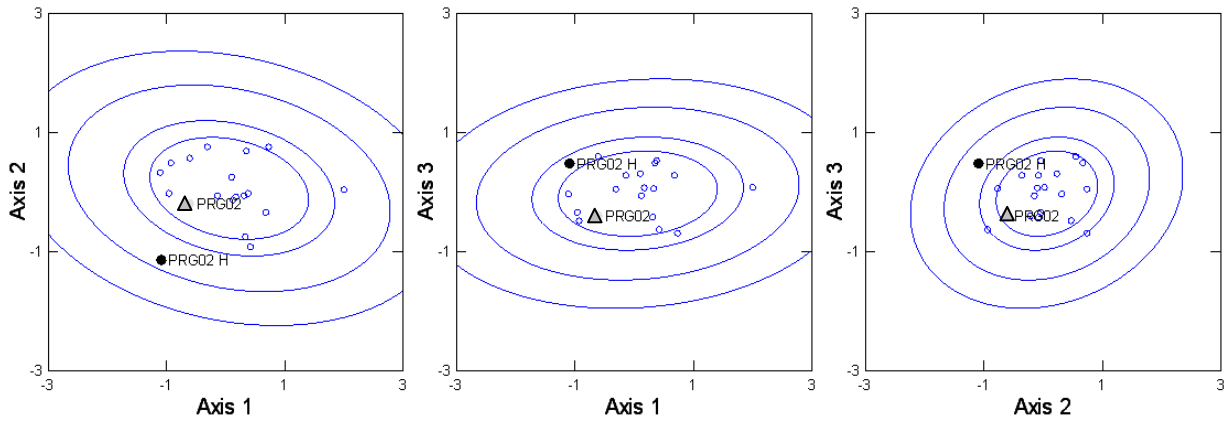

Figure 7. Ordination plots of a test site (PRG02H - black dot) and its corresponding original reference site (PRG02 - grey triangle), along with the rest of the reference sites in its sample group (open circles). Four confidence ellipses are plotted (75%, 90%, 99% and 99.9%). The grey reference site falls within all ellipses, therefore a Type I error is not made at any level of α. The black test site falls outside the 90% ellipse in at least one plot, indicating that at α=0.25 and α=0.10 (represented by the two innermost ellipses) a Type II error is not made, but at α=0.01 and α=0.001 a Type II error exists (since the test site falls inside these two outer ellipses in all three plots).

Because the assessment approach involved comparison of a test site to a specific group of reference sites, the results of the bioassessments are reported for each of the four different groups. Table 5 summarizes Type I and Type II error rates at each impact level for each group.

Table 5

Type I and Type II error rates at low, moderate and high levels of disturbance for each of the four groups in the Skeena predictive model.

	Disturbance Level					
	Low		Moderate		High	
α	Type I	Type II	Type I	Type II	Type I	Type II
Group 1 (n=10)						
0.25	0.20	0.80	0.20	0.40	0.10	0.20
0.1	0.00	1.00	0.00	0.60	0.00	0.30
0.01	0.00	1.00	0.00	0.80	0.00	0.80
0.001	0.00	1.00	0.00	1.00	0.00	0.80
Group 2 (n=84)						
0.25	0.50	0.49	0.49	0.14	0.49	0.01
0.1	0.26	0.75	0.23	0.37	0.25	0.05
0.01	0.05	0.96	0.05	0.76	0.05	0.23
0.001	0.01	0.99	0.01	0.89	0.01	0.40
Group 3 (n=17)						
0.25	0.41	0.59	0.47	0.29	0.41	0.06
0.1	0.12	0.88	0.12	0.47	0.12	0.12
0.01	0.00	1.00	0.00	0.88	0.00	0.41
0.001	0.00	1.00	0.00	1.00	0.00	0.59
Group 4 (n=18)						
0.25	0.56	0.39	0.56	0.28	0.39	0.00
0.1	0.17	0.78	0.11	0.50	0.11	0.00
0.01	0.00	0.94	0.00	0.89	0.00	0.28
0.001	0.00	1.00	0.00	1.00	0.00	0.72

Note. n=sample size (number of sites in each group). α=Type I error probability, set by drawing ellipses (e.g. α =0.25 indicates that the 75% confidence ellipse is used to read the ordination plot).

The results show that as α decreases, the actual Type I error rate decreases as well. At α=0.001 the Type I error rate is at or very close to zero and, except in the largest group (Group 2), the error rate is also zero at α=0.01. Type I error rates are of similar magnitude for Groups 2, 3 and 4, but are slightly lower in Group 1. Type I error rates change slightly depending on the

level of disturbance, which is an effect of having a test site plotted with the reference groups (test sites are known to have some effect on how the reference sites ordinate – S. Bennett, personal communication, 2010; J. Bailey, personal communication, 2010).

Type II error rates increase as the value of α decreases. Type II error rates are variable between groups, with Group 1 generally displaying the highest levels. Groups 2 and 4 have slightly lower Type II error rates than Groups 1 and 3. Overall, all groups show a decreasing error rate as the level of simulated impact increases. Error rates are quite high at low levels of disturbance. The lowest error rate at α=0.25 is 0.39 for Group 4 and the highest is 0.80 for Group 1. Figures 8 to 10 show the error rates for the different groups at each disturbance level.

