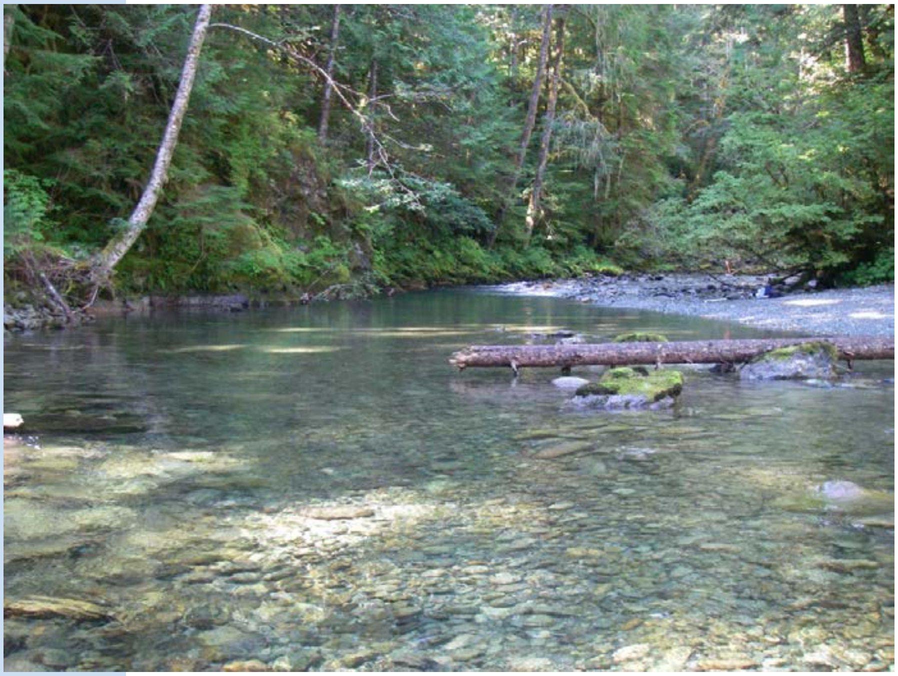


Evaluating Physical Habitat and Water Chemistry Data from Statewide Stream Monitoring Programs to Establish Least-Impacted Conditions in Washington State

DEPARTMENT OF
ECOLOGY
State of Washington

April 2015
Publication No. 15-03-011

Publication and contact information

This report is available on the Department of Ecology's website at
https://fortress.wa.gov/ecy/publications/SummaryPages/1503011.html

The Activity Tracker Code for this study is 08-532. The updated EIM Study IDs as of April 27, 2017 are WHM_BIO, WHM_ERR0, WHM_SEN, WHM_WAM0.

For more information contact:

Publications Coordinator
Environmental Assessment Program
P.O. Box 47600, Olympia, WA 98504-7600
Phone: (360) 407-6764

Washington State Department of Ecology - www.ecy.wa.gov

o Headquarters, Olympia	(360) 407-6000
o Northwest Regional Office, Bellevue	(425) 649-7000
o Southwest Regional Office, Olympia	(360) 407-6300
o Central Regional Office, Yakima	(509) 575-2490
o Eastern Regional Office, Spokane	(509) 329-3400

Evaluating Physical Habitat and Water Chemistry Data from Statewide Stream Monitoring Programs to Establish Least-Impacted Conditions in Washington State

by

Siri K. C. Wilmoth, MS, U.S. Geological Survey
Kathryn M. Irvine, PhD, U.S. Geological Survey
Chad A. Larson, PhD, Washington State Department of Ecology

U.S. Geological Survey
Northern Rocky Mountain Science Center
2327 University Way, Suite 2
Bozeman, MT 59715

Washington State Department of Ecology
Environmental Assessment Program
300 Desmond Dr SE
Olympia, WA 98504

This page is purposely left blank

Table of Contents

Page

List of Figures

Page

List of Tables

Abstract

An important step in the process of evaluating stream habitat condition in a state-wide survey is establishing criteria for reference or least-disturbed condition. Establishing these criteria sets a benchmark from which all other data is compared.

We evaluated data collected from the first four years of a state-wide stream monitoring survey in Washington State. This data set consisted of various GIS-generated, land-use predictor variables, physical habitat metrics, and water chemistry variables from 75 reference sites and 351 randomly sampled sites. We evaluated which of these predictor variables were effective at discriminating reference from random sites within level III ecoregions throughout the state using a combination of multivariate clustering and ordination techniques. We describe average observed conditions for a subset of predictor variables determined to be effective at discriminating reference and random sites.

We propose statistical criteria for establishing reference conditions for stream habitat in Washington, using these criteria to determine whether any of the random sites meet expectations for reference condition and whether any of the established reference sites fail to meet expectations for reference condition.

Acknowledgements

This work was funded by a cooperative agreement No. C1400001 between the Washington State Department of Ecology and the U.S. Geological Survey, Northern Rocky Mountain Science Center, to provide statistical support for the Biological Monitoring Program.

The authors thank many people, including Chris Hartman, who graciously compiled GIS data, and those involved in collecting the survey data: Glenn Merritt, Dan Dugger, Eiko Urmos-Berry, Mitch Wallace, Brian Gallagher, Jill Lemmon, Karen Adams, and Amnis Opes Institute, LLC.

Introduction

Background

Beginning in 2009, the Washington State Department of Ecology (Ecology) began extensive sampling efforts to measure habitat, water chemistry, and biological assemblages of streams throughout Washington. These efforts encompass three separate, but inclusive, projects in the Environmental Assessment Program, each employing the same methods and protocols.

- The first two, the Ambient Biological and the Sentinel projects, sample hand-picked, *least disturbed*, or reference sites, throughout the state, with Sentinel sites sampled annually. Ambient Biological sites are sampled on a rotating panel. Eight to ten sites are sampled within eight of Washington's nine Level III Omernik Ecoregions (Omernik, 1987), with two to three ecoregions sampled annually. Willamette Valley is excluded.
- The third project, Watershed Health Monitoring (WHM), uses a probabilistic sampling design where randomly selected sites stratified by stream order are sampled throughout the state on a rotating panel. Fifty sites are sampled per salmon recovery region, with one to two salmon recovery regions sampled annually. The probabilistic design is intended to ensure unbiased statistical inferences and representation of streams across the state, excluding federal and tribal lands.

During the selection of specific locations for hand-picked sites, a candidate list was made based on aerial photographs, a GIS pre-screening process, and, where possible, prior data and knowledge of the site. A general attempt was made to distribute sites spatially across the state, though no stratification scheme was used. All candidate sites were investigated by a field reconnaissance team. We considered these characteristics for all sites: hydrologic alteration and stream regulation; road densities; logging, mining, agricultural activities; and measures of urbanization at least 300 meters upstream and 50 meters downstream from each site (Bailey et al., 2004). Only those sites with the lowest level of anthropogenic stressors in good ecological condition were retained as best professional judgment (BPJ) reference sites. These sites provide the baseline against which the ecological conditions in all other streams in the region would be measured.

After four years of data collection, our aim was to decide which of the measured variables (e.g., GIS and site-specific measures) best characterize *least-disturbed* condition in Washington. This information will be used as a foundation for building habitat and bioassessment models that will then be used to assess the status and trends of streams throughout Washington.

The data represent 351 unique, random sites and 75 hand-picked sites categorized as reference by BPJ, as inventoried by Ecology. The number of unique reference sites varied from zero to 12 for each Omernik level III Ecoregion (Omernik and Gallant, 1986; Omernik, 1987; Pater et al., 1998; U.S. EPA., 2005). The number of random sites varied from 7 to 76 within each Omernik level III Ecoregion (Table 1). A limited number of sites were assessed twice each year to assess seasonal variation and sampling variability during the summer-fall index period. From these sites with repeat visit data, just one visit was randomly selected for modeling purposes. Sites are plotted at their locations across Washington in Figure 1.

Table 1. The number of unique sites in each of Washington's nine Omernik level III Ecoregions (U.S. EPA, 2005), by site type (random or reference).

Ecoregion	Site Type	
	Random	Reference
Blue Mountains	10	9
Cascades	36	10
Coast Range	57	9
Columbia Plateau	76	4
Eastern Cascades Slopes and Foothills	26	10
North Cascades	51	11
Northern Rockies	46	12
Puget Lowland	42	10
Willamette Valley	7	0
Total	351	75

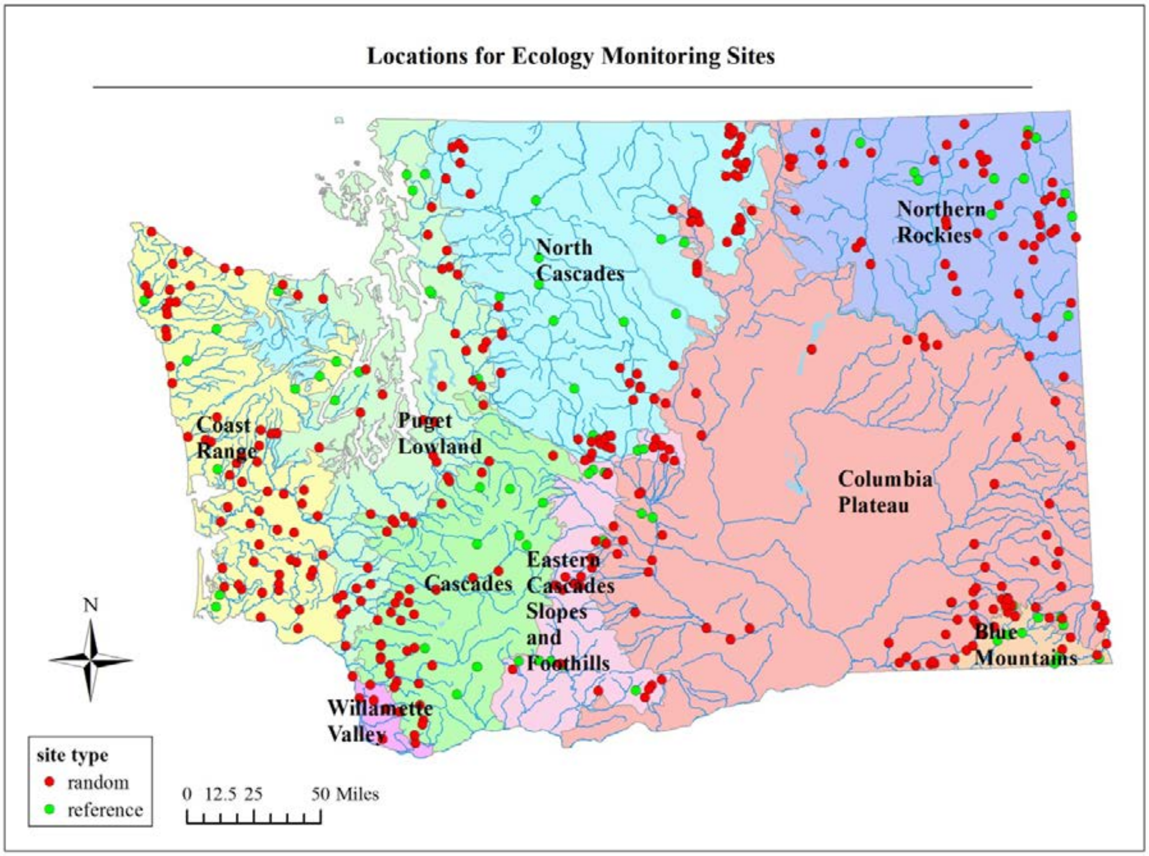


Figure 1. Site locations sampled by Ecology's Ambient Biological, Sentinel, and Watershed Health Monitoring project by ecoregion and site type (reference vs. random).

Data encompass sites sampled from 2009 through 2012

Reference sites were distributed across a broad geographic area of Washington, encompassing distinct ecological conditions and disturbance regimes. West of the Cascade Range (Coast Range, Puget Lowland, Cascades, and Willamette Valley Ecoregions), reference sites on the wetter west side had a median precipitation of 90.7 cm/year, median elevation of 152 m, and median percent slope of 36.28%. Urbanization ranged from minimal to extensive, with over 3.6 million inhabitants within the Seattle-Tacoma-Bellevue metropolitan area (U.S. Census Bureau, 2014). This region was also characterized by high road densities, development, and impervious cover. East of the Cascades (Blue Mountains, Eastern Cascades, North Cascades, Northern Rockies, and Columbia Plateau Ecoregions), reference sites had much lower median precipitation levels (34.65 cm/year), higher median elevations (769 m), and lower median slope (33.84%). The eastern half of the state was typically characterized by cultivated crops, pasture, and grassland, and has roughly one-third the population of western Washington (U.S. OFM, 2011). These distinct natural settings and the extent and type of human disturbance were important factors in setting expectations for reference conditions.

A common approach to setting expectations for reference conditions is to use filtering criteria. Filtering criteria are specific thresholds of physical, chemical, or biological variables that reference sites must be below or above. For example, sites with road density greater than a certain value within a catchment are filtered (removed) from the list of candidate sites in least-disturbed condition. Waite et al. (2000) determined specific values of chemical water quality parameters and physical habitat (Rapid Bioassessment Protocol score) to screen-out impacted sites from a candidate set of 577 probabilistic sites. Waite et al. (2000) used a single set of criteria for a complex and heterogeneous landscape (the Mid-Atlantic Highlands), which contains parts of eight distinct Omernik Level III ecoregions (Davis and Scott, 2000). Similarly, Herlihy et al. (2008) set chemical and physical-habitat filters to identify least-disturbed sites in a nationwide bioassessment of 1,655 sites. In the western states, Whittier et al. (2007) used water-quality, physical-habitat, and disturbance metrics to select a list of least-disturbed sites, to examine the assumption that BPJ sites are all in least-disturbed condition. Whittier et al. (2006) and Whittier et al. (2007) specified variable filtering criteria based on relationships with a natural gradient (e.g., slope, elevation).

Our general approach differs from that of previous investigators who impose thresholds and filtering criteria to define the set of least-disturbed sites (Waite et al., 2000; Herlihy et al., 2008). As a first step, we used a regionalization scheme to attempt to minimize the naturally occurring variation in environmental variables (Hawkins et al., 2000). Then, similar to Herlihy et al. (2008), we used chemical and physical habitat variables to identify least-disturbed sites from among the BPJ and random locations. The method of identifying these least-disturbed sites is where our philosophy differs. Our approach relies on the assumption that existing BPJ sites represent least-disturbed condition. By characterizing the physical and chemical environment at the majority of BPJ sites with multivariate ordination and clustering techniques, we can define statistical criteria for least-disturbed conditions at these sites. Moving forward, these environmental characteristics can then be used to identify new locations that also meet these criteria. Although natural differences, e.g., erosion potential and geology, within ecoregions exist, we think that this characterization of sites using selected physical habitat and chemistry predictors yields a reasonably broad depiction of the environment of a site.

Objectives

The objectives of this report are to describe least-disturbed condition, defined by Stoddard et al. (2006) as "the best available physical, chemical and biological habitat conditions given today's state of the landscape." Specifically, we:

1. Describe the environmental characteristics at a set of ambient and sentinel sites, chosen by BPJ. Use these to define statistical criteria for least-disturbed condition based on catchment, reach-level and point metrics for GIS, habitat, and chemistry data.

2. Determine which random sites, if any, meet defined statistical criteria for least-disturbed condition. Use in conjunction with BPJ reference sites for subsequent development of multimetric indices, e.g., B-IBI (Karr, 1991) and multivariate bioassessment models, e.g., RIVPACS (Wright, 1995).

3. Identify reference sites dissimilar in multivariate environmental space to the majority of BPJ reference sites. Investigate these further for their status as least-disturbed condition.

Statistical Methods

Variable reduction

Finding the best set of variables to describe reference condition was an iterative process, but it was driven by *a priori knowledge* and the perception of biologists. A candidate set of metrics was filtered based on skewness, univariate normality, and amount of missing data. We took care with correlated variables, noting the results of Ketchen and Shook (1996) who found that correlated metrics tended to skew clustering solutions in the direction of that construct. When correlated predictors were retained, we verified that this underlying construct either directly represented a disturbance metric (e.g., road density), or had a substantial influence on macroinvertebrate communities (e.g., percent sand/fines, embeddedness and median particle size). See Richards et al., 1993; Wood and Armitage, 1997; Bilotta and Brazier, 2008; and many other sources. We used analysis of variance (ANOVA) on site type (random vs. reference), blocked by ecoregion, to determine a variable's efficacy at differentiating between site type after accounting for ecoregion effects. We interpreted ANOVA results with a conservative bonferroni approach.

Metric repeatability also played a role in the variable reduction process. We examined an additional subset of data containing repeat site visits to compute the coefficient of variation (CV), a dimensionless measure of variability: $CV = \left(\frac{sd(x)}{mean(x)}\right)$. We followed the guidelines determined by Roper et al. (2010), who suggested that variables with a CV > 35% have low consistency, adding random noise to analyses. For comparison purposes, we also provide the signal-to-noise (S:N) ratio, a precision metric outlined in Kaufmann et al. (1999). The S:N ratio compares the within-year variance among streams (signal) relative to the variance between repeat stream visits within the same year (noise) from a random-effects ANOVA model for the probability sample streams. The higher the value of S:N, the more useful the predictor may be for discerning trends or changes. For example, an S:N ratio of 1 indicates that the within-year variation among streams is equal to the variation between repeat visits. Roper et al. (2010) suggest that the likelihood of detecting environmental heterogeneity is low when the S:N ratio is less than 2.5, moderate when the S:N ratio is between 2.5 and 6.5, and high when the ratio is greater than 6.5.

Appropriate spatial extent

To determine the appropriate spatial extent for the analysis, we pooled all observations for a statewide analysis and examined ordination and cluster membership for geographic patterns. We used a multivariate ANOVA (MANOVA) to examine differences in multivariate means between geographic regions and to test the sufficiency of a statewide analysis. Then, we applied Dixon's nearest-neighbor method (Dixon, 1994) to the scaled site scores from a principal components ordination along the first two axes. Dixon's method compares the geographic region of each site with the geographic region of its nearest neighbor in Euclidean principal component analysis space, constructing a contingency table by site type. Expected cell counts are calculated under a null hypothesis of random labeling and depend only on the numbers of each type (N_i), and the total number of sites (N). For example, for type *i* sites, the expected number of nearest neighbors of type *j* is given by $[N_{ij}] = N_i \frac{N_j}{(N-1)}$ if $i \neq j$, or alternatively $[N_{ij}] = N_i \frac{N_j - 1}{(N-1)}$ if $i = j$ (Dixon, 2002). Variances of cell counts under the null hypothesis of random labeling for two groups are as follows (for extensions to more than 2 groups, see Dixon, 1994; and Dixon, 2002):

$$Var[N_{ij}] = \begin{cases} (N+R)P_{aa} + (2N-2R+Q)P_{aaa} + (N^2-3N-Q+R)P_{aaaa} - (NP_{aa})^2 & \text{for } i=j \\ NP_{ab} + QP_{aab} + (N^2-3N-Q+R)P_{aabb} - (NP_{ab})^2 & \text{for } i \neq j \end{cases}$$

where $N, R,$ and Q are quantities derived from the set of locations such that N is the total number of points, R is the number of reflexive nearest neighbors, and Q is the number of shared neighbors. Probabilities (P_{ij}) are those for random permutations. Under the null hypothesis of random labeling, the test statistic is $z_{ij} = \frac{N_{ij}-E[N_{ij}]}{\sqrt{Var[N_{ij}]}}$, assuming an asymptotic normal distribution with mean 0 and variance 1.

Multivariate ordination and clustering

Principal component analysis (PCA) and *k*-means clustering were used in tandem to produce plot overlays and confirm ordination results. The two methods use similar approaches and are both based on Euclidean distance, thus reinforcing conclusions. For PCA, only complete cases were used, and variables were transformed to optimize univariate normality (either $log_{10}(x+1)$ for volume and density data, or the $arcsin\sqrt{x}$ for proportion data; Sokal and Rohlf, 1995). Variables were centered and scaled $\left(\frac{x-\bar{x}}{s_x}\right)$, and PCA scores and variable loadings (eigenvectors) were computed in R. For *k*-means clustering, data were ranged prior to analysis (Milligan and Cooper, 1988; Everitt and Hothorn, 2011), such that each observation of variable x_j had range $[c, c+1]$: $x_{rge} = \frac{x_i}{x_{max}-x_{min}}$. We used 200 random starting locations for each *k*-means clustering run due to algorithm sensitivities to local minima. We selected the number of groups (*k*) to which to allocate sites based on the algorithm's ability to allocate the maximum number of reference sites to a single cluster while also discriminating between reference and random locations. Clusters were renamed according to the number of reference sites within each group, such that cluster 1 always contained the bulk of the reference sites and cluster *k* the fewest.

To determine sites most similar to the bulk of the BPJ reference sites, we drew a convex hull around reference sites within cluster 1 in the two-dimensional space created by site ordinations along the first and second principal components (PC1 and PC2). Cluster 1 sites within and outside of the convex hull (both reference and random) were recorded, while those sites within the convex hull were used to set the expectations for reference condition. This method allowed us to hone in on the (presumed) best sites and directly combine ordination and clustering procedures. MANOVA tests examined the evidence for differences in multivariate means between the proposed groups.

We used Fisher's Linear Discriminant Analysis (LDA, Fisher, 1936) to find the linear combination of variables that best separated the groups derived from clustering, PCA and convex hull inclusion. We computed discriminant functions to determine the relative importance of metrics using standardized data $\left(\frac{x-\bar{x}}{s_x}\right)$ with the lda() function in the MASS package (R Core Team, 2014). Each site was ordinated in the space of the linear discriminant axes, and the discriminant function coefficients from standardized data were interpreted as each variable's relative contribution to the discrimination between site types (Borcard et al., 2011). The identification (or classification) functions were computed with the original (*not* standardized) descriptors. These functions are used to find the group with which a new site likely belongs (Legendre and Legendre, 1998). There are as many identification functions as there are groups, and the functions are of the form:

$$S_i = c_i + w_{i1} \times x_1 + w_{i2} \times x_2 + \cdots + w_{ip} \times x_p$$

where S_i denotes a new site's identification score for group i; c_i denotes a constant for the ith group; w_{ij} denotes the weight for the ith group and the jth variable for $j = 1,2, \dots p$; and x_i denotes the observed value for the jth variable of the new site. We attribute a new site to the group for which it has the highest identification score. If agreement is sufficient, then the identification function can be used to assess new sites for potential inclusion into the ambient/sentinel monitoring program, and complex analyses do not have to be replicated.

We used repeat visits to a unique location as hold-out data to examine the accuracy of site allocation using the identification function. Here, we treated group membership derived from PCA, clustering and convex hull inclusion as the true group membership and compared the predicted group derived from the identification function.

Results

Variable reduction

GIS data

The data set contained 59 different GIS metrics, which were reduced to a set of 6 candidate metrics. Many metrics were not considered for ordination because they did not meet statistical assumptions, such as skewed distributions with a large number of sites with small values and a very few with high values (e.g., primary highway road densities or high-intensity development). Several metrics were redundant, since they represented the same measure at different spatial scales (e.g., impervious area for catchment vs. a 100-m stream-level buffer). However, we considered two metrics for road density (both secondary highway and unimproved roads).

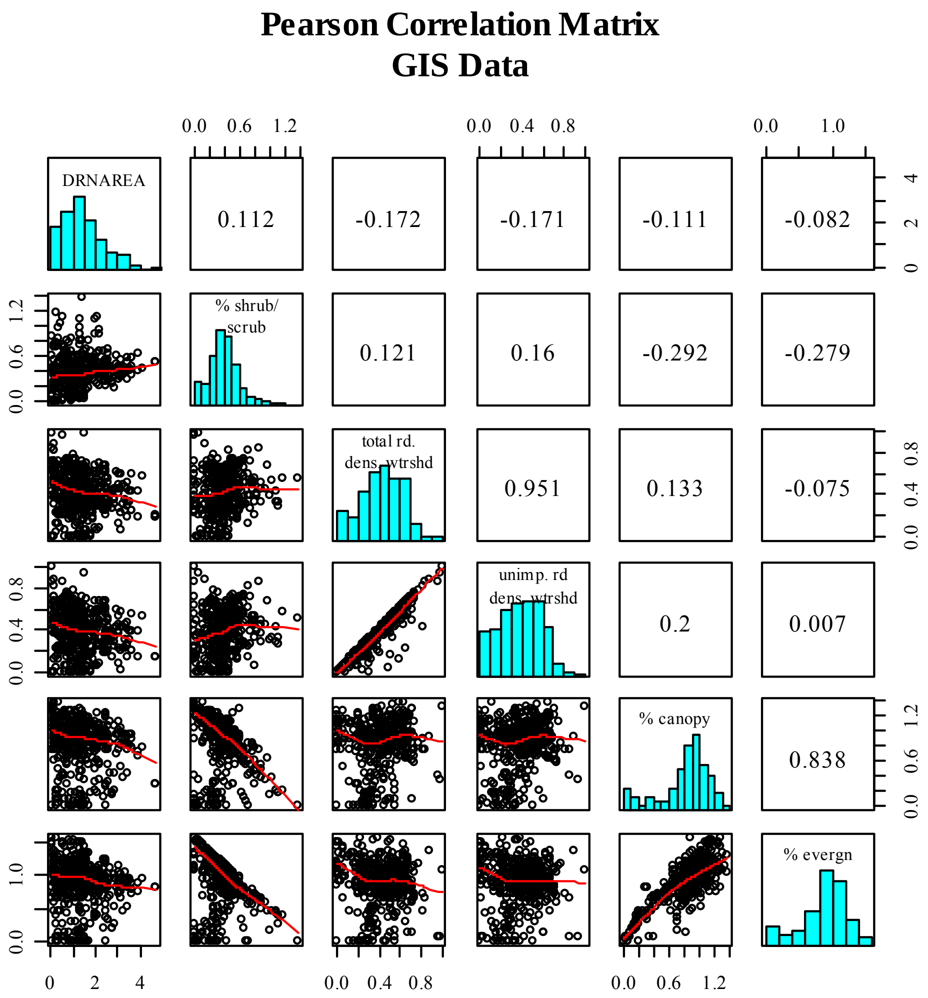
Six GIS metrics exhibited approximately normal distributions, and a two-way ANOVA revealed that 5 of the 6 had a significant effect of site type (random vs. reference), after accounting for ecoregion and differing type effects between ecoregions (interaction term). The bonferroni adjusted *p*-values for the effects of ecoregion, site type (random or reference) and an interaction between the two are in Table 2.

Table 2. Bonferroni adjusted (n=18) p-values for selected GIS metrics from an ANOVA for site type (reference vs. random) after accounting for ecoregion and an interaction between ecoregion and type.

Analysis included all sites across Washington. For metric definitions, see Appendix A.

Selected GIS Metrics	**Ecoregion**	**Type**	**Ecoregion: Type**
$\arcsin\sqrt{\%\text{ canopy}}$	< 0.0001	0.0002	1
$\arcsin\sqrt{\%\text{ evergreen}}$	< 0.0001	< 0.0001	1
$\arcsin\sqrt{\%\text{ shrub/scrub}}$	0.0001	0.0011	0.0019
$\log_{10}[\text{drainage area}]$	< 0.0001	1	1
$\log_{10}[\text{Total Unimp. Rds-wtrshd} + 1]$	< 0.0001	< 0.0001	< 0.0001
$\log_{10}[\text{Total Rds-wtrshd} + 1]$	< 0.0001	< 0.0001	< 0.0001

Figure 2 illustrates the univariate and bivariate distributions for these 6 GIS metrics. These and other metrics of *a priori* interest are plotted by ecoregion in Appendix C.



Pearson Correlation Matrix
GIS Data

Figure 2. Pairwise scatterplots and histograms for 6 selected GIS metrics, as well as their Pearson correlation coefficients.

A smoothing line (lowess) is added to bivariate plots.

Habitat data

Similar to the GIS data, redundancy played a big role in variable reduction of the candidate pool of 260 habitat and reach-level metrics. Ninety-four metrics for large woody debris (LWD) varied by counts/volumes and various size classes. The most generic of these was selected for the volume of total LWD per 100 m (Robison, 1997). Metrics for fine particles were of *a priori* interest, and four of the 39 total metrics for substrate were tested with an ANOVA for their ability to discriminate between reference and random sites (Table 3). Despite their correlations, we tested each of these substrate variables because of their varying quantitative and qualitative methods. This resulted in different reference vs. random effects. Riparian cover and disturbance metrics were also of interest *a priori*, though some variables were not selected because of missing data. For example, shade at the center of the stream was missing for 60 sites. Data were inconsistent. See Appendix B for example of IDX.Canopy that had a median CV of 33% for repeat visits. Other variables were not easy for us to differentiate between random and reference sites after accounting for the effect of Ecoregion. See Table 3 for results from a two-way ANOVA. Metrics are plotted in Figure 3.

Table 3. Bonferroni adjusted (n=18) p-values for selected habitat and reach-level metrics from an ANOVA for type (reference vs. random) after accounting for ecoregion and an interaction.

Analysis included all sites across Washington. For metric definitions, see Appendix A.

Selected Habitat and Reach-level Metrics	Ecoregion	Type	Ecoregion: Type
$\arcsin\sqrt{\%\ \text{sand}}$	< 0.0001	1	0.0512
$\arcsin\sqrt{\%\ \text{sand/fines}}$	< 0.0001	< 0.0001	0.8180
$\arcsin\sqrt{\%\ \text{embeddedness}}$	< 0.0001	0.1974	0.3616
Median Substrate Size	< 0.0001	0.0002	1
Bare Ground Cover	< 0.0001	< 0.0001	0.6100
$\log_{10}[\text{LWD vol.} + 1]$	< 0.0001	< 0.0001	1
$\log_{10}[\text{PWP. All} + 1]$	< 0.0001	< 0.0001	1
$\log_{10}[\text{Relative Bed Stability}]$	< 0.0001	0.2044	1

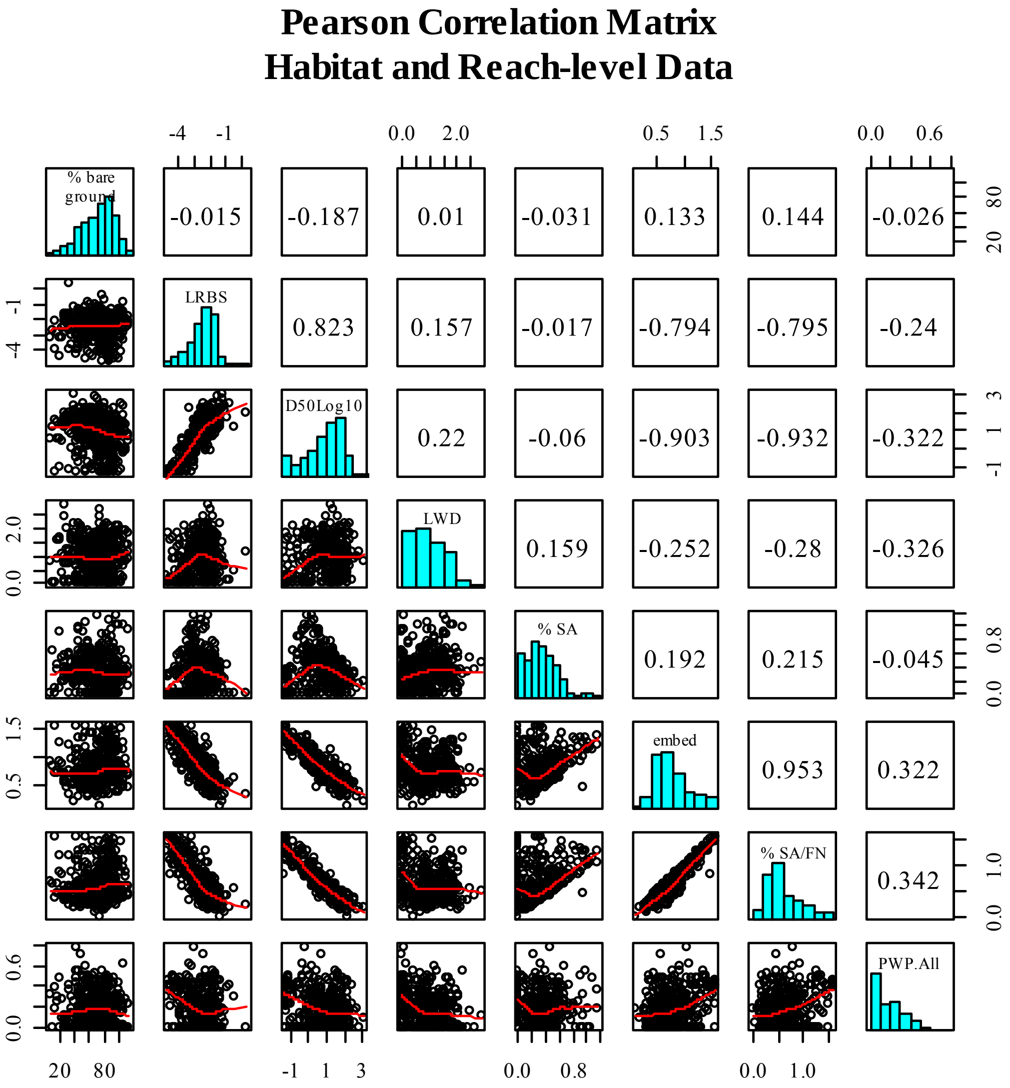

Pearson Correlation Matrix
Habitat and Reach-level Data

Figure 3. Pairwise scatterplots and histograms for 8 selected reach-level and habitat metrics, as well as their Pearson correlational coefficients.

A smoothing line (lowess) is added to scatterplots.

Chemistry data

The chemistry data were made up of 56 metrics of water and sediment chemistry, and variable reduction was largely guided by the quantity of missing data. Fifteen variables were missing data for more than 100 sites, and only 10 variables were missing data for fewer than 45 sites. Of those 10 remaining variables, one was redundant (two measures for dissolved oxygen), and another was inconsistent (turbidity had a median CV of 63% for repeat visits, shown in Appendix B). The remaining candidate metrics are plotted in Figure 4.

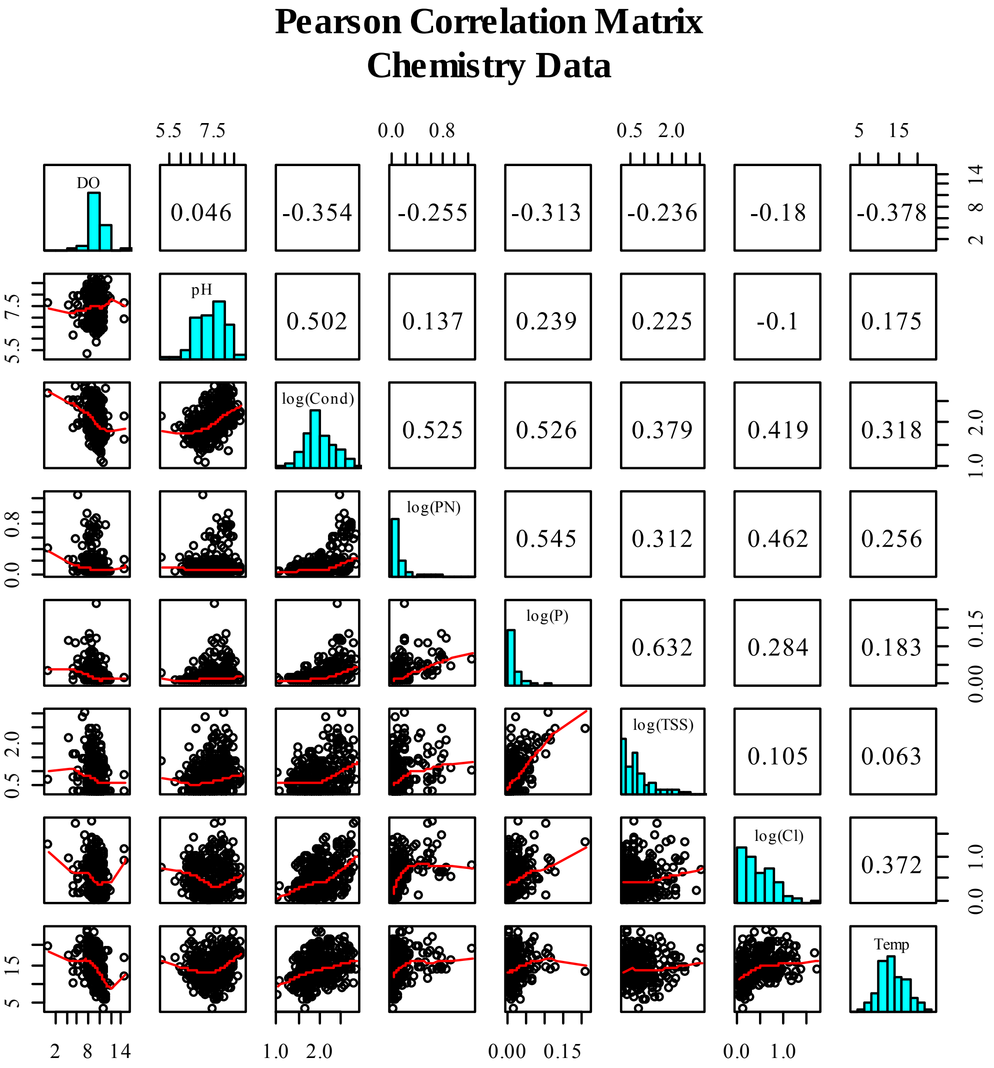


Figure 4. Pairwise scatterplots and histograms for 8 selected chemistry variables, as well as their Pearson correlational coefficients.

A smoothing line (lowess) is added to scatterplots.

Only 4 of this set of 8 variables met the (approximate) normality requirements for ANOVA, and only two (dissolved oxygen and temperature) showed evidence of an effect of site type (random vs. reference) after accounting for the effect of ecoregion (Table 4).

Table 4. Bonferroni adjusted p-values (n=18) for selected chemistry variables from an ANOVA for type (reference vs. random) after accounting for ecoregion and an ecoregion by type interaction.

Analysis included all sites across Washington. For metric definitions, see Appendix A.

Selected Chemistry Variables	Ecoregion	Type	Ecoregion: Type
DO (mg/L)	0.0005	0.0009	1
pH	< 0.0001	1	1
Water Temp. ($°C$)	< 0.0001	< 0.0001	1
$\log_{10}$[Conductivity]	< 0.0001	1	1

Appropriate spatial extent

Results from statewide clustering and ordination revealed underlying regional similarities.

Sites in the Columbia Plateau Ecoregion appear systematically different from sites in Washington's other 8 ecoregions. Random locations in the Columbia Plateau Ecoregion occupy their own distinct space opposite the bulk of the reference sites in cluster 1 (red crosses marked *5* in Figure 5, opposite green circles) and account for 59 of the 74 sites allocated to cluster 3 (Table 5). The principal component eigenvectors associated with this space are greater human disturbance (PWP.All), higher water temperatures, and increased sand/fines, with lower amounts of cover (evergreen cover, canopy cover and volume of LWD in-stream), lower levels of dissolved oxygen, and lower road densities. When arrayed in the space of the first two principal components, there was significant statistical evidence for this spatial segregation of Columbia Plateau sites. Columbia Plateau sites tended to be their own nearest neighbor in PC space ($z = 9.75$, $p < 0.0001$, from Dixon's contingency table test comparing expected and observed counts of nearest neighbors by type, as shown in Table 6. Rather than being interspersed with other sites, they are a group unto themselves, justifying the use of a separate model and separate expectations for least-disturbed condition.

There were also differences between sites in the western and eastern regions of Washington. For example, 60% of sites grouped with cluster 2 were in Western Washington (116 of 194, Table 5): 61% of sites grouped with cluster 1 were in Eastern Washington (86 of 140, Table 5). This is readily seen on a map displaying sites with their cluster membership (Figure 6), where there are more triangles (cluster 2 sites) in the Western side of the state and more circles (cluster 1 sites) in the Eastern side of the state. Again, the exception is the Columbia Plateau, shown with red crosses (cluster 3 sites) scattered throughout.