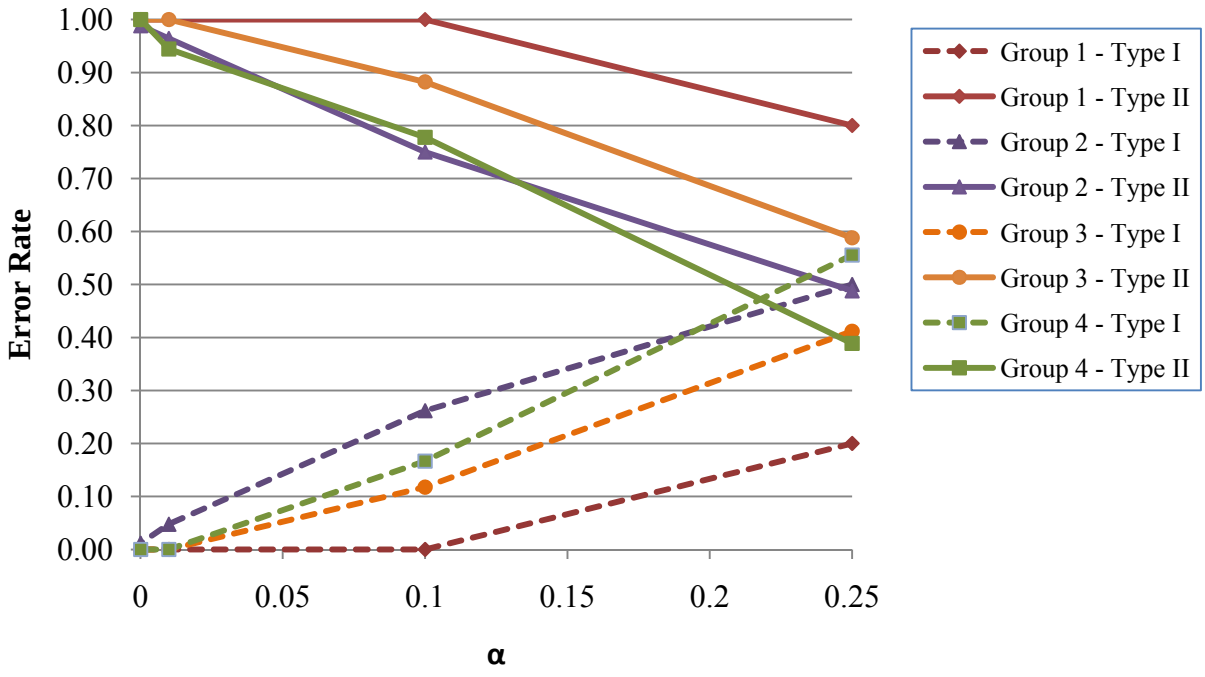

Figure 8. Type I and Type II error rates for various levels of α (probability of Type I error) at low simulated disturbance.

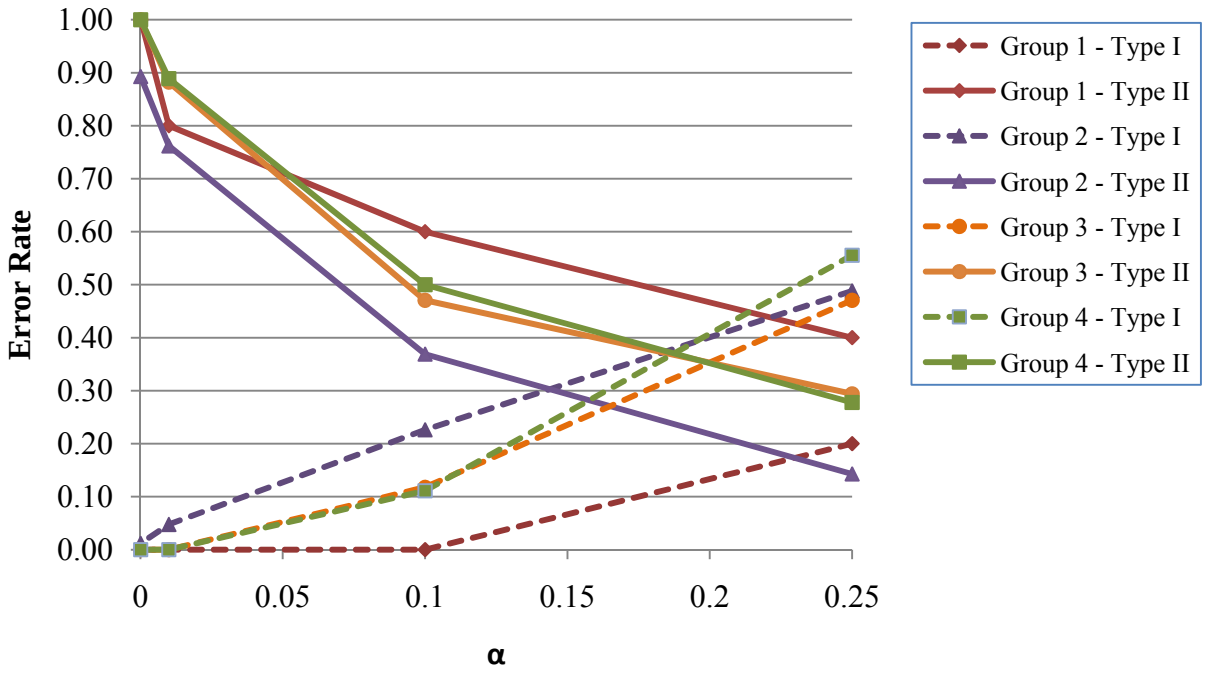

Figure 9. Type I and Type II error rates for various levels of α (probability of Type I error) at moderate simulated disturbance.