The principal component eigenvectors (Figure 5) associated with cluster 1 sites (mostly Eastern Washington sites) were: larger substrate sizes (i.e., larger D50Log10, less sand/fines), more dissolved oxygen, greater canopy and evergreen cover, and more LWD. Also, cluster 1 sites had lower road densities, less bare ground, lower water temperatures, and less human disturbance (PWP.All). The principal component eigenvectors associated with cluster 2 sites (mostly Western Washington sites) were: higher road densities, increased bare ground cover, and to some extent more LWD and greater canopy cover.

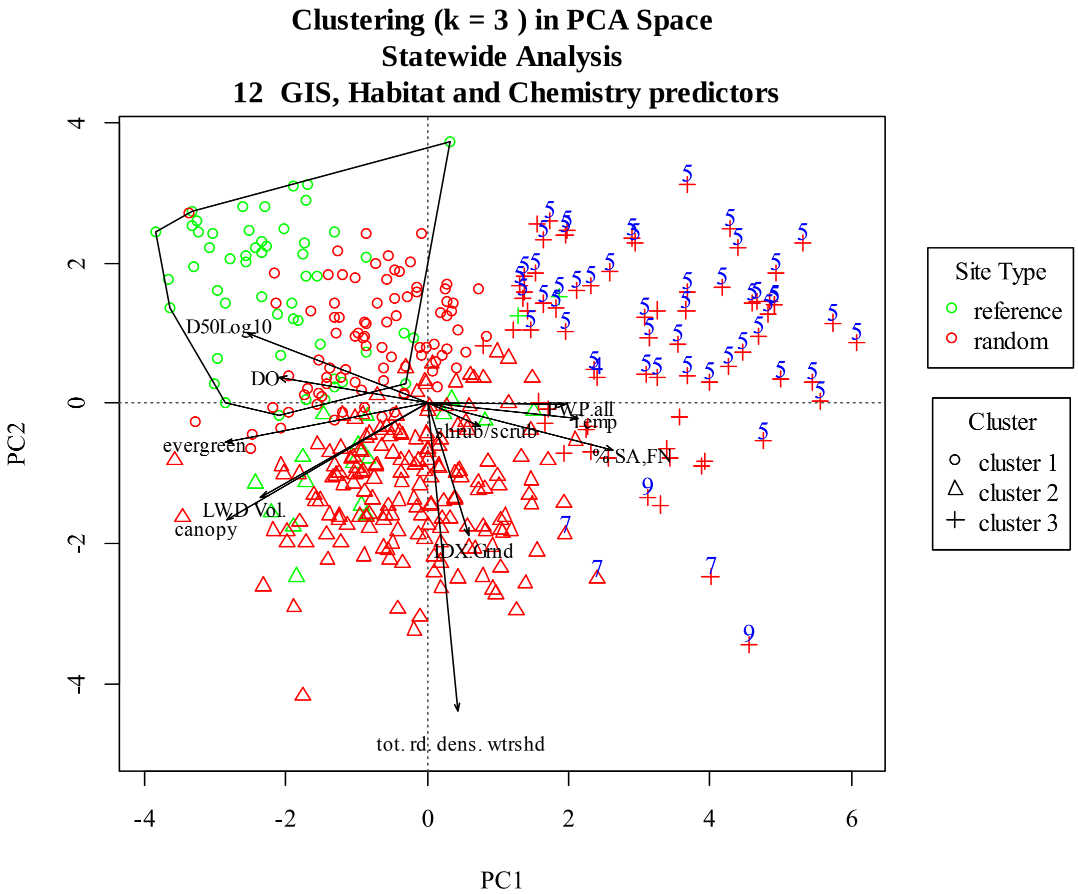


Figure 5. The distribution of individual sites in principal component space (PC1 and PC2).

Color indicates site type, and symbols indicate cluster. Selected points are labeled by ecoregion (5 for Columbia Plateau; 7 for Puget Lowland; and 9 for Willamette Valley). The convex hull marks the boundary in PC space of all reference sites assigned to cluster 1 (k-means clustering for k=3). Arrows represent variable loadings (eigenvectors) for the first 2 PCs. One vector is not pictured here (road 4 density, identical to total road density). PCA and cluster analysis were of 12 GIS, habitat, and chemistry metrics jointly. The first two PCs explain 0.4892 of their variability.

Table 5. Group membership of sites to clusters from a k-means cluster analysis (k=3) of 336 random and 72 reference sites across 9 ecoregions in Washington.

Cluster analysis was for 12 GIS, habitat, and chemistry metrics jointly.

Ecoregion	**Random Sites**			**Reference Sites**		
	1	**2**	**3**	**1**	**2**	**3**
Blue Mountains	7	2	1	7	2	0
Cascades	7	29	0	10	0	0
Coast Range	15	40	0	4	4	0
Columbia Plateau	13	5	57	1	1	2
Eastern Cascades	7	17	1	7	3	0
North Cascades	24	21	1	11	0	0
Northern Rockies	13	26	7	10	1	0
Puget Lowland	1	32	3	3	6	0
Willamette Valley	0	5	2	0	0	0
Total	87	177	72	53	17	2

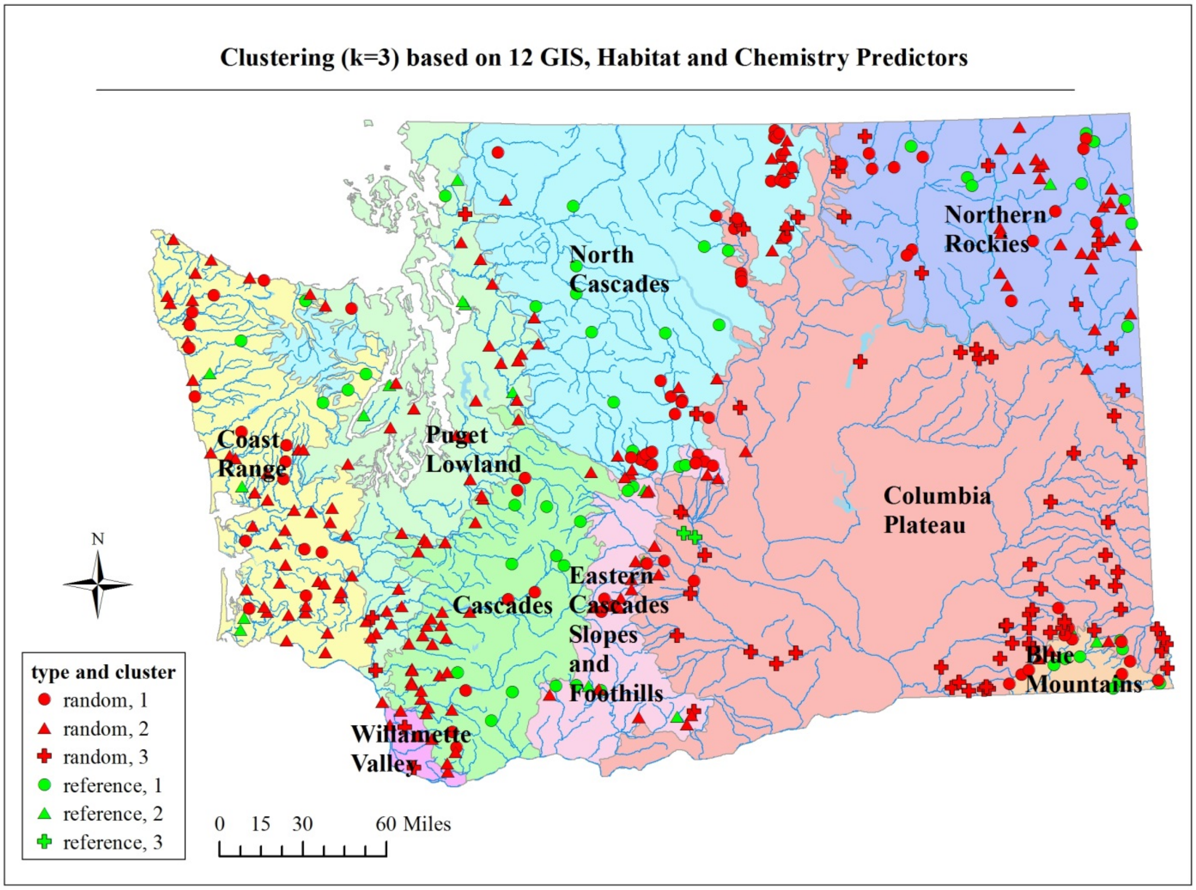


Figure 6. Cluster membership of random and reference sites by geographic location.

Cluster analysis (k-means clustering for k=3) was of 12 GIS, habitat, and chemistry metrics

An overall test of spatial segregation provided irrefutable evidence against the null hypothesis of random labeling of points (i.e., region) in principal components space ($p < 0.00001$, $\chi^2=123.52$, $df=6$, from an overall segregation test; Dixon, 1994). For both Western and Eastern Washington, there was very strong statistical evidence against the null hypothesis of random labeling of nearest neighbor sites arrayed in principal component space. In other words, sites in Western Washington tended to have PCA ordination scores that immediately juxtaposed them with other sites from Western Washington ($z = 5.31$, $p < 0.0001$, from Dixon's contingency table test comparing expected and observed counts of nearest neighbors by type, as shown in Table 6). More concretely, sites in Western Washington had 100 observed nearest neighbor sites also from Western Washington from their ordination in PCA space, when only 63.3 were expected (Table 6). Given its asymptotic normal distribution (Dixon, 2002), this provides strong evidence against the null hypothesis of random labeling. Similarly, for Eastern Washington there was also strong evidence that nearest neighbor sites tended to be from the same region ($z = 2.13$, $p = 0.0166$, from Dixon's contingency table test comparing expected and observed counts of nearest neighbors by type, as shown in Table 6).

Table 6. Results for region-specific tests of spatial segregation for sites using Dixon's (2002) nearest neighborhood method in PCA space.

First two principal components are from analysis of 12 GIS, habitat, and chemistry metrics jointly. Region West includes Coast Range, Puget Lowland, Willamette Valley and Cascades Ecoregions. Region East includes Eastern Cascades Slopes and Foothills, North Cascades, Northern Rockies and Blue Mountains Ecoregions. Columbia includes only Columbia Plateau Ecoregion.

From (*i*)	**To (*j*)**	**Obs. Count**	**Exp. Count**	z_{ij}	**Significance**
West	West	100	63.3	5.31	*
	East	57	66.5	-1.49	
	Columbia	4	31.3	-5.64	*
East	West	68	66.5	0.24	
	East	84	68.9	2.13	*
	Columbia	16	32.6	-3.37	*
Columbia	West	5	31.3	-6.00	*
	East	18	32.6	-3.31	*
	Columbia	56	15.1	9.75	*

Gathered together, multivariate analyses and tests of spatial segregation validate the use of a regional model to best accommodate the geographic similarity in the data. A multivariate ANOVA (MANOVA) provided overwhelming evidence against the null hypothesis of pooling sites across the state, suggesting that the alternative (regional models) would be more appropriate (Wilk's lambda = 0.1641, approximate F-statistic of 48.21 with 24 and 405 df, $p < 0.0001$). Separate models based on ecoregion could also remedy problems with autocorrelation, but we chose to model the expectations for least-disturbed conditions by region rather than by ecoregion for two reasons. First, pooling across a larger geographical area maximizes sample sizes and statistical power. And perhaps most importantly, there were *a priori* biological reasons for partitioning sites by region, due to their distinct climatic and disturbance regimes.

Variable reduction by region

To select variables for separate regional analyses, a two-way ANOVA was used to determine which variables had significant difference in means between reference and random sites after accounting for ecoregion effects. For the Columbia Plateau, only a type effect was calculated, since it represents a single ecoregion. ANOVA results are displayed in Table 7 with the significance levels of selected variables highlighted. Note that all but the Columbia Plateau have been adjusted for multiple testing errors. The subsets of variables selected for the western and eastern regions are similar, but the Columbia Plateau is best described by a somewhat different set of variables (Table 7).

Table 7. By region, the bonferroni adjusted p-values from a two-way ANOVA (n=19) for type (reference vs. random) and ecoregion effects for selected variables.

Interaction effects were also calculated but are not listed here. For Columbia Plateau, results from a one-way ANOVA were not bonferroni adjusted. The p-values highlighted in gray correspond to the variables selected for analyses; metrics were selected when significant at the α = 0.10 level.

Variable	**West**		**East**		**Columbia**
	Ecoregion	**type**	**Ecoregion**	**Type**	**Type**
$\arcsin\sqrt{\%\text{ canopy}}$ (catchment)	< 0.0001	0.1327	0.0004	0.0094	0.3361
$\arcsin\sqrt{\%\text{ evergreen}}$	< 0.0001	0.0003	< 0.0001	0.0086	0.1700
$\arcsin\sqrt{\%\text{ canopy}}$ (reach)	0.0112	1	1	0.9873	0.3931
$\arcsin\sqrt{\%\text{ sand}}$	1	1	0.0087	1	0.5965
$\arcsin\sqrt{\%\text{ sand/fines}}$	0.0077	0.1024	0.0014	0.0587	0.0302
$\arcsin\sqrt{\%\text{ shrub/scrub}}$	1	0.0059	0.0000	0.0001	0.0160
$\arcsin\sqrt{\%\text{ embeddedness}}$	0.0013	0.5081	0.0014	1	0.0748
Median Substrate Size	0.0169	0.0992	0.0001	0.0832	0.0548
DO (mg/L)	0.0047	0.0836	1	0.0026	0.3480
Bare Ground Cover	0.4825	0.0429	1	0.0001	0.6081
$\log_{10}$[Conductivity]	< 0.0001	1	1	1	0.2775
$\log_{10}$[drainage area]	0.0045	1	1	1	0.8535
$\log_{10}$[LWD vol. + 1]	0.0694	0.0262	1	0.0010	0.0114
PWP.All	0.0003	0.3279	0.6876	0.0014	0.0270
$\log_{10}$[Unimp. Rds-wtrsd+1]	1	< 0.0001	0.9069	<0.0001	0.2492
$\log_{10}$[Total Rds-wtrsh + 1]	0.1022	< 0.0001	0.2932	<0.0001	0.6486
LRBS	1	0.3633	< 0.0001	1	0.1260
pH	0.7471	1	1	1	0.7906
Water Temp. ($^\circ C$)	< 0.0001	0.0044	0.1722	0.0133	0.0691

Reference conditions by region

Western Washington

Ordination and clustering revealed which sites were most similar to the BPJ reference sites. Reference sites had similar ordination scores along PC1 and PC2 (shown in Figure 7 plotted in green), and nearly two-thirds (17 of 27) of all reference sites were grouped with cluster 1. See Table 8. In addition, a total of 23 random sites were grouped with cluster 1 and had similar conditions for the selected variables. Six of these 23 random sites were within the convex hull and were most similar to the reference sites with respect to the selected environmental variables. These are listed as candidates for least-disturbed condition (Figure 7 and Appendix F).

The principal component eigenvectors associated with cluster 1 were primarily low road densities, greater evergreen cover and in-stream LWD, less bare ground, and lower water temperatures. These sites also generally had higher levels of dissolved oxygen (DO) and fewer sand/fines (i.e., higher D50Log10 values). Cluster 2 was associated with more sand/fines (i.e., lower D50Log10 values), higher road densities, and lower levels of DO. Cluster 3 was associated with higher water temperatures, higher road densities, increased bare ground, less evergreen cover and less in-stream LWD.

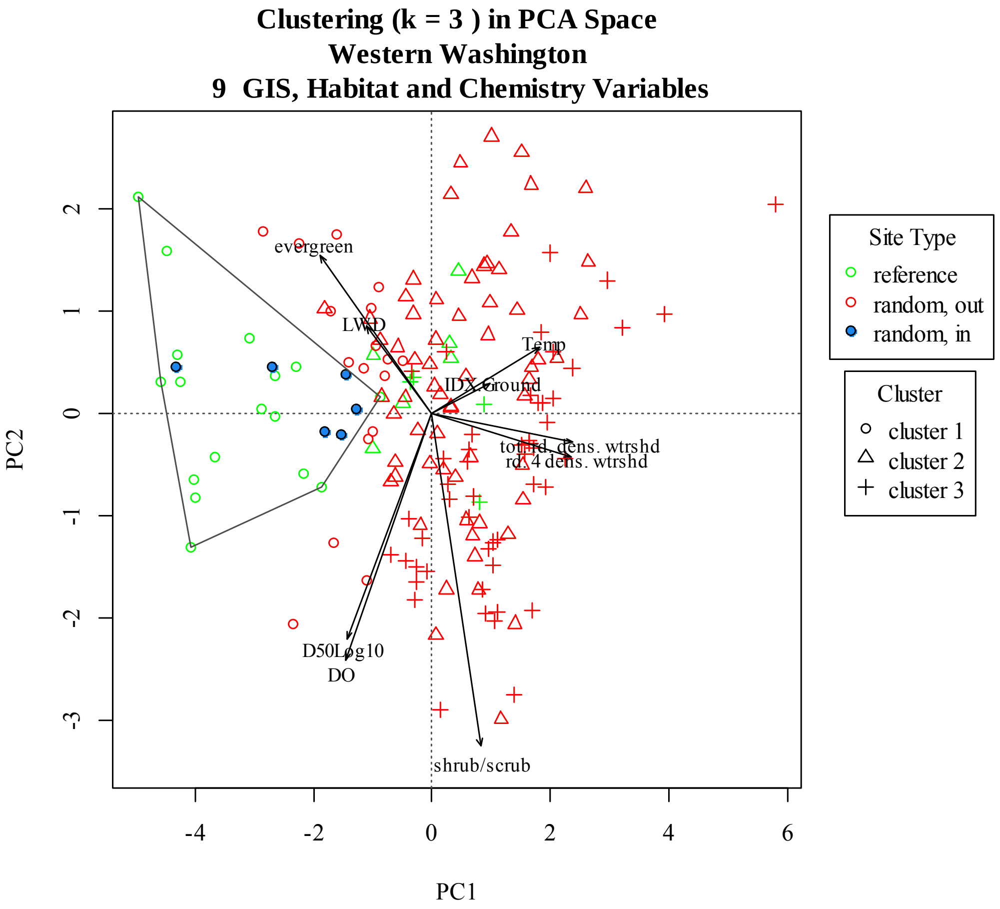


Figure 7. For Western Washington, the distribution of individual sites in principal component space (PC1 and PC2).

Color indicates site type, and symbols indicate cluster. The convex hull marks the boundary in PC space of all reference sites assigned to cluster 1 (k-means clustering, k=3). Arrows represent variable loadings (eigenvectors) for the first 2 PCs. PCA and cluster analysis were of 9 GIS, habitat, and chemistry variables jointly, and the first 2 PCs explain 0.492 of their variability.

Table 8. Group membership of sites to clusters from a k-means cluster analysis (k=3) of 136 random and 27 reference sites across 4 ecoregions in Western Washington.

Sites with missing data were excluded. Cluster analysis was for 9 GIS, habitat, and chemistry metrics jointly.

Ecoregion	**Random Sites**			**Reference Sites**		
	1	**2**	**3**	**1**	**2**	**3**
Cascades	7	15	14	10	0	0
Coast Range	13	30	12	4	4	0
Puget Lowland	3	16	19	3	2	4
Willamette Valley	0	2	5	0	0	0
Total	23	63	50	17	6	4

A linear discriminant analysis of the standardized site data from Western Washington clearly separated groups defined by PCA, clustering and inclusion within the convex hull (Figure 8). There was overwhelming evidence against the null hypothesis of equal means, suggesting the spatial separation of these four groups in multivariate space ($p < 0.0001$, from a multivariate ANOVA, Wilk's lambda = 0.0974, approximate F-statistic of 19.96 with 27 and 442 *df*). The main separators of least-disturbed sites were disturbance metrics. Sites within the convex hull of reference sites for cluster 1 ($n = 23$; Table 9) were characterized primarily by lower road densities but also by higher levels of evergreen forest and LWD. Cluster 1 sites not included within the convex hull ($n = 17$) generally had more bare ground, higher water temperatures and road densities, and smaller substrate sizes (D50Log10) than those within the convex hull. Sites in clusters 2 and 3, to varying degrees, had greater road densities, less evergreen cover and more shrub/scrub cover, more bare ground and smaller particles (D50Log10), as well as less dissolved oxygen and increased water temperatures (Table 10). Particularly in Western Washington, percent shrub/scrub cover can be an indication of harvested timber lands.

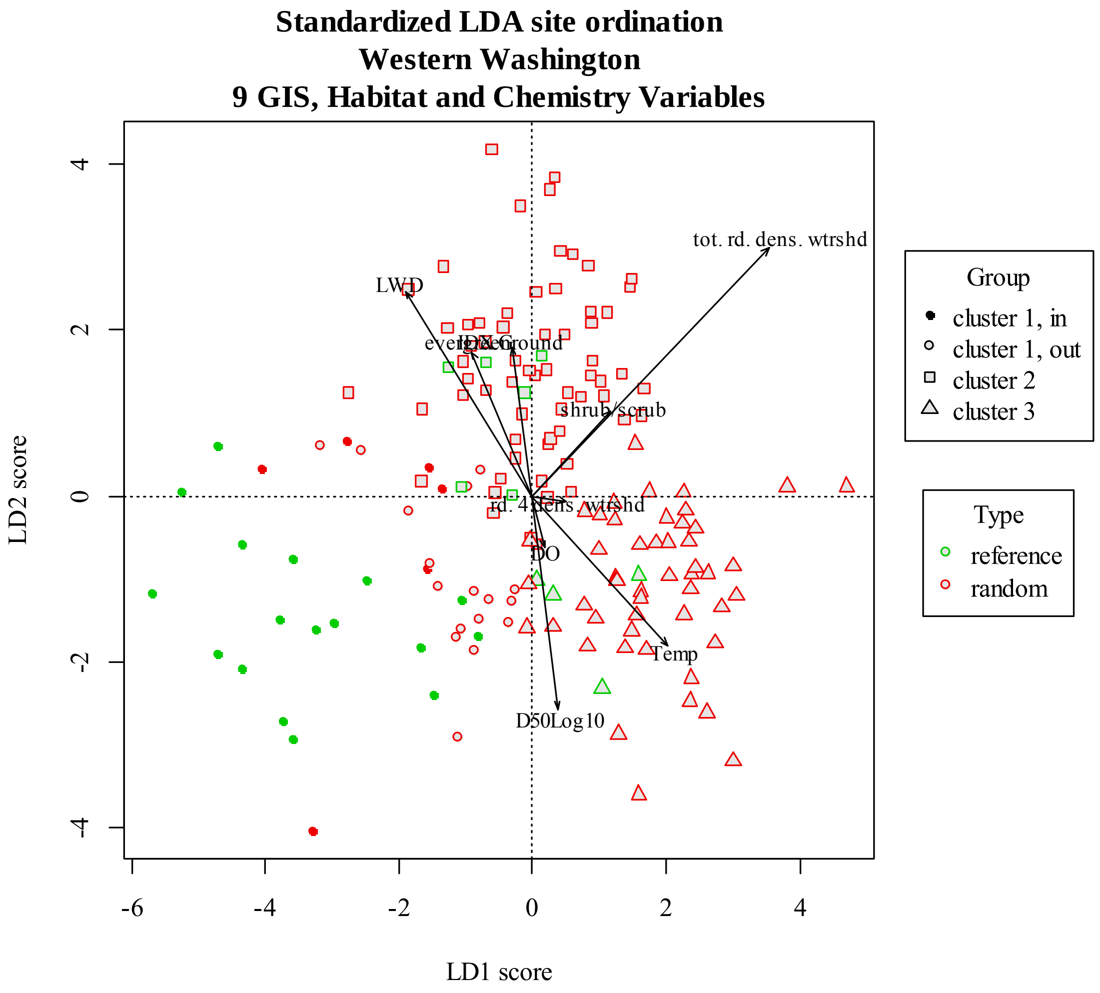


Figure 8. Linear Discriminant Analysis of standardized site data for the Western Washington sites by group.

This is derived from clustering and principal component ordination using a convex hull around the bulk of the reference sites to determine which sites were within (in) the hull, and which were not (out).

The Fisher's linear discriminant coefficients (Table 10) provide a predictive tool to classify a new site using the 9 specified GIS, habitat, and chemistry metrics.

Table 9. For Western Washington, group means for 9 GIS, habitat, and chemistry metrics.

For comparison, group means for all BPJ reference sites and all random sites are provided first (including sites missing data for some metrics). Means are on the transformed scale, where applicable.

Metric extent	**Metric**	**Western Washington**					
		BPJ ref.	**random**	**cluster 1, in**	**cluster 1, out**	**cluster 2**	**cluster 3**
		n=29	*n*=141	*n*=23	*n*=17	*n*=69	*n*=54
Catchment	$\arcsin\sqrt{\%\ \text{evergreen}}$	1.07	0.88	1.15	1.10	0.93	0.75
Reach	D50Log10	1.57	0.91	1.75	1.32	0.54	1.23
Point	DO (mg/L)	10.27	9.67	10.38	10.11	9.54	9.62
Reach	$\log_{10}$[LWD vol.+1]	1.48	1.19	1.56	1.20	1.53	0.73
Reach	Bare Ground	69.40	81.02	67.04	71.92	88.30	75.21
Point	Water Temp. $(^\circ C)$	11.75	14.32	10.57	13.69	13.29	16.15
Catchment	$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	0.26	0.39	0.27	0.26	0.39	0.42
Catchment	$\log_{10}[\text{Total Rds.} + 1]$	0.31	0.55	0.16	0.36	0.59	0.61
Catchment	$\log_{10}[\text{Unimp. Rds.} + 1]$	0.29	0.51	0.15	0.31	0.55	0.57

We illustrate the application of the identification function (Table 10) for Western Washington. A reference site in the Cascades Ecoregion (BIO06600-BIGC04) was surveyed multiple times. When building the linear discriminant predictive model, the site was ordinated within the convex hull of cluster 1 reference sites based on a randomly selected survey visit (Sept. 20, 2012). Modeled and predictive survey data from the same season are displayed in Table 11, and the scores for each group are given in Table 12.

Table 10. Fisher's Linear Discriminant Identification function coefficients for classification of a new site in Western Washington.

Fisher's Linear Discriminant Identification Function Coefficients	cluster 1, in	cluster 1, out	cluster 2	cluster 3
Intercept	-107.171	-119.102	-126.872	-126.483
$\log_{10}[\text{LWD vol.} + 1]$	7.365	5.395	7.642	2.779
$\arcsin\sqrt{\%\ \text{evergreen}}$	24.858	26.320	27.691	20.925
DO (mg/L)	12.274	12.615	12.056	12.551
D50Log10	-3.575	-3.945	-5.424	-3.171
Water Temp. $(°C)$	3.306	3.809	3.489	4.242
$\log_{10}[\text{Total Rds-wtrshd} + 1]$	45.936	65.926	75.350	72.651
$\log_{10}[\text{Unimp. Rds-wtrshd} + 1]$	-27.702	-35.408	-28.283	-26.187
$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	17.248	20.659	26.931	26.010
IDX.Ground	0.152	0.147	0.209	0.132

Table 11. Survey data for reference site BIO06600- BIGC04 in the Cascades Ecoregion, Western Washington.

Repeat visit data were randomly selected from all visits for modeling (9/20/2012) and predictive purposes (8/28/2012).

Variable	Modeling Survey Data 9/20/2012	Predictive Survey Data 8/28/2012
$\log_{10}[\text{LWD vol.} + 1]$	1.12	1.74
$\arcsin\sqrt{\%\ \text{evergreen}}$	1.17	1.17
DO (mg/L)	10.15	10.45
D50Log10	2.84	2.77
Water Temp. $(°C)$	10.2	11.0
$\log_{10}[\text{Total Rds-wtrshd} + 1]$	0.433	0.433
$\log_{10}[\text{Unimp. Rds-wtrshd} + 1]$	0.428	0.428
$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	0.310	0.310
IDX.Ground	68.4	68.9

Table 12. Scores for each group from Fisher's Linear Discriminant Identification Functions for reference site BIO06600- BIGC04 in the Cascades Ecoregion of Western Washington.

Scores are based on repeat survey data from Aug. 28, 2012.

Site Name	Score cluster 1, in	Score cluster 1, out	Score cluster 2	Score cluster 3
BIO06600-BIGC04	113.33	113.83	111.42	109.31

Based on group scores, the predictive data place this site in the *cluster 1, out* group, though just barely. During the modeling exercise, this same site was grouped within the convex hull of reference sites for cluster 1. The discrepancy is small. Minor changes in reach- and point-level metrics weren't entirely responsible for its exclusion from the group within the convex hull. Some of these changes were: increase in water temperature, minor increase in bare ground, and minor decrease in substrate size (D50Log10). Observed values of LWD increased, as did DO. These metrics have moderate-to-high consistency between repeat visits (CV < 29% and S:N ratio > 3.1 for each variable). See Appendix B for details. We feel confident with this predictive classification. Catchment-level metrics (i.e., road density and cover) have not changed, but this site has higher road densities than reference sites in Western Washington. Table 9 shows 0.31 for the mean of all $\log_{10}[\text{Total Rds-wtrshd} + 1]$ for reference sites in Western Washington. Since road densities were a decisive factor in discriminating group membership, this site already represented the fringe of the distribution for least-disturbed conditions. This is likely why small changes in other disturbance factors allocated this BPJ site to a different group (*cluster 1, out*).

For all sites in Western Washington with repeat visit data, the identification function correctly predicted group membership for 20 of the 27 sites. All misclassifications occurred between adjacent categories. For example, 4 sites modeled within the convex hull of cluster 1 were placed in the *cluster 1, out* group; 1 site modeled within cluster 2 was placed in the *cluster 1, out* group; and 2 sites modeled within cluster 2 were placed in the cluster 3 group.

Eastern Washington

Multivariate ordination and clustering were very effective at grouping reference sites for Eastern Washington. Nearly two-thirds of all the reference sites in Eastern Washington (Table 13) were allocated to cluster 1 and appear within the convex hull, arrayed in the negative space for the first principle component (Figure 9). There were also 11 random sites co-occurring within this convex hull of reference sites in PCA space, depicting their similarity with respect to these metrics and thus candidates for least-disturbed condition. Appendix F lists these least-disturbed sites. Conversely, reference sites spatially distant from this convex hull may not represent least-disturbed condition, since they are dissimilar in this multivariate environmental space.

Table 13. Group membership of sites to clusters from a *k*-means cluster analysis (*k*=3) of 127 random and 41 reference sites across 4 ecoregions in Eastern Washington.

Sites with missing data were excluded. Cluster analysis was for 12 GIS, habitat, and chemistry metrics jointly.

Ecoregion	**Random Sites**			**Reference Sites**		
	1	**2**	**3**	**1**	**2**	**3**
Blue Mountains	2	7	1	3	5	1
Eastern Cascades	3	19	3	6	3	1
North Cascades	9	17	20	10	0	1
Northern Rockies	8	7	31	8	0	3
Total	22	50	55	27	8	6

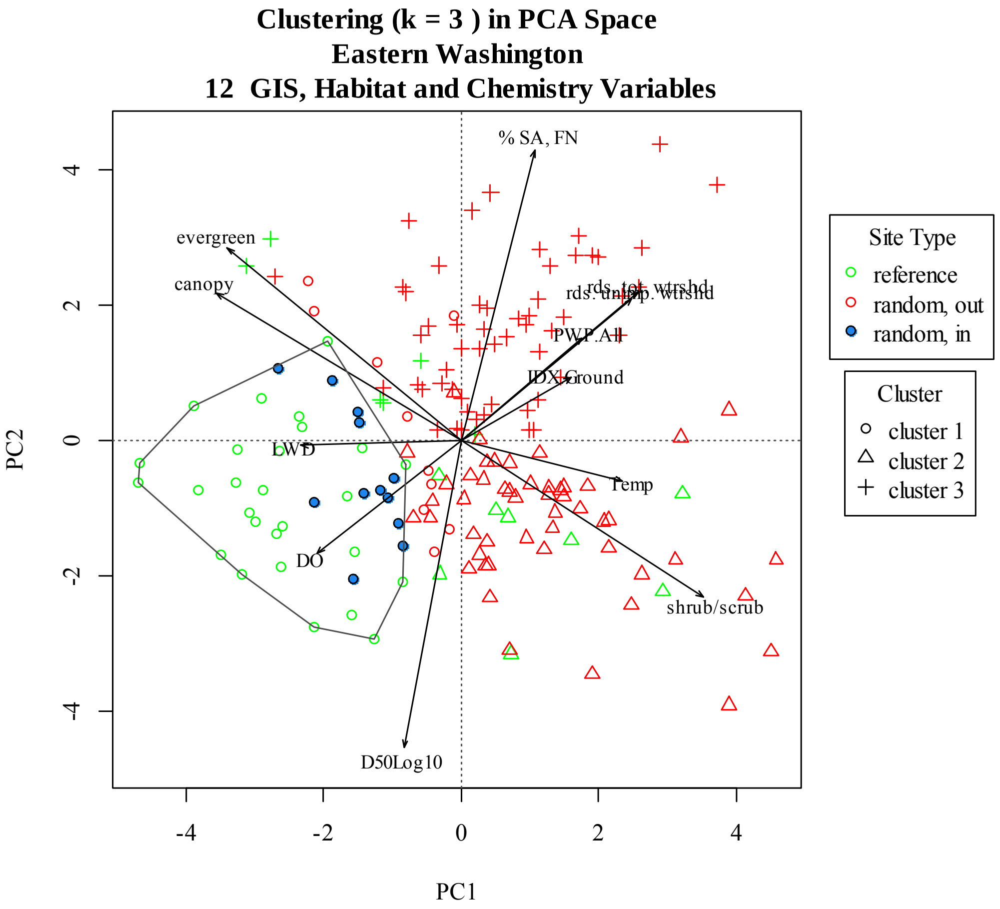


Figure 9. For Eastern Washington, the distribution of individual sites in principal component space (PC1 and PC2).

Color indicates site type and symbols indicate cluster.

The convex hull marks the boundary in PC space of all reference sites assigned to cluster 1 (k-means clustering for k=3). Arrows represent variable loadings (eigenvectors) for the first 2 PCs. PCA and cluster analysis were of 12 GIS, habitat, and chemistry variables jointly, and the first 2 PCs explain 0.506 of their variability.

The principal component eigenvectors associated with cluster 1 and the bulk of the reference sites are as follows: high levels of LWD and DO, and high levels of canopy and evergreen cover (Figure 9). These sites were also associated with larger substrate sizes (high D50Log10, low percent sand/fines). Sites allocated to cluster 1 also had low road densities, low human disturbance levels and low bare ground cover. Cluster 2 sites, by contrast, were associated with lower levels of evergreen and canopy cover and higher levels of shrub/scrub cover, lower levels of DO, higher water temperatures, and increased bare ground. Cluster 3 sites were characterized by high levels of sand fines (small D50Log10 values) and bare ground, high road densities, and increased human disturbance (PWP.All).

A linear discriminant analysis of the standardized site data from Eastern Washington clearly separated groups defined by PCA, clustering and inclusion within the convex hull (Figure 10). There was overwhelming evidence against the null hypothesis of equal means, suggesting the spatial separation of these four groups in multivariate space ($p < 0.0001$, from a multivariate ANOVA, Wilk's lambda = 0.0808, approximate F-statistic of 16.90 with 36 and 453 *df*). Sites within the convex hull of reference sites for cluster 1 ($n = 39$; Table 14) were characterized primarily by low road densities, low water temperatures, and low sand/fines, but also by higher levels of evergreen and forest cover. Cluster 1 sites not included within the convex hull ($n = 10$) generally had less LWD, smaller substrate sizes (smaller D50Log10 and larger percent sand/fines), more bare ground, higher water temperatures and less DO than those within the convex hull. Sites in clusters 2 and 3, most notably, had greater road densities, increased water temperatures, and greater sand/fines than sites grouped with cluster 1. (See Table 14.)

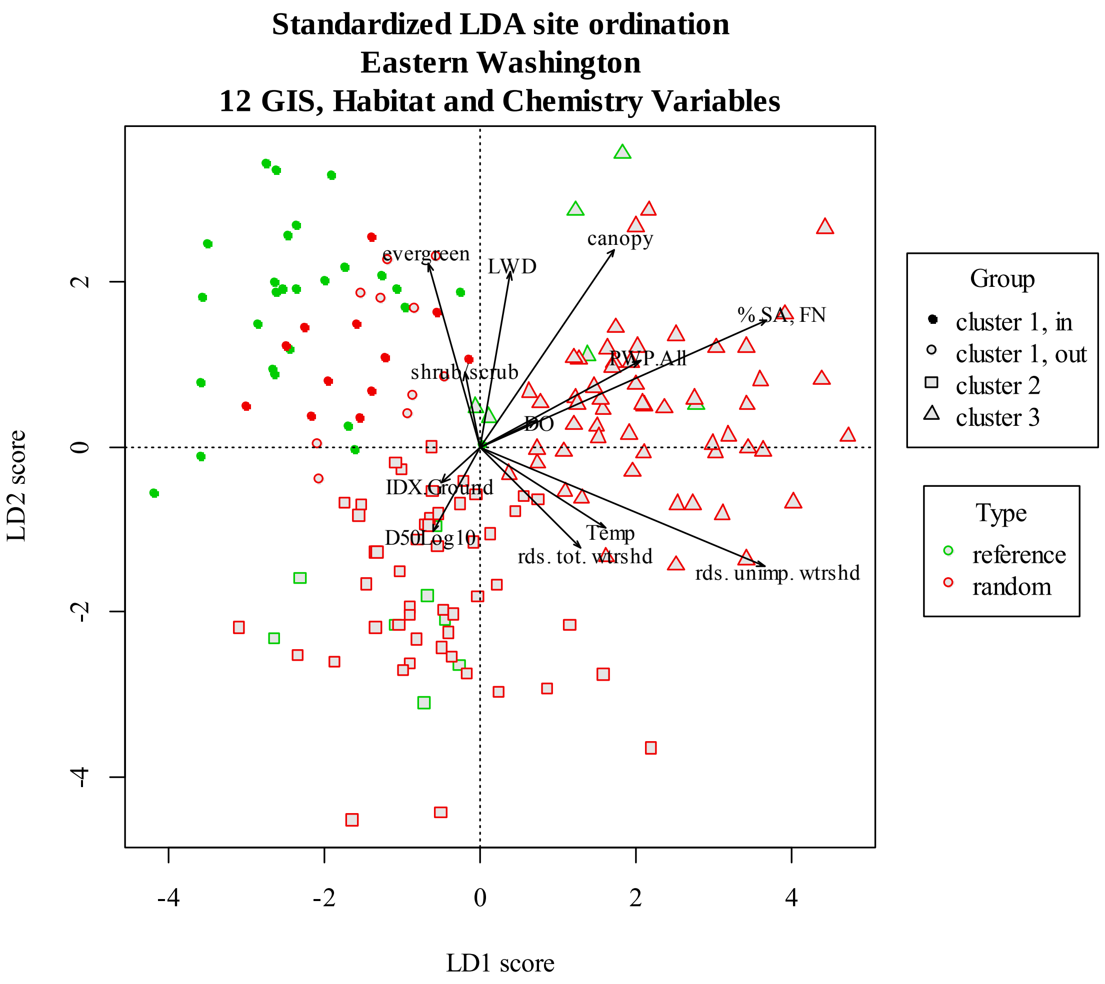


Figure 10. Linear Discriminant Analysis of standardized site data for the Eastern Washington sites by group.

This is derived from clustering and principal component ordination using a convex hull around the bulk of the reference sites to determine which sites were within (in) the hull, and which were not (out).