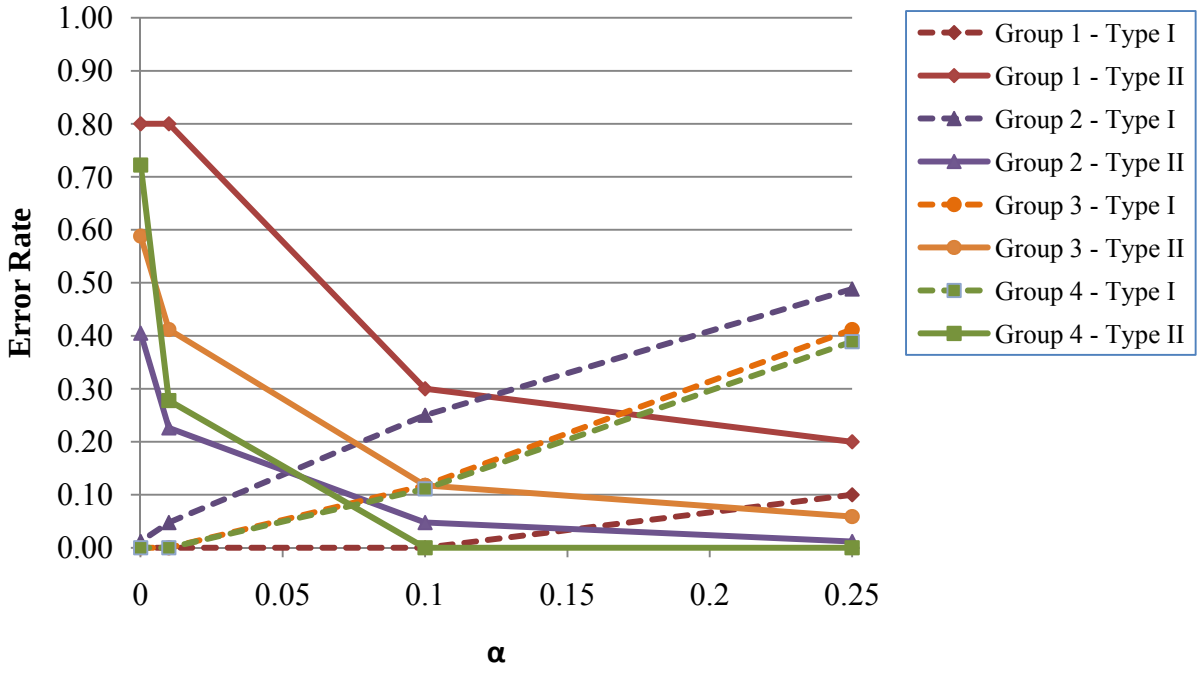

Figure 10. Type I and Type II error rates for various levels of α (probability of Type I error) at high simulated disturbance.

Chapter 4 - Discussion

4.1 Methodology for Simulating Disturbances

The concept of simulating disturbances to test RCA bioassessments has been published in the literature for more than five years (Bailey et al., 2004; Bailey et al., 2008; Cao & Hawkins, 2005; Mazor et al., 2006); however, biologists involved in studying RCA bioassessments have yet to agree on the best approach to do this. Some progress was made at a workshop in June 2010 when many bioassessment practitioners agreed that impact simulations could be a valuable tool for assessing bioassessment performance (J. Bailey, personal communication, 2010), and since then, Bailey (J.) and others have continued working to develop and test different methods.

My methodology of removing individuals and entire families based on tolerance values was derived from work done by Bailey et al. (2008) in the Yukon. However, I needed to make decisions about how to specifically apply the methodology to the Skeena data set and these decisions have undoubtedly affected the results. Influential decisions included the choice of tolerance values for individual families, the magnitude of abundance and richness changes at different levels of disturbance, and deciding how many families to remove and which ones. In addition, choices I made about how many sites to plot at once in the ordinations had some effect on the resulting error rates.

The purpose of this project was to evaluate bioassessment performance in the context of mining impacts, so the disturbance simulations that I used needed to create BMI communities that resemble those which have been subjected to mining-related effects. Unfortunately, the most commonly used tolerance value lists are based on the ability of BMI taxa to withstand organic pollution and nutrient enrichment, rather than metal pollution. Clements (personal

communication, 2010) indicated that any comprehensive tolerance lists for metal pollution are rare and one is certainly not available for British Columbia. In the absence of an accepted metal-tolerance value list, I examined a number of lists for organic pollution and other contaminants. The five lists (for Idaho, Ohio, North Carolina, Wisconsin and Mid Atlantic Coast) in Barbour et al. (1999) were very similar to lists developed for Nova Scotia (Mandaville, 2002), California (Aquatic Bioassessment Lab, 2003) and Kansas (Huggins & Moffett, 1988) suggesting that any particular list or combination of lists may provide data at a sufficient level of accuracy for this project. Furthermore, Blocksom & Winters (2006) evaluated different methods for creating tolerance values. After examining and analyzing lists from Delaware, Kentucky, Maryland, Massachusetts and a United States Environmental Protection Agency (USEPA) laboratory, they found that at the family level of taxonomic resolution there "were not systematic differences in tolerance values, and correlations were high among lists" (Blockstrom & Winters, 2006, p. iii).

Huggins and Moffett (1988) indicate that tolerance values for different pollutants are not directly comparable; however there are some general similarities among pollution types, such as Ephemeroptera, Plecoptera and Trichoptera taxa being more sensitive than other orders. Clements (1994) warns that BMI may respond differently to toxic chemicals than they do to organic pollution, and suggests that caution is required when using biotic indices based on tolerance to nutrient enrichment for assessing metal pollution. When selecting the Idaho organic pollution tolerance list, I reviewed differences between the commonly-used Idaho values and corresponding Kansas metal pollution tolerance values and found that most Skeena BMI families fell within the same tolerance category in both lists. Only seven families would have fallen in a different category had I used the Kansas values, but in all cases the difference was a single

category. In approximately half of the cases the families were more sensitive, while in the other half they were less sensitive.

I also reviewed numerous mining-related studies to ensure that the Idaho tolerance values did not misrepresent impacts expected from mining activities. Many studies found that pollution from mining activities resulted in a response similar to that from organic pollution for most BMI taxa, and sensitivities of specific taxa have been shown to be similar by researchers such as: Clements (2004) who estimated toxic concentrations for various taxa; Clements, Carlisle, Lazorchak, & Johnson (2000) who described the responses of various dominant taxa to metals; Beasley and Kneale (2003) who developed lists of the most sensitive and tolerant families for a range of environmental variables including many metals; and Marque et al. (2003) who developed a list of families that are sensitive to disturbance in the water column and metal content in the sediment. Marque et al. indicate that "generally speaking, the families that are sensitive or tolerant to this mining disturbance are also sensitive or tolerant to organic pollution" (2003, p. 381). In addition, Tripole, Vallania, and del Carmen Corigliano (2008) studied BMI response to acidic stress in Argentina and developed taxa tolerance values based on the occurrence of taxa at different pH values. Their study results were consistent with earlier studies on acidification, which observed the disappearance or decreased abundance of sensitive taxa such as individuals in the orders of Ephemeroptera, Plecoptera and Trichoptera (EPT). Based on this review of research, and because my methodology involved lumping the tolerance values into three categories (0-2, 3-6 and 7-10), I believe that minor differences in values that could have resulted from using a different tolerance list did not significantly affect the results.

Previous research is inconclusive respecting quantifying the effects of mining on BMI community abundance and richness. However, a large number of studies have found that

increased metal concentrations are generally associated with decreases in richness and abundance, and sensitive taxa such as EPT are lost or replaced by more tolerant taxa (Clements, 1994; Clements et al., 2000; Hirst, Jüttner & Ormerod, 2002; Maret et al., 2003; Marques et al., 2003; Pollard & Yuan, 2006; Smolders, Lock, Van der Velde, Medina Hoyos & Roelofs, 2003). In a series of mesocosm experiments in the Colorado Rockies, Clements (2004) showed that at low to moderate metals exposure, total abundance and richness did not change significantly, however, EPT taxa richness and abundance generally showed measurable decreases. At higher metal loadings, abundance and richness were typically both affected but, overall, measures of abundance were found to be more sensitive to metals than measures of richness, especially at lower concentrations. In northwestern British Columbia, Limnotek Research and Development (1992) conducted a mesocosm study at Equity Mine where acid mine drainage (AMD) was added at varying concentrations. They found that densities decreased with increasing AMD, especially once a threshold was reached. They also found that richness did not change as significantly as abundance (Limnotek Research and Development, 1992).

To accurately reflect mining impacts, the disturbance simulation rules I used in this project were developed so they would result in greater shifts in abundance than in richness. BMI abundance needed to be affected to a small degree at low levels, and be increasingly affected at higher levels. Total richness needed to be only minimally affected at low levels, but to show greater effects at the highest levels of disturbance. EPT richness and abundance needed to show greater changes than total richness and abundance, with effects beginning at relatively low levels of disturbance. Examination of metric results for the four groups indicates that this was achieved with the methodology for this study.

The magnitude of the abundance and richness changes clearly affected the resulting test site data set, and therefore had some effect on the assessments and error rates. Determining the extent to which small differences in these magnitudes affect the final assessments and error rates is beyond the scope of this project, but the metrics indicate that the rules are reasonable for creating a mine-influenced data set for bioassessment evaluations. The magnitudes that I chose created a low disturbance scenario which resulted in a minor but measurable shift in community structure and small changes in some metrics. The moderate and high disturbance scenarios created greater changes in community structure and are levels of impact that a resource manager would likely want to avoid, especially when environmental values are high.

Another key decision in applying my methodology was determining how many families to remove from each site when applying the richness changes. The simplest approach would have been to identify particular families from a list of all that were present and then remove those same families from the entire data set, regardless of how many different families were at each site to start with and whether or not the selected family was actually present. The four Skeena groups have somewhat different richness characteristics, so removing a fixed number of families from every site would have a greater effect on Group 1 sites, which have substantially lower richness values than those in other groups. A more complex approach would be to create test site data sets where the number of families removed at a test site is based on the richness of its original reference site. The approach I used in this study involved calculating the average richness in each reference group, then applying the percent reduction based on this number. I believe that this was a reasonable compromise between calculating richness at each site and relying on the average richness of all sites together, since it allowed for consideration of differences among the groups.

To ensure that all sites were actually being disturbed, only families that were present at a particular site were considered for removal. While families that I removed from each site were randomly selected, the total number of families removed from sites in a particular group was constant according to the percentage change identified in the rules. At some sites, very abundant families were removed while at others, lesser-abundant families were removed. In addition, some sites lost families that were relatively rare in the data set and therefore not likely an important factor in defining the expected community in a reference group while other sites lost families that were widely present. I decided that applying a random approach to the removals effectively provided a realistic sense of what could happen when a certain number of BMI families go missing from a site as a result of an impact.