Table 14. For Eastern Washington, group means for 12 GIS, habitat, and chemistry metrics. *For comparison, group means for BPJ reference sites and for all random sites are provided first and include sites missing data for some metrics). Means are on the transformed scale, where applicable.*

Metric extent	Metric	Eastern Washington					
		BPJ ref.	**random**	**cluster 1, in**	**cluster 1, out**	**cluster 2**	**cluster 3**
		n=42	*n*=133	*n*=39	*n*=10	*n*=58	*n*=61
Catchment	$\arcsin\sqrt{\%\text{ canopy}}$	0.91	0.82	0.96	0.94	0.69	0.90
Catchment	$\arcsin\sqrt{\%\text{ evergreen}}$	1.12	1.01	1.18	1.12	0.85	1.12
Reach	D50Log10	1.37	0.94	1.44	0.72	1.52	0.32
Point	DO (mg/L)	10.08	9.36	10.07	8.96	9.47	9.32
Reach	$\log_{10}[\text{LWD vol.} + 1]$	1.17	0.87	1.27	0.92	0.65	0.95
Reach	IDX.Ground	51.48	68.29	56.69	73.88	66.97	65.32
Point	Water Temp. ($°C$)	10.72	12.91	10.15	11.22	13.95	12.38
Catchment	$\arcsin\sqrt{\%\text{ shrub/scrub}}$	0.30	0.44	0.28	0.37	0.53	0.36
Catchment	$\arcsin\sqrt{\%\text{ sand/fines}}$	0.46	0.63	0.43	0.66	0.43	0.85
Catchment	$\log_{10}[\text{Total Rds.} + 1]$	0.24	0.42	0.15	0.21	0.40	0.52
Catchment	$\log_{10}[\text{Unimp. Rds.} + 1]$	0.20	0.40	0.12	0.16	0.37	0.49
Reach	$\log_{10}[\text{PWP.All} + 1]$	0.06	0.18	0.06	0.10	0.13	0.25

The identification function correctly predicted group membership for 15 of 19 sites from the holdout, repeat visit data. Only one of the four misclassifications was not between adjacent groups. Two sites modeled within the convex hull of cluster 1 were predicted in other groups (*cluster 1, out* and *cluster 3*). One site modeled in cluster 2 was predicted to belong to *cluster 3*. And one site from cluster 3 was predicted to belong to cluster 2. For prediction of group membership of a future site in Eastern Washington (excluding the Columbia Plateau), Table 15 lists the Fisher's linear discriminant coefficients.

Table 15. Fisher's Linear Discriminant Analysis identification function coefficients for classification of a new site in Eastern Washington.

Fisher's Linear Discriminant Coefficients	**Group**			
	cluster 1, in	**cluster 1, out**	**cluster 2**	**cluster 3**
Intercept	-309.758	-305.594	-299.855	-337.365
$\log_{10}[\text{LWD vol.} + 1]$	11.987	10.454	8.598	11.628
$\arcsin\sqrt{\%\text{ canopy}}$	57.780	64.368	51.342	65.456
$\arcsin\sqrt{\%\text{ evergreen}}$	143.773	141.557	134.382	138.121
DO (mg/L)	18.129	17.410	18.006	18.679
D50Log10	31.443	29.420	31.857	30.713
Water Temp. $(^{\circ}C)$	3.834	3.872	4.219	4.379
$\arcsin\sqrt{\%\text{ sand/fines}}$	125.307	123.163	124.945	136.122
$\log_{10}[\text{PWP.All} + 1]$	21.154	22.265	20.330	32.969
$\log_{10}[\text{Total Rds-wtrshd} + 1]$	35.228	43.806	43.846	44.711
$\log_{10}[\text{Unimp. Rds-wtrshd} + 1]$	-22.549	-26.600	-11.208	-2.326
$\arcsin\sqrt{\%\text{ shrub/scrub}}$	173.804	178.497	170.157	172.220
IDX.Ground	0.096	0.111	0.107	0.080

Columbia Plateau Ecoregion, Washington

In general, we found different expectations for reference conditions in the Columbia Plateau than for other regions of Washington. Results for the Columbia Plateau illustrate a clear separation in PCA space for its 4 reference sites from the 75 random sites (Figure 11 and Table 16). There were no random sites that could be proposed for least-disturbed condition, applying the same criteria as previously for the western and eastern regions of the state. Sites appear to have been allocated to clusters along a gradient corresponding perfectly with PC1, and cluster 2 occupies the neutral space in the plot.

The component eigenvectors associated with sites in cluster 1 (Figure 11) were low sand/fines and embeddedness, and high D50Log10 (large substrate sizes). These sites were also associated with the eigenvectors for increased LWD, shrub/scrub cover, lower water temperatures, and lower levels of human disturbance.

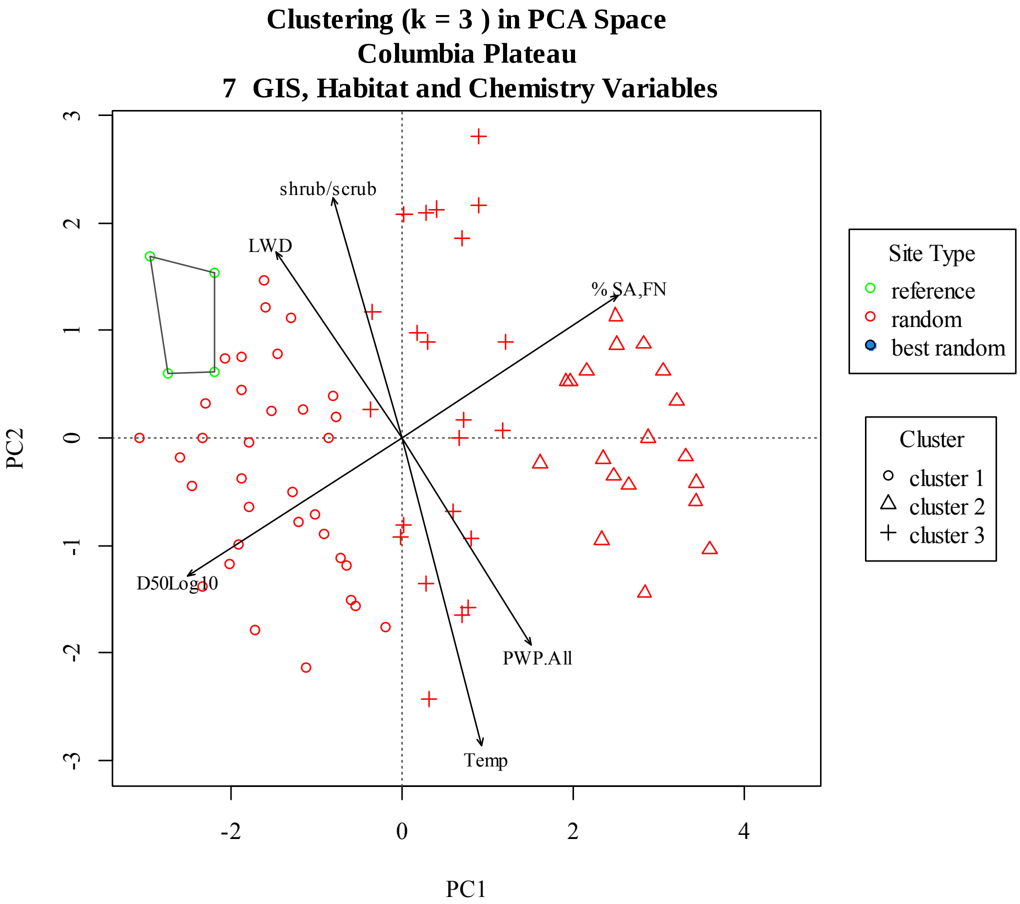


Figure 11. For the Columbia Plateau, the distribution of individual sites in principal component space (PC1 and PC2).

Color indicates site type and symbols indicate cluster. The convex hull marks the boundary in PC space of all reference sites assigned to cluster 1 (k-means clustering, k=3). Arrows represent variable loadings (eigenvectors) for the first 2 PCs. One vector is not pictured here (Embeddedness nearly identical to % sand/fines). PCA and cluster analysis were of 7 GIS, habitat, and chemistry metrics jointly, and the first 2 PCs explain 0.674 of their variability.

Table 16. Group membership of sites to clusters from a *k-means* cluster analysis (*k*=3) of 75 random and 4 reference sites in the Columbia Plateau.

Sites missing data were excluded from analysis. Cluster analysis was for 7 GIS, habitat, and chemistry metrics jointly.

Ecoregion	**Random Sites**				**Reference Sites**		
	1	**2**	**3**		**1**	**2**	**3**
Columbia Plateau	34	22	19		4	0	0

A linear discriminant analysis of the standardized site data from the Columbia Plateau clearly separated groups defined by PCA, clustering and inclusion within the convex hull (Figure 12). There was overwhelming evidence against the null hypothesis of equal means, suggesting the spatial separation of these four groups in multivariate space ($p < 0.0001$, from a multivariate ANOVA, Wilk's lambda = 0.0500, approximate F-statistic of 17.39 with 21 and 199 *df*). Sites within the convex hull of reference sites for cluster 1 ($n = 4$; Table 17) were characterized primarily by low water temperatures, low sand/fines, and low levels of human disturbance, but also by higher levels of LWD and shrub/scrub cover. Cluster 1 sites not included within the convex hull ($n = 34$) generally had less LWD, less cover from shrub/scrub, higher water temperatures, and greater levels of human disturbance (PWP.All) than those within the convex hull. Sites in clusters 2 and 3, most notably, had more sand fines, increased human disturbance, and higher water temperatures (Table 17) than sites grouped with cluster 1.

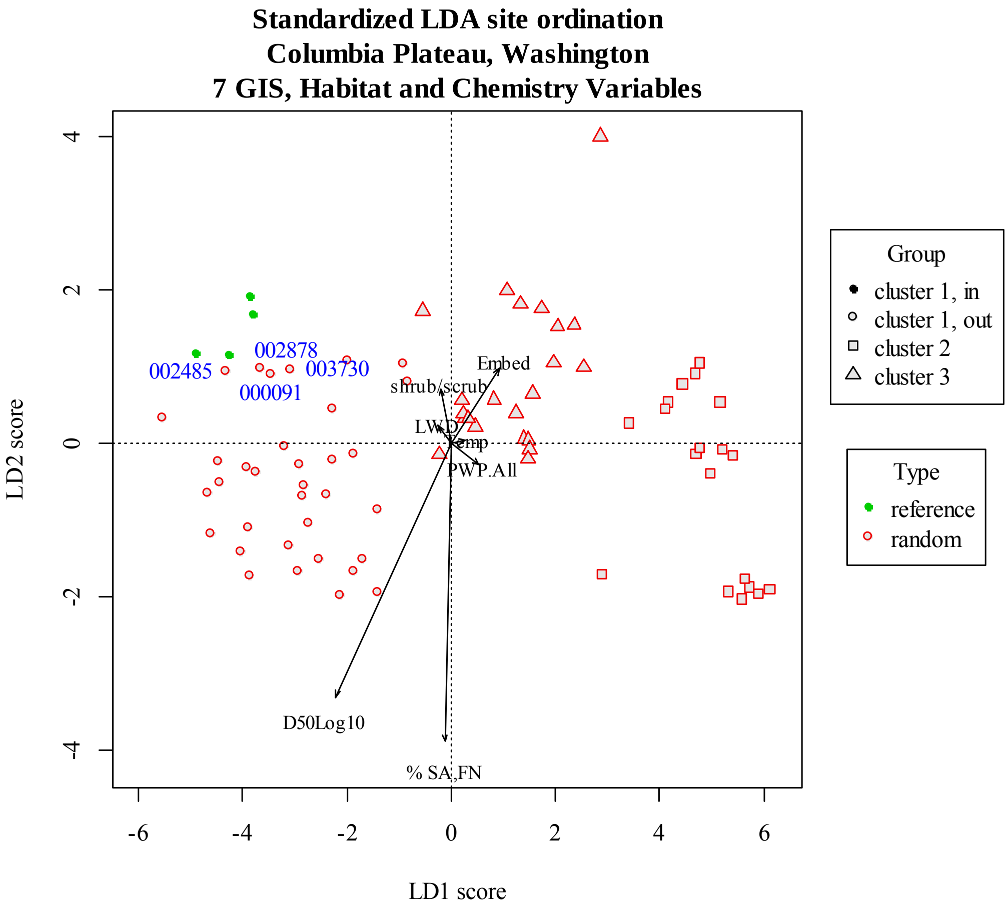


Figure 12. Linear Discriminant Analysis of standardized site data for the Columbia Plateau sites by group.

This is derived from clustering and principal component ordination using a convex hull around the bulk of the reference sites to determine which sites were within (in) the hull, and which were not (out).

Four selected points were labeled by site name as potentially "least disturbed" random sites.

Table 17. For the Columbia Plateau Ecoregion, group means for 7 GIS, habitat, and chemistry metrics.

For comparison, group means for BPJ reference sites and for all random sites (n=76) are provided first, including sites missing data for some metrics. Means are on the transformed scale, where applicable.

Metric extent	**Metric**	**Columbia Plateau**					
		BPJ ref.	**random**	**cluster 1, in**	**cluster 1, out**	**cluster 2**	**cluster 3**
		n=4	*n*=76	*n*=4	*n*=34	*n*=22	*n*=19
Reach	D50Log10	1.44	0.26	1.44	1.37	-0.18	-1.17
Reach	$\log_{10}[\text{LWD vol.} + 1]$	0.65	0.25	0.65	0.34	0.28	0.04
Point	Water Temp. $(^{\circ}C)$	13.84	16.67	13.84	15.95	16.76	17.87
Catchment	$\arcsin\sqrt{\%\ \text{shrub/scru}}$	0.82	0.43	0.82	0.46	0.51	0.28
Catchment	$\arcsin\sqrt{\%\ \text{sand/fines}}$	0.39	0.86	0.39	0.50	0.94	1.40
Reach	$\arcsin\sqrt{\%\ \text{Embed}}$	0.62	0.96	0.62	0.63	1.05	1.45
Reach	$\log_{10}[\text{PWP.All} + 1]$	0.10	0.27	0.10	0.21	0.26	0.40

The identification function correctly predicted group membership for 8 of 10 sites from the holdout, repeat visit data (Table 18). All misclassifications occurred between adjacent categories. One site modeled within the convex hull of cluster 1 was allocated to *cluster 1, out* by the identification function, and one site modeled in cluster 2 was allocated to *cluster 1, out*.

Table 18. Fisher's Linear Discriminant Analysis identification function coefficients for classification of a new site in the Columbia Plateau Ecoregion, Washington.

Fisher's Linear Discriminant Coefficients	**Group**			
	cluster 1, in	**cluster 1, out**	**cluster 2**	**cluster 3**
Intercept	-59.430	-61.279	-96.900	-63.724
$\log_{10}[\text{LWD vol.} + 1]$	14.797	11.222	4.556	8.377
D50Log10	31.697	31.657	16.318	19.259
Water Temp. $(^{\circ}C)$	1.126	1.498	2.247	1.926
$\arcsin\sqrt{\%\ \text{sand/fines}}$	29.608	50.565	50.737	37.575
$\arcsin\sqrt{\%\ \text{Embed}}$	59.443	44.636	63.233	56.669
$\log_{10}[\text{PWP.All} + 1]$	2.443	7.272	34.369	19.049
$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	-2.651	-7.395	-13.256	-7.309

Discussion

Strengths of the multivariate ordination approach

Group means for environmental variables (outlined in Tables 9, 14, and 17 for each region) successfully and consistently tracked ecological expectations. In other words, the *best* set of sites (*cluster 1, in*) within each region consistently had fewer sand fines, less bare ground cover, lower road densities, lower water temperatures, greater dissolved oxygen, and greater canopy cover. In contrast, those sites not grouped with the bulk of the reference sites (i.e., clusters 2 and 3) consistently had group means suggesting disturbance (greater sand fines, more bare ground cover, higher road densities, higher water temperatures, less dissolved oxygen, and less canopy cover). These groups were defined by clustering, PCA, and convex hull inclusion based on their similarities with the bulk of the BPJ reference sites with respect to the environmental variables. In essence, statistically derived groups reveal (dis-) similarities with the bulk of the reference sites without needing to determine specific thresholds of disturbance.

Our general approach differs from that of previous investigators (Herlihy et al. 2008; Waite et al. 2000; Whittier et al. 2006; Whittier et al. 2007), but we understand the value of thresholds. They provide clear guidelines for site selection and are straightforward. Our ordination method, however, relies solely on trends observed in BPJ sites to set the expectations for reference condition. Here, *a priori* specifications have been made in the selection-process for BPJ sites, but are not required for any threshold specifications. If traditional thresholds are desired, they can easily be interpolated from quantiles of predictors for the optimal group (*cluster 1, in*, Appendix E).

Another strength of this multivariate ordination approach is its flexibility. We illustrate its plasticity with an examination of a random site (WAM06600-000672) from the North Cascades Ecoregion of Eastern Washington. When building the linear discriminant predictive model, the site was ordinated within the convex hull of cluster 1 reference sites based on a Sept. 23, 2009 survey. This was a randomly selected survey from among all repeat surveys for this site. Modeled and predictive survey data from the same season (Aug. 4, 2009) are displayed in Table 19, and the scores for each group are given in Table 20.

Many sites have decisive scores from the identification functions, but this random site in the North Cascades Ecoregion is different. The difference between the highest and lowest score is only 2.38 points. For Eastern Washington hold-out data, the median difference between function scores was 7.9 (n = 23), with a maximum difference of 18.00 and a minimum difference of 2.00.

Table 19. Survey data for random site WAM06600-000672 in the North Cascades Ecoregion, Eastern Washington.

Repeat visit data were randomly selected from all visits for modeling (9/23/2009) and predictive purposes (8/04/2009).

Variable	Modeling Survey Data 9/23/2009	Predictive Survey Data 8/04/2009
$\log_{10}[\text{LWD vol.} + 1]$	1.64	1.69
$\arcsin\sqrt{\%\ \text{canopy}}$	0.997	0.997
$\arcsin\sqrt{\%\ \text{evergreen}}$	1.05	1.05
DO (mg/L)	10.2	9.1
D50Log10	1.91	1.84
Water Temp. $(°C)$	7.70	14.3
$\arcsin\sqrt{\%\ \text{sand/fines}}$	0.275	0.249
$\log_{10}[\text{PWP.All} + 1]$	0.241	0.320
$\log_{10}[\text{Total Rds-wtrshd} + 1]$	0.437	0.437
$\log_{10}[\text{Unimp. Rds-wtrshd} + 1]$	0.437	0.437
$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	0.500	0.500
IDX.Ground	49.8	74.9

Table 20. Scores for each group from Fisher's Linear Discriminant Identification Functions for random site WAM06600-000672 in the North Cascades Ecoregion of Eastern Washington.

Scores are based on repeat survey data from Aug. 4, 2009.

Site Name	Score cluster 1, in	Score cluster 1, out	Score cluster 2	Score cluster 3
WAM06600-000672	333.32	334.68	333.74	335.70

The identification function allocated this site to cluster 3, though the modeling analysis placed it within the convex hull of cluster 1 sites. Differences in reach- and point-level survey data have tipped the balance. Notably, there was a large spike in water temperature and bare ground cover for the predictive data. Also, both surveys had very high levels of human disturbance (PWP.All) with higher levels in the predictive data (Table 16). Though catchment metrics did not vary between surveys, this site also had high road densities compared to means shown in Table 14. The identification function has performed adequately at allocating a site of questionable quality to cluster 3, since an examination of the predictive survey data reveals strong differences from other reference sites in Eastern Washington. This example illustrates the capacity and flexibility of the multivariate method at allocating sites to groups based on the whole picture as opposed to a series of stipulations (threshold method).

Digging deeper into this specific example, we found insightful notes from field forms for the site WAM06600-000672. We found that additional variability in some of the metrics may have been unintentionally brought on by sampling slightly different reach lengths. Despite this, precision analysis of these various metrics indicated that these metrics generally performed reasonably well. Repeatability of these variables was moderate to high (S:N ratio > 3, Appendix B).

While DO, LWD, and temperature vary with natural conditions, DO and LWD are expected to be higher and temperature lower under least-disturbed condition relative to impacted sites. The approach taken in this analysis describes the expected average conditions of these variables under least-impacted condition. The deliberate, pre-screening process of sites minimized important anthropogenic stressors among the hand-picked BPJ locations. As a result, we believe that environmental conditions at BPJ sites do represent the range of those expected for sites in least-disturbed condition.

Our approach can show whether DO, LWD, and water temperature lie outside the distribution of BPJ sites. We feel this captures the range of natural variability in these variables. Otherwise, we would rely on a threshold value to determine potential impairment. To some degree, data on human disturbance and natural gradients may have been confounded. Reducing the scope of the classification scheme could minimize these impacts. However, small sample sizes constrain the regionalization scheme.