I believe that this aspect of the methodology had the greatest effect on the results of the RCA assessments and corresponding error rates, but the random selection of families to remove reduced bias in the taxa removal process. To better understand how the choices related to the removals affected the final results, further investigation would be required. For example, I did not explore the possibility of adding tolerant families that may have been absent from a particular reference site, and this could be an area of future study.

Finally, I calculated error rates in this study by comparing the location of a single test site with its assigned group of reference sites. Both the Type I and Type II errors were tabulated at the same time, from the same ordination plots (i.e. in each plot, the locations of both the test site and its corresponding reference site were recorded). An alternative approach that I could have used for calculating error rates involves plotting multiple sites at once, or plotting only the reference sites by themselves to measure the Type I error rates. I conducted some cursory trials of other error calculation approaches, and found that they yielded similar but slightly different

results. I believe that the approach I used in this study is the best method for calculating error rates since the two corresponding error types are determined at the same time in the same assessment, and the addition of more test sites to the ordination plots is known to have some effect on how the reference sites ordinate.

4.2 Tradeoffs Between Type I and Type II Error Rates

Historically, many statistical hypothesis tests set a probability threshold for rejecting a null hypothesis at p=0.05, which is a Type I error probability of 5% (Mapstone, 1995). When applying RCA, the null hypothesis is that a test site is in reference condition and the threshold is commonly set at 10% by drawing a 90% probability ellipse around a group of reference sites (Bailey et al., 2004). Environmental biologists have recently started to move beyond just thinking about Type I errors and have begun to also think about the power of hypothesis tests and the probability of making a Type II error (β). Power of 80% has become a common decision point in cases where power is considered at all (Bailey et al., 2004; di Stephano, 2003). While some consideration is better than nothing, Bailey et al. (2004) argue that this arbitrary approach is flawed, that it is the relative importance of the two error types that really matters, and that their size relative to a decision point is what needs to be determined and communicated. This is similar to the argument made by Mapstone (1995), who suggested that decision points should be set based on the relative importance and costs of making the two different types of errors.

The results from this study clearly demonstrated that error rates in Skeena RCA bioassessments vary drastically depending on which value of α is used for drawing an ellipse. Both error types are affected by α and they behave in opposite ways, indicating that there are trade-offs to be made when setting the decision point of α. In Skeena RCA bioassessments, Type

I error rates are very low when α is small, and the rates increase as α increases. Type II error rates, on the other hand, are very high when α is small but tend to decrease as α increases.

The different groups show similar trends in error rates, but the exact magnitudes of the rates differ among them. The error rates for Group 1 differ most from the other three groups. This is likely because these sites tend to have different BMI communities than sites in the other groups: abundance and richness metrics are considerably lower and the overall community composition is also different. Lower Type I error rates in Group 1 suggest that perhaps this group has a more closely-clustered group of reference sites compared to the other groups. The generally lower Type II error rates in Groups 2 and 4 may be related to the fact that the average community composition at these sites includes a higher proportion of more sensitive families, so the disturbance simulations likely changed these communities more drastically than those in Groups 1 and 3. Overall, the fact that error rates differ among the groups highlights the importance of calculating error rates for each group separately, and not just for the overall bioassessment approach.

Type I error rates are not greatly affected as the level of disturbance changes. The largest group (Group 2) has quite consistent Type I error rates among disturbance levels, while the smaller groups show some minor variation since the location of a single site has a greater effect on the overall rates. The tightness of a cluster of reference sites will influence the location and size of the probability ellipse that is drawn around it, therefore affecting the resulting Type I error rate. Overall, Type I error rates are also generally higher than expected based on α, which represents the probability of making an error. Type I error is inherent in BEAST assessments because decision points are defined by drawing confidence ellipses that represent Type I error probabilities. However, results from this study should caution those interpreting assessment

results to keep in mind that the actual Type I error rate may be considerably higher than the expected Type I error rate.

Unlike the relatively constant Type I error rates, Type II error rates decreased as the level of impact increased because the effect size was larger. In most of the scenarios examined in this study, Type II error rates were greater than Type I error rates. The exceptions were in the high disturbance scenario when α was 0.10 or greater, and the moderate disturbance scenario when α was 0.25. Many RCA bioassessments, including those carried out with Environment Canada's CABIN online system, use α values of 0.10 and smaller (Environment Canada, 2010). In CABIN, even at the largest value of α=0.10, Type I and Type II error rates are not balanced when trying to detect low disturbances, and Type II error rates are very high. In these instances a better balance can only be achieved by setting a larger value for α, such as 0.25 or greater. Using α=0.10 for assessing moderately disturbed sites yields error rates that are a little more balanced, but it is with high disturbances that Type I and Type II error rates are roughly the same magnitude at α=0.10, when rates for both error types fall between 0 and 0.30 for all groups.

Power is the likelihood that a bioassessment will detect an impact when there is one and it is equal to *1-Type II Error Rate*. With the error rates observed in this study, the power of Skeena RCA bioassessments for detecting low levels of disturbance is relatively small when α=0.10. The highest power exists with Group 2 sites, where there is 25% power; the lowest power lies with Group 1, which actually had no power to detect low disturbance in this study. Power for detecting moderate disturbance is significantly better at 40-63% depending on the group and power for detecting high disturbance ranges from 70-100%.

4.3 Using Bioassessment Results for Management Decisions

When considering how to use the results from this study, it is important to bear in mind the sets of rules used to create the three increasing levels of artificial impacts. The impacts created from the disturbance simulations may or may not be significant in the ecological or management context of a particular real-world situation. Interpreting results from an experiment like this requires knowledge and understanding of the significance of the impacts to the context at hand. Decision-makers must first consider, from an ecological perspective, the level of impact or effect size that they want to detect and protect against; and only then should they interpret the error rates and trade-offs associated with that effect size. For example, the low disturbance scenario in this study did not significantly alter the BMI communities and in many cases, detecting this level of impact may not be important since ecological structure and function in the stream is probably not affected. On the other hand, the moderate disturbance simulation resulted in more substantial changes to the BMI community structure, and probably its function as well. In many cases, this level of impact may be worthy of protecting against. Determining the exact level of impact or effect size that a decision-maker should care about is complex, somewhat subjective, and is certainly beyond the scope of this study; however, the concept is worthy of consideration and perhaps further research

In addition to effect size, other factors should also be considered when interpreting bioassessment error rates and trade-offs in the context of environmental management decisions. Most importantly, decision-makers must consider the consequences of making a wrong judgement (Mapstone, 1995; Quinn & Keough, 2002). In cases where environmental protection is a primary goal, then the risk of making a Type II error is likely more significant than the risk of making a Type I error and bioassessments should be undertaken with a decision point that errs on the side of mistakenly judging a reference site to be impacted (i.e. a Type I error). This can be

easily justified when the consequences associated with undetected environmental damage (i.e. a Type II error) are serious and costly, but the implication of making a Type I error is often merely a requirement for more detailed review of the data and possibly additional assessment work to confirm the presence of an impact (Bailey et al., 2004). Data reviews and additional assessments can easily and often relatively cheaply be completed before activities are restricted or rehabilitation is undertaken.

In the mining industry, the perceived presence of an unacceptable environmental impact could lead to management actions such as installation of costly pollution control technologies, curtailment in production (and therefore loss of jobs and profits), and/or undertaking of expensive rehabilitation projects. Having to pursue any of these actions unnecessarily as a result of a Type I error is costly and not desirable from anyone's perspective. On the other hand, failing to take action in response to an undetected impact due to a Type II error could have devastating and irreversible impacts to critical resources such as drinking water supplies or salmon stocks. These examples illustrate that the potential costs to society of making judgement errors can be high and care must be taken when using bioassessment results in the mining sector.

This study revealed that Type I error rates are larger than expected. This means that decision-makers may encounter Type I errors more frequently than anticipated, and they should use caution when concluding that a site is not in reference condition. At α=0.10, anywhere from 0-26% of sites may be assessed as "not in reference condition" even though their watersheds are unaffected by human activities. If α is increased to 0.25, this jumps to 10-56%. These percentages reinforce the importance of conducting further confirmatory and diagnostic assessment work before making management decisions which will undoubtedly have costs associated with them - be they economic, social or political.

When used on their own, Skeena RCA bioassessments have a Type II error rate that may not meet environmental protection needs in northwestern British Columbia. For example, if assessments are conducted using the tools available in CABIN where $\alpha=0.10$, the power to detect moderate disturbances lies somewhere between 40 and 63% and the power to detect highly disturbed sites ranges from 70 to 100%. Missing a moderately disturbed site 37-60% of the time may be insufficient for ensuring sustainable development objectives, especially in instances where local environmental values are high such as in protected areas or around endangered and threatened species and ecosystems.

To help achieve environmental protection goals, RCA bioassessments should be combined with other impact assessment tools in a weight-of-evidence monitoring approach. Since they provide a relatively inexpensive and easy method to monitor the ecosystem itself, they should be considered as a central component of ongoing mining EEM programs that also include regular water and sediment quality monitoring, and periodic sampling of fish, plants and other ecosystem components. RCA bioassessments in the receiving environment below a mining discharge can be easily performed on a regular and defined schedule. If the assessment indicates that the site is in reference condition and other data supports this conclusion, it can be queued for routine follow-up testing in the future. If a site is deemed not to be in reference condition, further investigations of available data and/or additional experimentation can be conducted to confirm if this site is impacted and then to determine the cause of the impact. Since testing is relatively simple, repeat testing can also be used to monitor the success of remediation efforts and to track changes in environmental health over time (Perrin et al., 2007).

To minimize the likelihood of judgement errors and further improve our ability to monitor for impacts from the mining sector, the BMI data used for RCA bioassessments can also

be reviewed using other analytical techniques (such as the calculation of metrics and indices) to provide important additional interpretations. In addition, consideration should be given to increasing the sensitivity of CABIN RCA bioassessments by setting a more conservative decision point such as $\alpha=0.25$. At $\alpha=0.25$, Type I and Type II errors rates are more balanced, and for moderate disturbances the Type II error rates generally lie below Type I and the power of the assessments has increased to 60-86%. This will improve the chances that RCA bioassessments are able to provide an early warning of potential impacts.

Chapter 5 - Conclusions

The purpose of this research project was to analyze and summarize the performance of Skeena RCA bioassessments, focusing on their use in the mining sector. This was accomplished by creating a data set of sites with artificial impacts, and then testing whether or not the impacts were detected by the bioassessments. Type I and Type II error rates were determined for each group in the Skeena predictive model, and the trade-offs were examined and described in the context of resource management decision-making.