Exceptional cases with large changes in point-predictors, such as water temperature or bare ground, do not undermine the central tendencies observed in the data. In this example from the North Cascades, seasonal differences in reach- and point-level variables had a strong influence on group membership. We are confident in the ability of the multivariate ordination (LDA) to correctly classify sites because of the strong evidence of differences in means for point-level metrics between reference and random sites within each region. See ANOVA results, in Table 7.

For other studies, investigators have chosen to eliminate environmental variables subject to temporal variability (Reynoldson et al., 1997). Here, Ecology has limited the potential for confounding temporal effects by restricting sampling to the summer-fall index period. Hold-out data from within the same season validate the LDA model's ability to consistently allocate sites, regardless of survey date and seasonal variation. Forty-three out of 56 sites (77%) were classified by their respective regional identification function as belonging to the same group with the modeling data. Additionally, only 1 of 13 misclassified sites was allocated to a non-adjacent group. Out of an abundance of caution, we share a vignette of this situation above and illustrate in Tables 19 and 20.

Synthesis and concluding thoughts

The goal of this report was threefold: (1) to describe reference conditions using multivariate statistical methods and define statistical criteria for reference conditions, (2) to determine whether any random sites meet these criteria, and (3) to examine the similarities and dissimilarities of BPJ reference sites in environmental multivariate space. In this analysis, we used cluster membership and ordination scores along the first two principal component axes to define the set of random sites most similar to the existing BPJ reference sites. Data from these *best* random sites may be used in conjunction with the BPJ reference sites to build subsequent models for MMI and O/E analyses. The BPJ sites are also listed by group to highlight those that may be dissimilar to the bulk of the BPJ reference sites (Appendix G).

In total, 20 random sites from eight of the nine different ecoregions across Washington are proposed here for least-disturbed condition (Table 21). If these sites were considered in conjunction with BPJ reference sites, the increase in sample sizes by ecoregion would be very beneficial. This is especially the case for the Columbia Plateau. Here, the total number of least-disturbed sites would increase from 4 to 8, since many statistical tests require a minimum number of 5 observations for the validity of asymptotic results. The list of sites deemed least disturbed may require revisions, since some reference sites appear dissimilar from the bulk of the reference sites based on ordinations. Both lists of least-disturbed random sites and potentially disturbed reference sites are listed by site name in Appendices F and G.

Table 21. Counts of sites proposed for least-disturbed condition for watershed health assessment.

Ecoregion	BPJ reference sites	Least disturbed random sites	Total
Blue Mountains	9	0	9
Cascades	10	3	13
Coast Range	9	2	11
Columbia Plateau	4	4	8
Eastern Cascades	10	1	11
North Cascades	11	6	17
Northern Rockies	12	5	17
Puget Lowland	10	1	11
Willamette Valley	0	0	0
Total	75	22	97

We cannot overemphasize the importance of field validation of all proposed revisions to the list of sites in least-disturbed condition, whether additions or removals. No statistical model can incorporate that invaluable gut feeling, the innate knowledge of a resource manager seasoned in the field.

Another factor to be considered with this analysis is that all metrics were treated equally, though in reality some may be of greater importance. For example, a site may be pristine with respect to all metrics except human disturbance and thus have been included in the list of least-disturbed random sites. We have shown here that some *best* random sites (within the group cluster 1, in) exhibit higher levels of human disturbance (PWP.All), but were comparable to reference sites with respect to other ecological metrics. If efforts of field validation reveal that these least-disturbed random sites are not of sufficient quality, the human disturbance metric could be upweighted in the analysis or a specific threshold could be defined for this single metric. As an additional comparison, we provide quantiles in the form of plots for the empirical cumulative distribution function (ECDF plots, Appendix D) and tabled values (Appendix E) for the groups defined by multivariate ordination, clustering, and inclusion in the convex hull of cluster 1 reference sites.

There is some discussion in the literature on the adequacy of ecoregions to partition the natural variability, since relationships between stream biological conditions and stressors are complex (Maloney et al., 2009; Waite et al., 2000; Hawkins et al., 2000; Ode et al., 2008). Some environmental characteristics, e.g., substrate size will respond to both anthropogenic stressors and vary naturally. Hawkins and Vinson (2000) used macroinvertebrate data to compare the classification strengths of *a priori* groupings, such as by ecoregion, and *a posteriori* biotic grouping, such as using the macroinvertebrate data itself to group sites. Overall, they found low classification strengths for both types of groupings. Hawkins and Vinson (2000) noted that ecoregions and other *a priori* classifications schemes were “sufficiently heterogeneous in their environmental conditions to have significantly different assemblage composition.” They prefer the *a posteriori* classifications using the macroinvertebrate data, though the classification strengths for these were not much better. However, we find this logic somewhat circular. This point is emphasized in Bailey et al.’s (2004) foundational book on the reference condition approach. The authors urge investigators not to use the structure of the community itself to identify a site as reference or otherwise, since the objective is to characterize the range of variation among communities in reference condition and describe the characteristics of reference sites.

Another important factor to consider is that the current investigation, establishing reference conditions, is only the first step in the bioassessment process. Particularly for predictive models comparing observed/expected (O/E) indices, test locations will be compared with reference locations based on site-specific variables, such as % slope, elevation, and % fast-water habitat. Thus, natural environmental gradients that affect the distribution of macroinvertebrate taxa are explicitly specified in predictive models. Another interesting study compared trade-offs between large-scale models and regional-scale models. Authors found agreement among MMI scores for the same region when comparing large-scale models to regional-scale assessments (Ode et al. 2008).

We hope that the results and statistical framework presented here will accurately and precisely estimate reference condition at specific sites across Washington and improve the technical quality of watershed health monitoring.

References

Literature Cited

Bailey, R., R. Norris, and T. Reynoldson. 2004. Bioassessment of freshwater ecosystems: using the reference condition approach. Kluwer Academic Publishers, Boston, Massachusetts.

Bilotta, G. and R. Brazier. 2008. Understanding the influence of suspended solids on water quality and aquatic biota. Water Research, 42:2849-2861.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer, New York.

Davis, W. and J. Scott. 2000. Mid-Atlantic Highlands Streams Assessment: Technical Support Document. EPA/903/B-00/004. USEPA Mid-Atlantic Integrated Assessment, Ft. Meade, Maryland.

Dixon, P. 1994. Testing spatial segregation using a nearest-neighbor contingency table. Ecology, 75:1940-1948.

Dixon, P. 2002. Nearest-neighbor contingency table analysis of spatial segregation for several species. Ecoscience, 9:142-151.

Everitt, B., and T. Hothorn. 2011. An Introduction to Applied Multivariate Analysis with R. Springer, New York.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188.

Hawkins, C. and M. Vinson. 2000. Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessment. Journal of the North American Benthological Society, 19:501-517.

Hawkins, C., R. Norris, J. Gerritsen, R. Hughes, S. Jackson, R. Johnson and R. Stevenson. 2000. Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. Journal of the North American Benthological Society, 19:541-556.

Herlihy, A., S. Paulsen, J. Van Sickle, J. Stoddard, C. Hawkins, and L. Yuan. 2008. Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale. Journal of the North American Benthological Society, 27:860-877.

Janisch, J. 2013. Dictionary of Metrics for Physical Habitat: Definitions and Calculations Used for Watershed Health Monitoring and Related Studies. Washington State Department of Ecology, Olympia, Washington. https://fortress.wa.gov/ecy/publications/SummaryPages/1303033.html.

Karr, J. R. 1991. Biological integrity: a long-neglected aspect of water resource management. Ecological Applications, 1:66-84.

Kaufmann, P., P. Levine, E. Robison, C. Seeliger, and D. Peck. 1999. Quantifying Physical Habitat in Wadeable Streams. EPA/620/R-99/003. U.S. Environmental Protection Agency, Washington, D.C.

Ketchen, D., and C. Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal, 176:441-458.

Legendre, P., and L. Legrendre. 1998. Numerical Ecology. Developments in Environmental Modelling 20. 2nd English edition. Elsevier Science & Technology, New York.

Maloney, K., D. Weller, M. Russell, and T. Hothorn. 2009. Classifying the biological condition of small streams: an example using benthic macroinvertebrates. Journal of the North American Benthological Society, 28:869-884.

Merritt, G. 2009. Status and trends monitoring for watershed health and salmon recovery: Field data collection protocol, wadeable streams. Washington State Department of Ecology, Olympia, Washington.

Milligan, G., and M. Cooper. 1988. A study of standardization of variables in cluster analysis. Journal of Classification, 5:181-204.

Ode, P., C. Hawkins, and R. Mazor. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. Journal of the North American Benthological Society, 27:967-985.

Omernik, J. 1987. Ecoregions of the conterminous United States. Annals of the Association of American Geographers, 77:118-125.

Omernik, J., and A. Gallant. 1986. Ecoregions of the Pacific Northwest States. U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, OR.

Pater, D., S. Bryce, T. Thorson, J. Kagan, C. Chappell, J. Omernik, S. Azevedo, and A. Woods. 1998. Ecoregions of Western Washington and Oregon. U.S. Geological Survey, Reston, VA.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reynoldson, T., R. Norris, V. Resh, K. Day, and D. Rosenberg. 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, 16:833-852.

Richards, C., G. Host, and J. Arthur. 1993. Identification of predominant environmental factors structuring stream macroinvertebrate communities within a large agricultural catchment. Freshwater Biology, 29:285-294.

Robison, E. G. 1997. Reach scale sampling metrics and longitudinal pattern adjustments of small streams. Ph.D. Dissertation. Department of Forest Engineering and Hydrology. Oregon State University, Corvallis, OR. 254 pp. http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/9756/Robison,%20E%20George_1997_PhD.pdf?sequence=3

Roper, B., J. Buffington, S. Bennett, S. Lanigan, E. Archer, S. Downie, J. Faustini, T. Hillman, S. Hubler, K. Jones, C. Jordan, P. Kaufmann, G. Merritt, and C. Moyer. 2010. A comparison of the performance and compatibility of protocols used by seven monitoring groups to measure stream habitat in the Pacific Northwest. North American Journal of Fisheries Management, 30:565-587.

Sokal, R. and F. Rohlf. 1995. The principles and practice of statistics in biological research (3rd ed.). New York: Freeman and Company.

Stoddard, J., D. Larsen, C. Hawkins, R. Johnson, and R. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications, 16:1267-1276.

U.S. Census Bureau. 2014. Population Estimates – Metropolitan and Micropolitan Statistical Areas. Web. Accessed: 15 Sep. 2014.

U.S. EPA (Environmental Protection Agency). 2005. Level III ecoregions of the continental United States (revision of Omernik, 1987). U.S. Environmental Protection Agency, Corvallis, Oregon.

U.S. OFM (Office of Financial Management). 2011. Washington population growth slowed during last decade, but state is more diversified than in 2000. Web. Updated: 23 Feb. 2011. Accessed: 19 Nov. 2014.

Waite, I., A. Herlihy, D. Larsen, and D. Klemm. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. Journal of the North American Benthological Society, 19:429-441.

Whittier, T., J. Stoddard, R. Hughes, and G. Lomnicky. 2006. Associations among catchment- and site-scale disturbance indicators and biological assemblages at least-and most-disturbed stream and river sites in the western United States. American Fisheries Society Symposium, 48:641-664.

Whittier, T., J. Stoddard, D. Larsen, and A. Herlihy. 2007. Selecting reference sites for stream biological assessments: best professional judgment or objective criteria. Journal of the North American Benthological Society, 26:349-360.

Wood, P., and P. Armitage, 1997. Biological effects of fine sediment in the lotic environment. Environmental management, 21:203-217.

Wright, J. F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian Journal of Ecology, 20:181-197.

R Packages Used

`dixon`: De la Cruz, M. 2008. Methods for analyzing point data. In: Introduccion al Analisis Espacial de Datos en Ecologia y Ciencias Ambientales: Metodos y Aplicaciones (eds. Maestre, F. T., Escudero, A. y Bonet, A.), pp 76-127. Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos y Caja de Ahorros del Mediterraneo, Madrid. ISBN: 978-84-9849-308-5.

`gclus`: Hurley, C. 2012. gclus: Clustering Graphics. R package version 1.3.1. http://CRAN.R-project.org/package=gclus

`MASS`: Venables, W., and B. Ripley. 2002. Modern Applied Statistics with S. Fourth Edition. New York: Springer. ISBN 0-387-95457-0.

`plyr`: Wickham, H. 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

`spatstat`: Baddeley, A., and R. Turner. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software 12(6), 1-42. URL http://www.jstatsoft.org/v12/i06/.

`splancs`: Rowlingson, B., and P. Diggle. 2013. splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-34. http://CRAN.R-project.org/package=splancs.

`vegan`: Oksanen, J., F. Blanchet, R. Kindt, P. Legendre, P. Minchin, R. O'Hara, G. Simpson, P. Solymos, M. Henry, H. Stevens, and H. Wagner. 2013. vegan: Community Ecology Package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan.

Appendices

Appendix A. Glossary, Metric Definitions, Acronyms, and Abbreviations*

Glossary

Catchment: A drainage area or basin in which all land and water areas drain or flow toward a central collector such as a stream, river, or lake at a lower elevation.

Embeddedness: The fraction of a substrate particle's surface that is surrounded by (embedded in) sand or finer sediments (≤ 2 mm).

Percentile: A statistical number obtained from a distribution of data, below which the numerical percentage of the data exists. For this study it is discussed in terms of estimated regional stream length within which a value is at or below a given value.

pH: A measure of the acidity or alkalinity of water. A low pH value (0 to 7) indicates that an acidic condition is present, while a high pH (7 to 14) indicates a basic or alkaline condition. A pH of 7 is considered to be neutral. Since the pH scale is logarithmic, a water sample with a pH of 8 is ten times more basic than one with a pH of 7.

Reach-level: Measurements recorded along a length of stream.

Point-level: Measurements recorded at a specific location within a stream.

Riparian: Relating to the banks along a natural course of water.

Site: Entire survey reach for a given data collection event, including any stations along the reach where observations or sample collection occurred. Maximum site length is 2000 meters. Minimum site length is 150 meters.

Watershed: A drainage area or basin in which all land and water areas drain or flow toward a central collector such as a stream, river, or lake at a lower elevation.

Metric Definitions

GIS Metrics	
DRNAREA	Drainage area derived from GIS coverages. Unit = km^2.
CANOPY_PCT	Canopy percent coverage derived from GIS coverages.. Unit = percent.
developed.low.intensity.prcnt	Percent area at the catchment level characterized by low-intensity development; includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20 – 49 percent of total cover. These areas most commonly include single-family housing units. Data generated from GIS coverage using the 2006 National land cover data for each watershed. Unit = percent.
shrub.scrub.prcnt	Percent area at the catchment level largely comprised of shrub/scrub; includes areas dominated by shrubs less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in early successional stage and trees stunted from environmental conditions. Data generated from GIS coverage using the 2006 National land cover data for each watershed. Unit = percent.
evergreen.forest.prcnt	Percent area at the catchment level larger comprised of evergreen forest; includes areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75 percent of the tree species maintain their leaves all year such that canopy is never without green foliage. Data generated from GIS coverage using the 2006 National land cover data for each watershed. Unit = percent.
TIA_MEAN_wtrshd	Total impervious area within the catchment level. Unit = km^2.
TIA_MEAN_LCA	Total impervious area within the lower-contributing area, a 2.5km buffer around the sample point clipped to the catchment boundary. Unit = km^2.
TIA_MEAN_stream_buffer	Total impervious area within the stream buffer (100m). Unit = km^2.
TIA_MEAN_stream_buffer_LCA	Total impervious area within the stream buffer in the LCA (100m stream buffer in LCA, a 2.5km buffer around the sample point clipped to the catchment boundary) . Unit = km^2.
Road Type 1	Primary highway/all-weather/hard surface. Unit = km/km^2.
Road Type 2	Secondary highway/all-weather/hard surface. Unit = km/km^2.
Road Type 3	Light-duty road/all-weather/improved surface. Unit = km/km^2.
Road Type 4	Unimproved road/fair or dry weather. Unit = km/km^2.
Road Type 9	Primary highway/all-weather/hard surface. Unit = km/km^2.

Habitat and Reach-level Metrics	
IDX.Canopy	Average, of plot-level vegetative cover ratings, to total number of vegetation plots evaluated. Plots occur at the ends of each channel-spanning transect, for all transects associated with the main channel (channel 0). Unit = percent.
IDX.Ground	Average, of plot-level vegetative cover ratings, to total number of vegetation plots evaluated. Plots occur at the ends of each channel-spanning transect, for all transects associated with the main channel (channel 0). Unit = percent.
P_Lith_SA	Areal proportion, of channel bed, along the site reach, *composed of* lithic substrates sized as sand (0.68-2 mm), as observed, bank-to-bank, at channel-spanning transects, associated with the main channel (channel 0). A result of 1.0 indicates that the channel bed, as observed along the site reach, was *composed of* a single substrate particle diameter category. Unit = unitless.
P_Lith_FN	Areal proportion, of channel bed, along the site reach, *composed of* lithic substrates sized as fines (0.001-0.68 mm), as observed, bank-to-bank, at channel-spanning transects, associated with the main channel (channel 0). A result of 1.0 indicates that the channel bed, as observed along the site reach, was *composed of* a single substrate particle diameter category. Unit = unitless.
PCT.SandFines	Ratio, adjusted to percent, of channel bed, along the site reach, dominated by the specified particle-diameter categories, as observed, bank-to-bank, at channel-spanning transects, associated with the main channel (channel 0). Unit = percent.
X.Embed	Average, of all embeddedness observations associated with the main channel (channel 0), bank to bank, for all channel-spanning transects, where transects are perpendicular to stream current. Unit = percent.
D50Log10	An exponent, determined from a logarithmic average diameter of each lithic substrate category, and from the proportion of each lithic category observed along the site reach, bank-to-bank, at channel-spanning transects, associated with the main channel (channel 0). Unit = $\log_{10}$(mm).
LWDSiteVolume100m_ m3.100m	Normalized volume of large woody debris of all size classes combined, either as intersecting or contained within the bankfull zone of the main channel (channel 0), per 100 m of channel, as visually observed along the length of site reach. Any pieces of dead trees of minimum dimensions are counted, including coarse roots and large limbs if not attached to a bole. Length minima are 2 m, Western Washington, and 1 m, Eastern Washington. Diameter minimum in all cases is ≥ 10 cm. Unit = cubic meters per 100 m.
PWP.All	Average, of all categorical proximity weights assigned to all rated plots for the proximity to the stream channel of all 13 human-influence types, as observed during a visual search of the plots. Unit = unitless.
PCT.BankAny	Percent, of all rated plots where any human-influence type, as detected during a visual search of the plots, was observed up to 30 m from the bankfull channel margin. Unit = percent.
X.DensioBank	Average, of readings of shaded (i.e., non-sky) densiometer grid-line intersections, as observed where the bankfull margins intersect each transect, for all observations associated with the main channel (channel 0). Unit = percent.

Chemistry Metrics	
Dissolved.Oxygen_mg.L	Dissolved Oxygen, measured twice, at the start and at the end of a data collection event; both sets of in situ measurements are usually made near the middle elevation of the site, on the main channel. Unit = mg/L.
Chloride_mg.L	Chloride, measured once at the start of a data collection event at the index transect of the site. Unit = mg/L.
Conductivity_uS.cm.25C	Conductivity, measured twice, at the start and at the end of a data collection event; both sets of in situ measurements are usually made near the middle elevation of the site, on the main channel. Unit = µS/cm at 25 °C.
Total.Organic.Carbon	Total Organic Carbon, measured from a site-composite sediment sample take from three separate shallow-water stations in the site. Unit = percent.
Total.Persulfate.Nitrogen_mg.L	Persulfate Nitrogen, measured once at the start of a data collection event at the index transect of the site. Unit = mg/L.
Total.Phosphorus_mg.L	Phosphorus, measured once at the start of a data collection event at the index transect of the site. Unit = mg/L.
Total.Suspended.Solids_mg.L	Total Suspended Solids, measured once at the start of a data collection event at the index transect of the site. Unit = mg/L.
Turbidity_NTU	Turbidity, measured once at the start of a data collection event at the index transect of the site. Unit = NTU.
pH_pH	pH, measured twice, at the start and at the end of a data collection event; both sets of in situ measurements are usually made near the middle elevation of the site, on the main channel. Unit = pH units.
Temperature.water_deg.C	Temperature, measured twice, at the start and at the end of a data collection event; both sets of in situ measurements are usually made near the middle elevation of the site, on the main channel. Unit = deg. C.