I found that actual Type I error rates in Skeena RCA bioassessments are higher than expected based on the Type I error probability (α) decision points set in the assessments. There is a trade-off between Type I and Type II error rates, so decision-makers must carefully consider the risks associated with each error type and design their bioassessments accordingly. Given the potential costs to society of making Type II errors, some situations may warrant the use of larger values of α (such as 0.25) and settling for a higher Type I error rate so that the Type II error rate is reduced.

My study also demonstrated that impact simulations can be used to evaluate bioassessment performance relatively easily. Error evaluation is recommended for all bioassessment tools, and future research in the areas of defining ecologically relevant impacts to BMI communities, and then simulating disturbances to represent them, will help to further enhance the value of this approach.

Skeena RCA bioassessments provide an efficient way to assess aquatic ecosystem health in the vicinity of mining activities by measuring the actual biota. However, decision-makers are reminded that like other assessment approaches, RCA bioassessments are most effective when

they are applied in conjunction with other monitoring tools in a “weight-of-evidence” approach to increase the likelihood of drawing correct conclusions about environmental impacts.

Works Cited

Aquatic Bioassessment Lab, 2003. *List of Californian macroinvertebrate taxa and standard taxonomic effort*. Retrieved from http://www.nps.gov/yose/naturescience/upload/Macroinvertebrates.2003.pdf.

Bailey, J. L., Reynoldson, T. B., & Bailey, R. C. (2007). *Reference condition approach bioassessment of Yukon River basin placer mining streams sampled in 2006.* Report prepared for the Mining and Petroleum Environment Research Group.

Bailey, J. L., Reynoldson, T. B., & Bailey, R. C. (2008). *Power, sensitivity and error using reference condition approach stream bioassessment methods: A Yukon example.* (in review).

Bailey, R. C., Kennedy, M. G., Dervish, M. Z., & Taylor, A. R. M. (1998). Biological assessment of freshwater ecosystems using a reference condition approach: Comparing predicted and actual benthic invertebrate communities in Yukon streams. *Freshwater Biology*, 39(4), 765-774.

Bailey, R. C., Norris, R. H., & Reynoldson, T. B. (2004). *Bioassessment of freshwater ecosystems using the reference condition approach*. Amsterdam: Kluwer.

Barbour, M. T., Gerritson, J., Snyder, B. D., & Stribling, J. B. (1999). *Rapid bioassessment protocols for use in streams and wadeable rivers: Periphyton, benthic macroinvertebrates and fish second edition*. Washington, USA: EPA.

Beasley, G. & Kneale, P. (2003). Investigating the influence of heavy metals on macroinvertebrate assemblages using partial canonical correspondence analysis (pCCA). *Hydrology and Earth System Sciences*, 7(2), 221-233.

Bennett, S. A. (2009). *Bioassessment of streams in Northwest BC using the Skeena BEAST09*. Report prepared by Bio Logic Consulting for Gitxsan Forest Enterprises and West Fraser Mills Ltd.

Bennett, S. A. (2010). *Bioassessment of streams in the Nass TSA using the reference condition approach*. Report prepared by Bio Logic Consulting for West Fraser Mills Ltd.

Blockstrom, K. A. & Winters, L. (2006). *The evaluation of methods for creating defensible, repeatable, objective and accurate tolerance values for aquatic taxa*. Ohio, USA: EPA.

Cao, Y. & Hawkins, C. (2005). Simulating biological impairment to evaluate the accuracy of ecological indicators. *Journal of Applied Ecology,* 42, 954–965.

Clements, W. H. (1994). Benthic invertebrate community responses to heavy metals in the Upper Arkansas River basin, Colorado. *Journal of the North American Benthological Society*, 13(1), 30-44.

Clements, W. H., Carlisle, D. M., Lazorchak, J. M., & Johnson, P. C. (2000). Heavy metals structure benthic communities in Colorado mountain streams. *Ecological Applications*, 10(2), 626-638.

Clements, W. H. (2004). Small scale experiments support causal relationships between meal contamination and macroinvertebrate community responses. *Ecological Applications*, 14(3), 954-967.

Environment Canada. (2002). *Metal mining guidance document for aquatic environmental effects monitoring*. Retrieved from http://www.ec.gc.ca/esee-eem/default.asp?lang=En&n=D450E00E-1.

Environment Canada. (2010). *Canadian Aquatic Biomonitoring Network (CABIN) website*. Retrieved from http://www.ec.gc.ca/rcba-cabin/.

Environmental Assessment Office. (2010a). *Environmental Assessment Office website*. Retrieved from http://www.eao.gov.bc.ca/index.html.

Environmental Assessment Office. (2010b). *Environmental Assessment Office user guide*. Victoria, Canada: B.C. Environmental Assessment Office.

Fredericks, J., Grieve, D., Lefebure, D., Madu, B., Northcote, B. & Wojdak, P. (2009). *British Columbia mining and mineral exploration overview 2009*. Victoria, Canada: B.C. Ministry of Energy, Mines and Petroleum Resources.

Gerhardt, A., Janssens de Bisthoven, L., & Soares, A. M. V. M. (2004). Macroinvertebrate response to acid mine drainage: Community metrics and on-line behavioural toxicity bioassay. *Environmental Pollution*, 130(2), 263-274.

Hilsenhoff, W. L. (1988). Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society*, 7(1), 65-68.

Hirst, H., Jüttner, I., & Ormerod, S. J. (2002). Comparing the responses of diatoms and macro-invertebrates to metals in upland streams of wales and cornwall. *Freshwater Biology*, 47(9), 1752-1765.

Huggins and Moffett. (1988). *Proposed biotic and habitat indices for use in Kansas streams*. Retrieved from: http://www.cpcb.ku.edu/research/assets/KBSRept35b.pdf.