*****Metric definitions from Janisch, 2009; Merritt, 2009.

Acronyms and Abbreviations

ANOVA	Analysis of Variance
B-IBI	Benthic Index of Biotic Integrity
BPJ	Best Professional Judgment
CV	Coefficient of Variation
DO	Dissolved Oxygen
ECDF	Empirical Distribution Function
GIS	Geographic Information Systems
LDA	Linear Discriminant Analysis
LWD	Large Woody Debris
MANOVA	Multivariate Analysis of Variance
PC1	Principal Component Axis 1
PCA	Principal Component Analysis
PERMANOVA	Permutational Multivariate Analysis of Variance
RIVPACS	River Invertebrate Prediction and Classification System
WHM	Watershed Health Monitoring

Units of Measurement

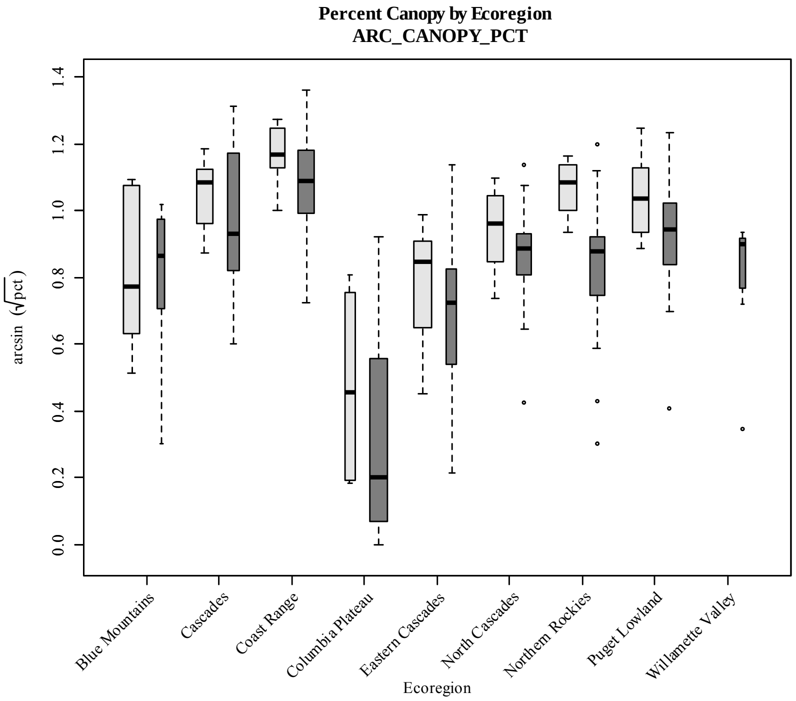
°C	degrees centigrade
cm	centimeter
km	kilometer, a unit of length equal to 1,000 meters
m	meter
mg/L	milligrams per liter (parts per million)
NTU	nephelometric turbidity units
uS/cm or µS/cm	microsiemens per centimeter, a unit of conductivity
%	percent

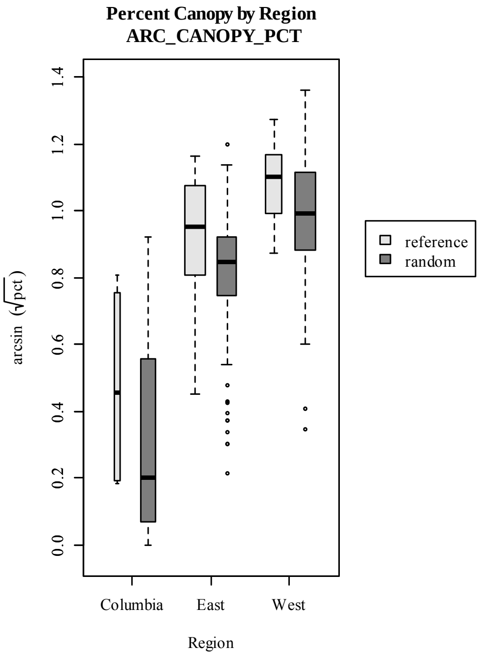
Appendix B. Coefficient of Variation for Habitat and Chemistry Metrics

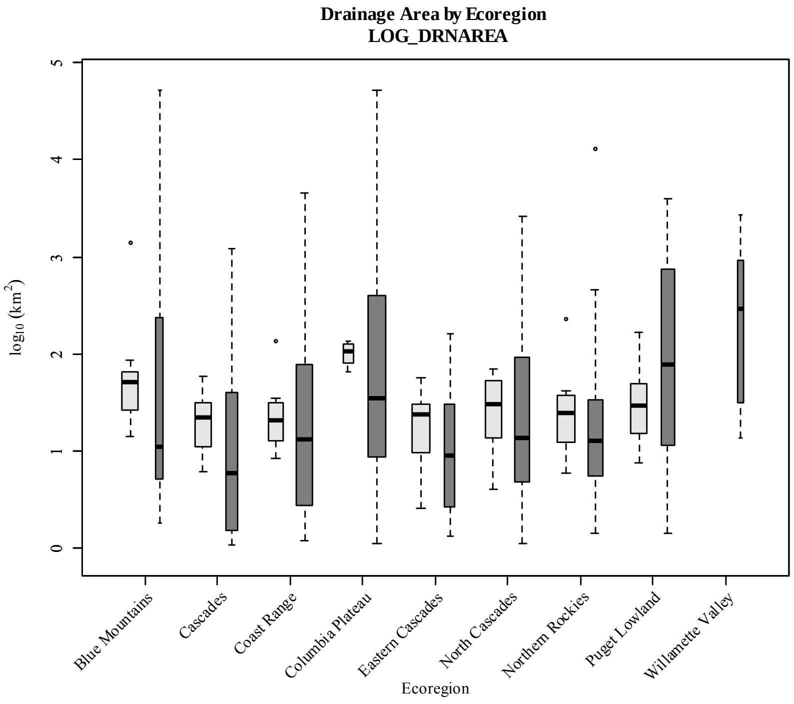
Spatial Extent	Metric	Mean CV (%)	Median CV (%)	S:N
reach-level	D50Log10	21.5	8.3	23.4
reach-level	IDX.Canopy	36.8	33.0	2.7
reach-level	IDX.Ground	18.3	15.4	3.1
reach-level	LRBS	9.7	7.6	4.1
reach-level	ARC_P_Lith_FN	43.1	14.7	15.5
reach-level	P_Lith_FN	55.4	28.3	18.0
reach-level	ARC_P_Lith_SA	26.8	19.0	2.2
reach-level	P_Lith_SA	44.6	35.8	2.4
reach-level	ARC_PPN.Canopy	8.3	4.4	6.8
reach-level	PPN.Canopy	8.6	3.3	5.9
reach-level	ARC_PCT.BankAny	62.0	26.1	2.0
reach-level	PCT.BankAny	71.2	47.1	1.6
reach-level	ARC_PCT.Fines	43.4	14.8	13.9
reach-level	PCT.Fines	55.8	28.3	16.7
reach-level	ARC_PCT.SandFines	13.6	12.2	14.8
reach-level	PCT.SandFines	24.6	22.3	17.9
reach-level	ARC_X.DensioBank	6.8	6.1	3.8
reach-level	X.DensioBank	6.0	3.6	5.8
reach-level	ARC_X.Embed	14.0	11.6	4.2
reach-level	X.Embed	23.7	20.3	3.5
reach-level	LOG_PWP.All	3281.6	28.9	3.6
reach-level	PWP.All	66.9	47.1	1.7
reach-level	LOG_LWDSiteVolume100m_m3.100m	42.1	15.3	2.1
reach-level	LWDSiteVolume100m_m3.100m	51.6	43.7	2.3
point-level	Dissolved.Oxygen_mg.L	7.5	4.6	6.3
point-level	pH_pH	4.1	3.0	0.4
point-level	Temperature.water_deg.C	10.7	9.0	3.5
point-level	LOG_Chloride_mg.L	43.6	11.5	30.7
point-level	Chloride_mg.L	16.2	12.5	8.5
point-level	LOG_Conductivity_uS.cm.25C	4.6	2.8	32.0
point-level	Conductivity_uS.cm.25C	19.1	12.7	12.3
point-level	LOG_Total.Organic.Carbon	122.2	46.9	2.8
point-level	Total.Organic.Carbon	40.6	34.6	1.4
point-level	LOG_Total.Persulfate.Nitrogen_mg.L	20.5	5.9	9.9
point-level	Total.Persulfate.Nitrogen_mg.L	27.3	19.3	5.6
point-level	LOG_Total.Phosphorus_mg.L	5.0	1.6	1.0

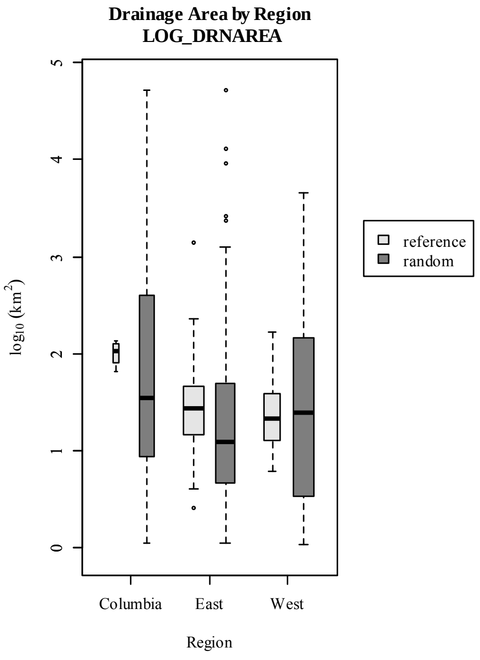
Spatial Extent	Metric	Mean CV (%)	Median CV (%)	S:N
point-level	Total.Phosphorus_mg.L	32.0	22.5	0.8
point-level	LOG_Total.Suspended.Solids_mg.L	54.7	30.0	0.4
point-level	Total.Suspended.Solids_mg.L	46.4	43.8	0.0
point-level	LOG_Turbidity_NTU	472.3	63.0	2.5
point-level	Turbidity_NTU	65.2	66.3	0.6

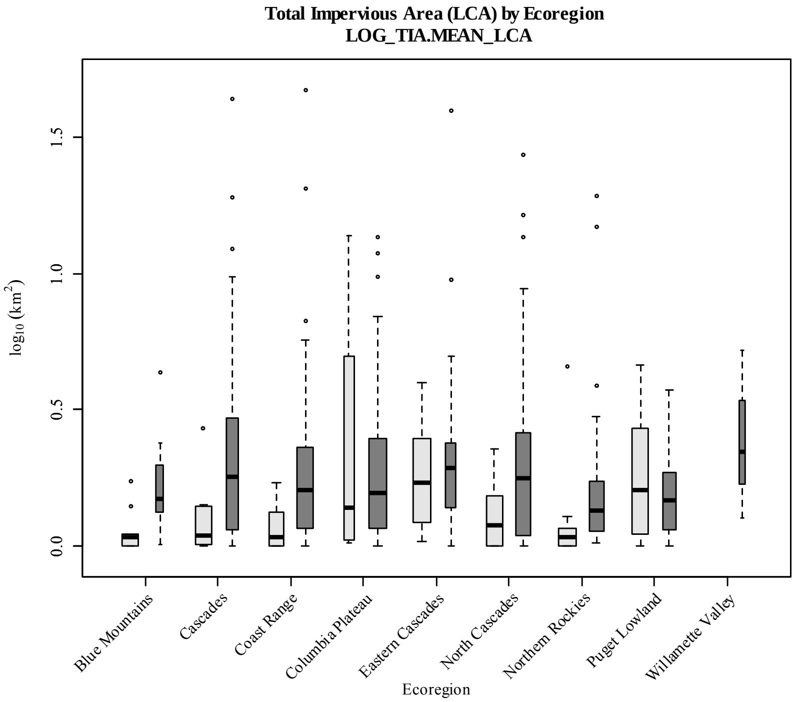
Appendix C. Plots of Selected GIS, Habitat, and Chemistry Metrics