International Network for Acid Prevention. (2009). *Global acid rock drainage guide*. Retrieved from http://www.gardguide.com.

Limnotek Research & Development. (1992). *Stream community responses to a gradient of acid mine drainage additions*. Prepared for British Columbia Acid Mine Drainage Task Force.

Linke, S., Norris, R. H. & Robinson, W. (2004). Impairator – a program to simulate the impacts of disturbance on biological assemblages. Version 1.0 [computer program]. University of Canberra: Canberra, Australia.

Mandaville, S. M. (2002). *Benthic macroinvertebrates in freshwaters – taxa tolerance values, metrics and protocols*. Retrieved from: http://www.chebucto.ns.ca/ccn/info/Science/SWCS/H-1/tolerance.pdf.

Mapstone, B. D. (1995). Scalable decision rules for environmental impact studies: effect size, type I and type II errors. *Ecological Applications*, 5(2), 401-410.

Maret, T. R., Cain, D. J., MacCoy, D. E., & Short, T. M. (2003). Response of benthic invertebrate assemblages to metal exposure and bioaccumulation associated with hard-rock mining in northwestern streams, USA. *Journal of the North American Benthological Society*, 22(4), 598-620.

Marqués, M. J., Martínez-Conde, E., & Rovira, J. V. (2003). Effects of zinc and lead mining on the benthic macroinvertebrates of a fluvial ecosystem. *Water, Air & Soil Pollution*, 148(1-4), 363-388.

Mazor, R. D., Reynoldson, T. B., Rosenberg, D. M., & Resh, V. H. (2006). Effects of biotic assemblage, classification, and assessment method on bioassessment performance. *Canadian Journal of Fisheries & Aquatic Sciences, 63*(2), 394-411.

Microsoft Excel (Version 2007) [Computer software]. Redmond, U.S.A.: Microsoft Corporation.

Millennium Ecosystem Assessment. (2005). *Ecosystems and human well-being: wetlands and water synthesis*. Washington, DC: World Resources Institute.

Mining Association of B.C. (2010). *Mining facts website*. Retrieved from http://www.mining.bc.ca/mining_facts.htm.

Ministry of Environment. 2007. *Environmental trends in British Columbia: 2007*. Victoria, Canada: B.C. Ministry of Environment.

Perrin, C. J., Bennett, S., Linke, S., Downie, A. J., Tamblyn, G., Ells, B., Sharpe, I., & Bailey, R. C. (2007). *Bioassessment of streams in north-central British Columbia using the reference condition approach*. Report prepared by Limnotek Research and Development Inc. and B.C. Ministry of Environment for the B.C. Forest Science Program. Retrieved from http://www.env.gov.bc.ca/epd/regions/skeena/water_quality/benthic/bio_streams_RCA_07.pdf

Pollard, A. I., & Yuan, L. (2006). Community response patterns: Evaluating benthic invertebrate composition in metal-polluted streams. *Ecological Applications*, 16(2), 645-655.

PRIMER (Version 6) [Computer software]. Plymouth, U.K.: Primer-E Ltd.

Quinn, G. P. & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

Reece, P. S. & Richardson, J.S. (2000). Biomonitoring with the reference condition approach for the detection of aquatic ecosystems at risk. In L. M. Darling (editor). *Proceedings of a conference on the biology and management of species and habitats at risk, Kamloops, B.C., 15 - 19 Feb.,1999. Volume Two*. (pp. 549-552). Victoria, Canada: B.C. Ministry of Environment.

Reynoldson, T. B., Bailey, R. C., Day, K. E., & Norris, R. H. (1995). Biological guidelines for freshwater sediment based on BEnthic assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. *Australian Journal of Ecology, 20*(1), 198-219.

Reynoldson, T. B., Norris, R. H., Resh, V. H., Day, K. E., & Rosenberg, D. M. (1997). The reference condition: A comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society, 16*(4), 833-852.

Rosenberg, D. & Resh, V. (1993). Introduction to freshwater biomonitoring and benthic macroinvertebrates. In D. Rosenberg & V. Resh (editors). *Freshwater biomonitoring and benthic invertebrates*. (pp. 1-9). New York: Routledge, Chapman & Hall.

Samuels, M.L. & Witmer, J.A. (2003). *Statistics for the life sciences.* New Jersey: Prentice Hall.

Smolders, A. J. P., Lock, R. A. C., Van, der Velde, G., Medina Hoyos, R. I., & Roelofs, J. G. M. (2003). Effects of mining activities on heavy metal concentrations in water, sediment, and macroinvertebrates in different reaches of the pilcomayo river, south america. *Archives of Environmental Contamination & Toxicology*, 44(3), 0314-323.

Systat (Version 13) [Computer software]. Chicago, U.S.A.: Systat Software Inc.

Tripole, S., Vallania, A. & del Carmen Corigliano, M. (2008). Benthic macroinvertebrate tolerance to water acidity in the Grande river sub-basin (San Luis, Argentina). *Limnetica*, 27 (1), 29-38.

Wojdak, P. (2010). Northwest Region. In *Exploration and mining in British Columbia 2009*. Victoria, Canada: B.C. Ministry of Energy, Mines and Petroleum Resources.

World Conservation Monitoring Centre. (1998). *Freshwater biodiversity: a preliminary global assessment*. Cambridge, UK: WCMC - World Conservation Press.

Wright, J. F. (1995). Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Australian Journal of Ecology*, 20(1), 181-19.