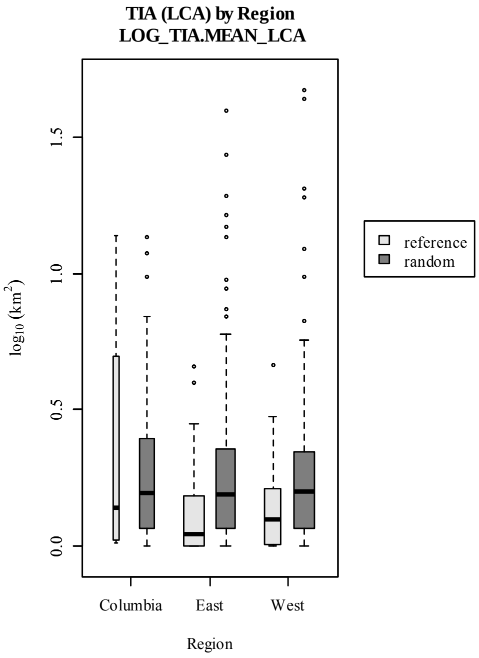


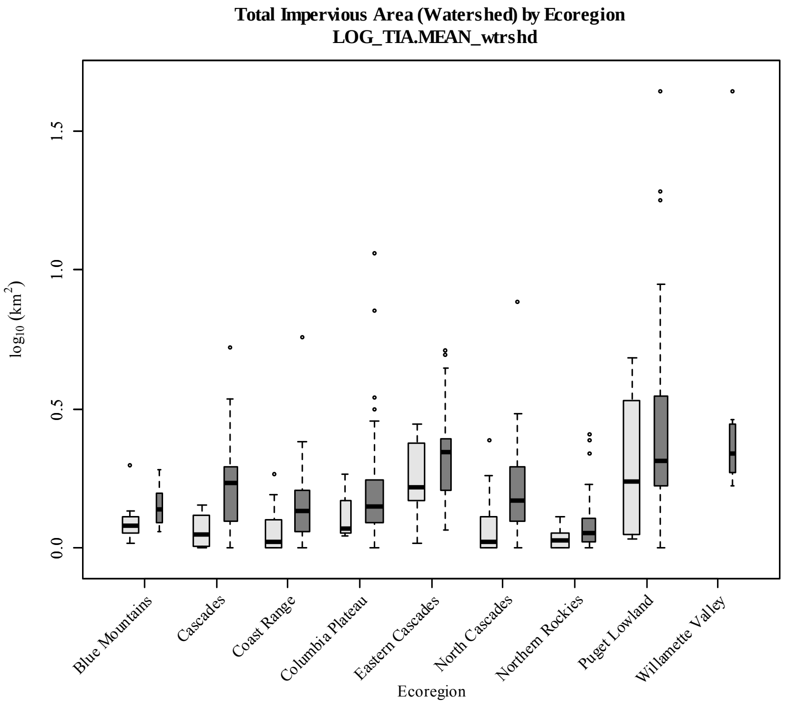


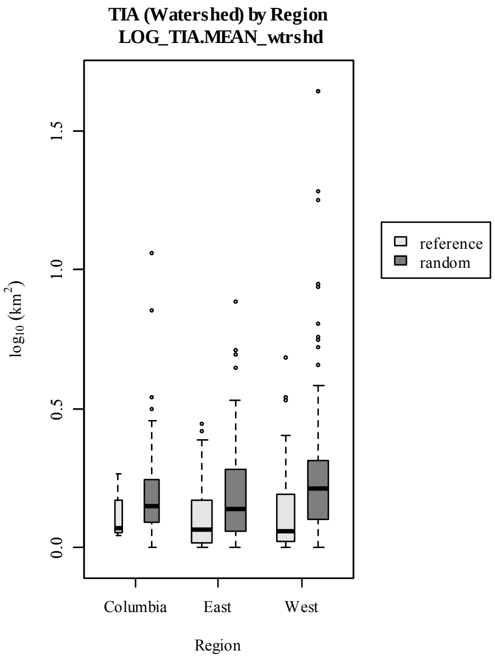


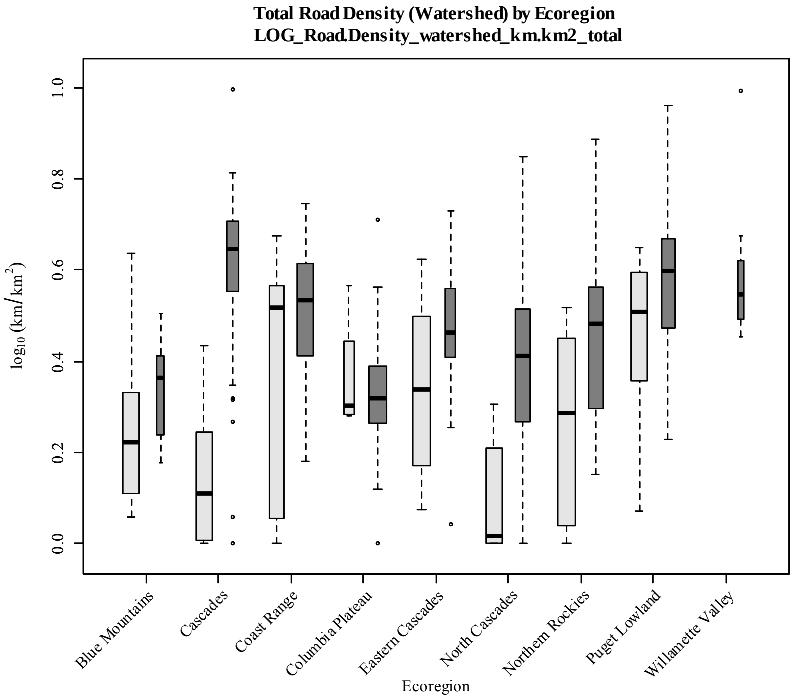


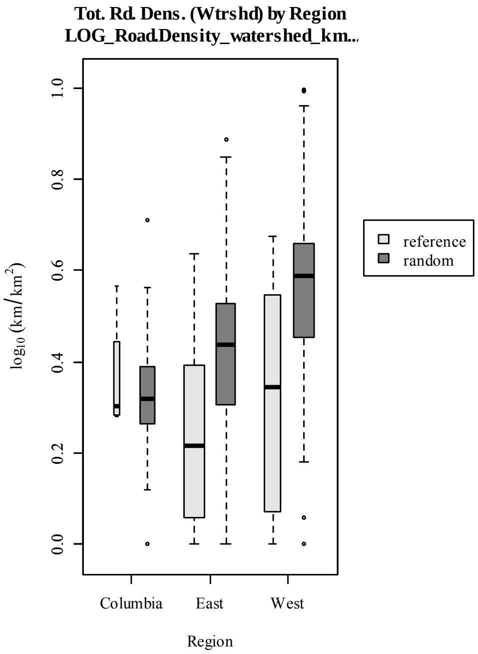

Total Impervious Area (LCA) by Ecoregion
LOG_TIA.MEAN_LCA

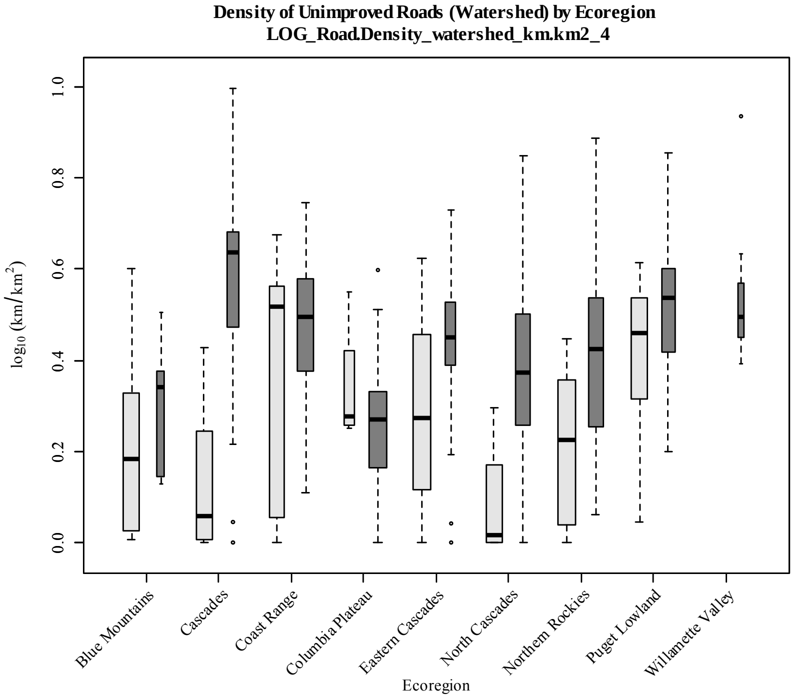

TIA (LCA) by Region
LOG_TIA.MEAN_LCA

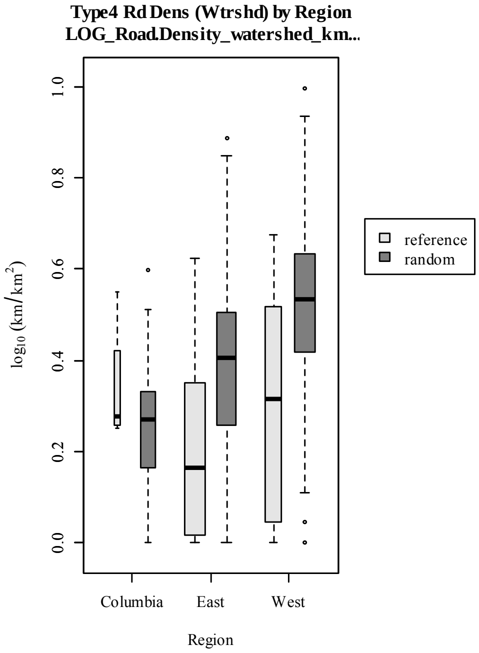

Total Impervious Area (Watershed) by Ecoregion
LOG_TIA.MEAN_wtrshd

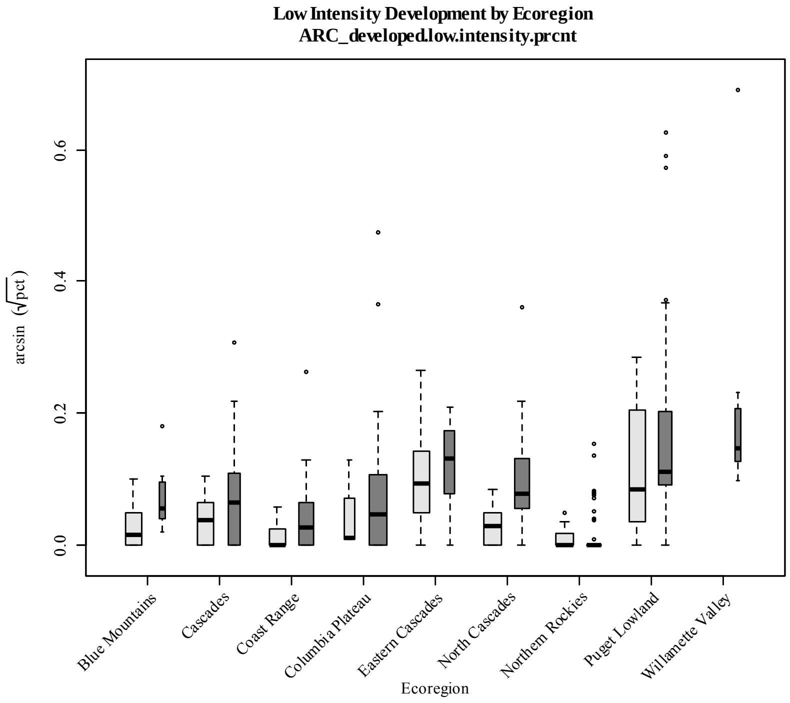

TIA (Watershed) by Region
LOG_TIA.MEAN_wtrshd

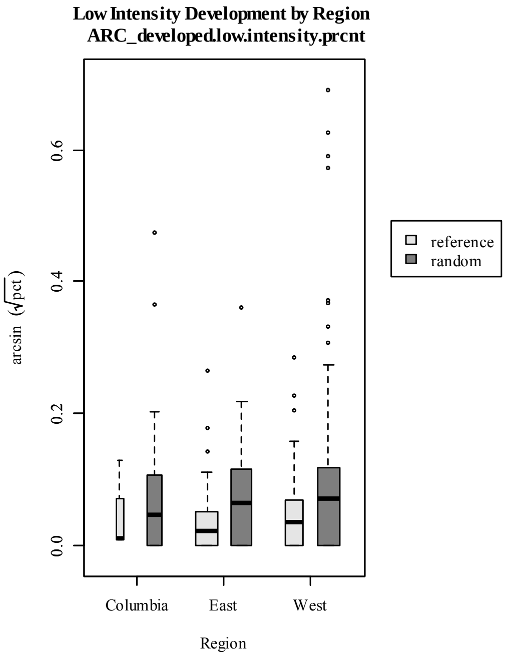

Total Road Density (Watershed) by Ecoregion
LOG_Road.Density_watershed_km.km2_total
$\log_{10}$ (km/km^2)
0.0
0.2
0.4
0.6
0.8
1.0
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

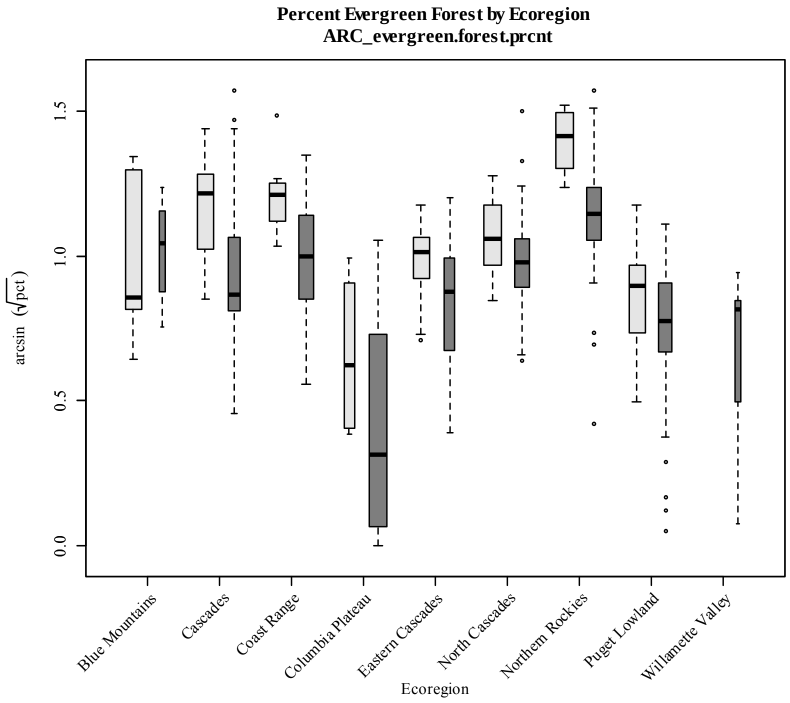

Tot. Rd. Dens. (Wtrshd) by Region
LOG_Road.Density_watershed_km...
$\log_{10}$ (km/km^2)
0.0
0.2
0.4
0.6
0.8
1.0
Columbia
East
West
Region
reference
random

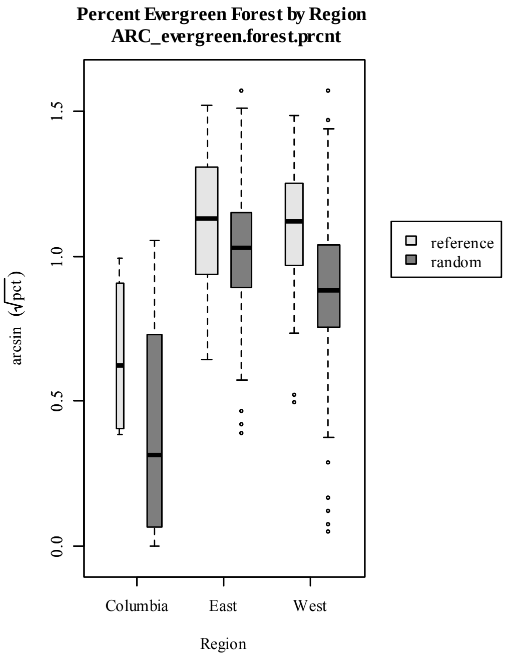

Density of Unimproved Roads (Watershed) by Ecoregion
LOG_Road.Density_watershed_km.km2_4
$\log_{10}$ (km/km^2)
0.0
0.2
0.4
0.6
0.8
1.0
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

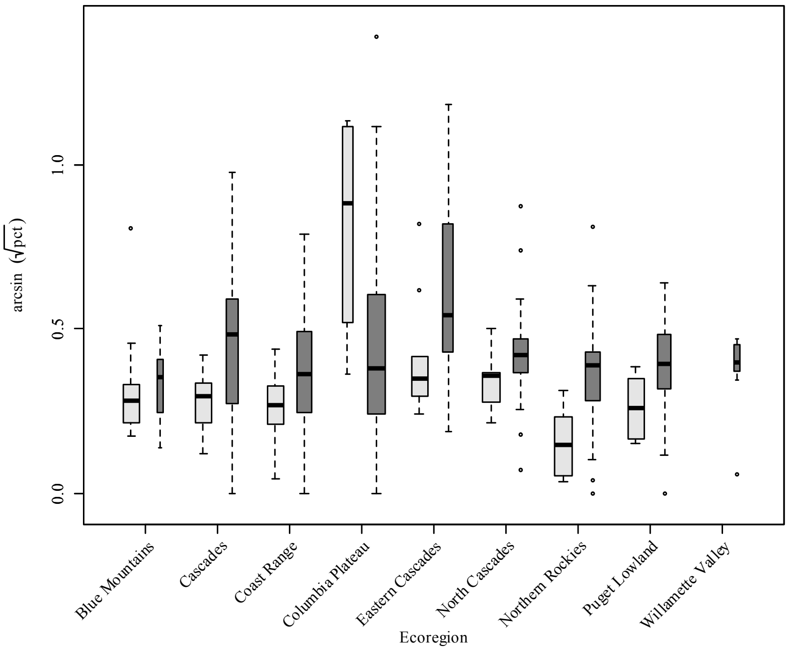

Type4 Rd Dens (Wtrshd) by Region
LOG_Road.Density_watershed_km...
$\log_{10}$ (km/km^2)
0.0
0.2
0.4
0.6
0.8
1.0
Columbia
East
West
Region
reference
random


Low Intensity Development by Ecoregion
ARC_developed.low.intensity.prcnt

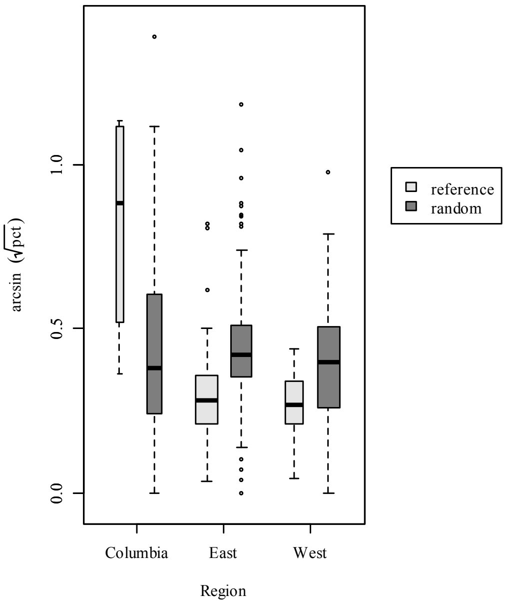

Low Intensity Development by Region
ARC_developed.low.intensity.prcnt


Percent Evergreen Forest by Ecoregion
ARC_evergreen.forest.prcnt

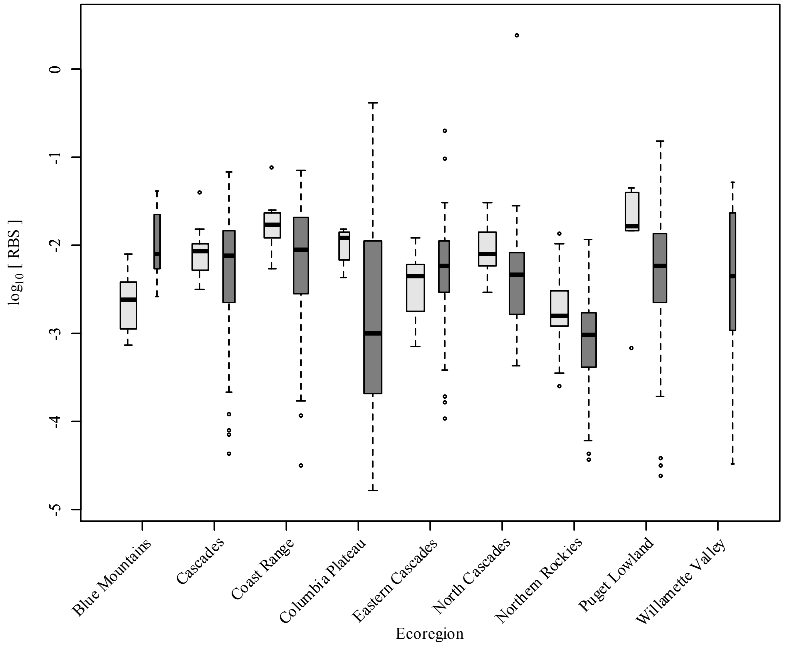

Percent Evergreen Forest by Region
ARC_evergreen.forest.prcnt



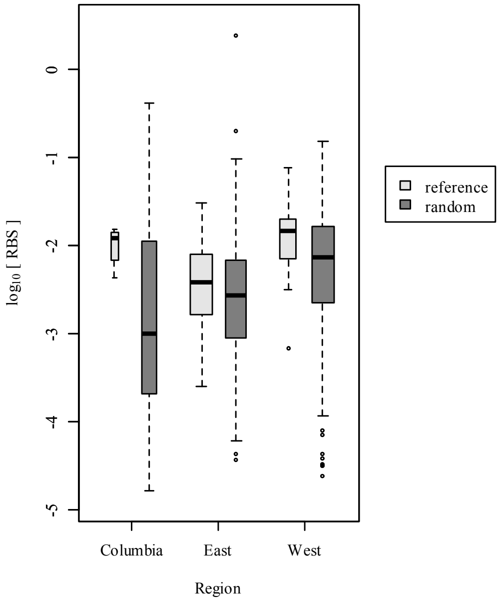
Percent Shrub/Scrub by Ecoregion
ARC_shrub.scrub.prcnt



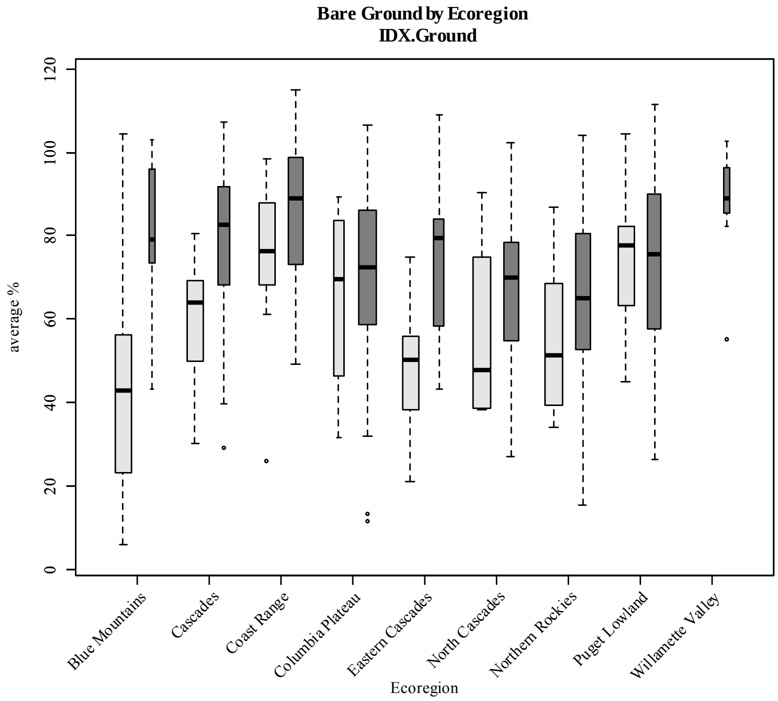
Percent Shrub/Scrub by Region
ARC_shrub.scrub.prcnt

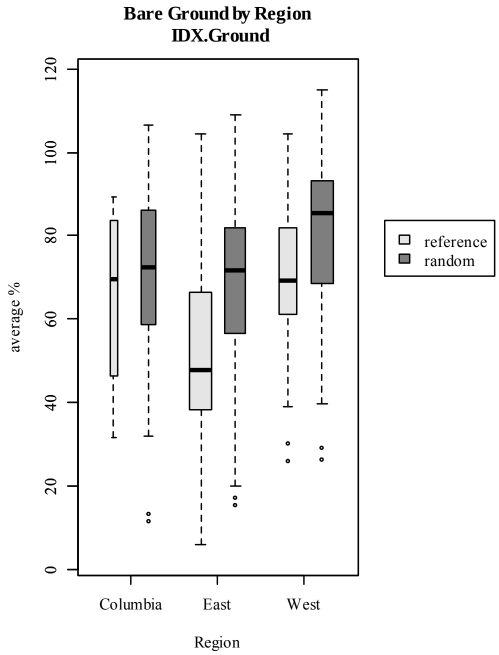


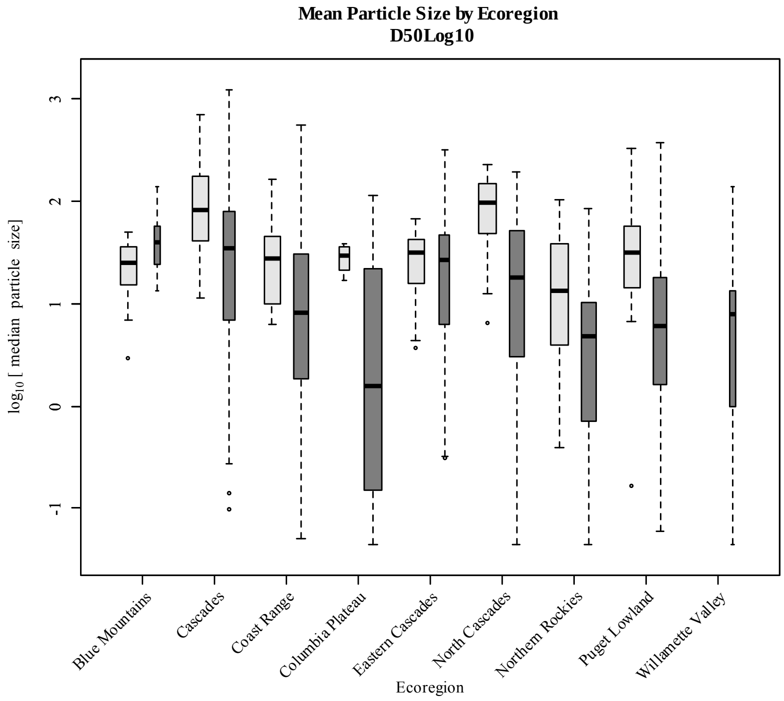
Relative Bed Stability by Ecoregion
LRBS

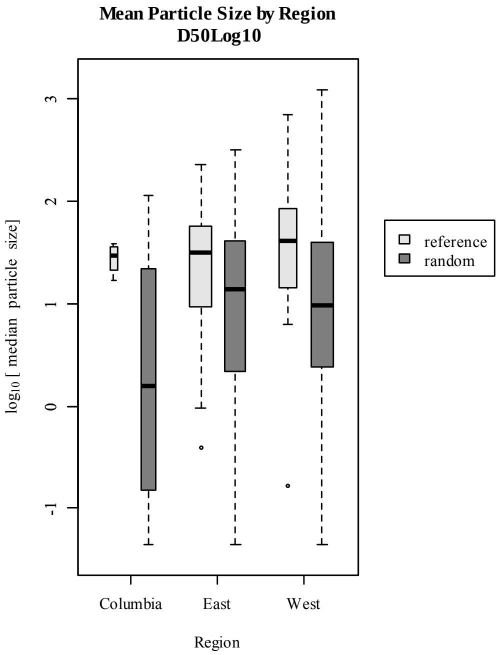


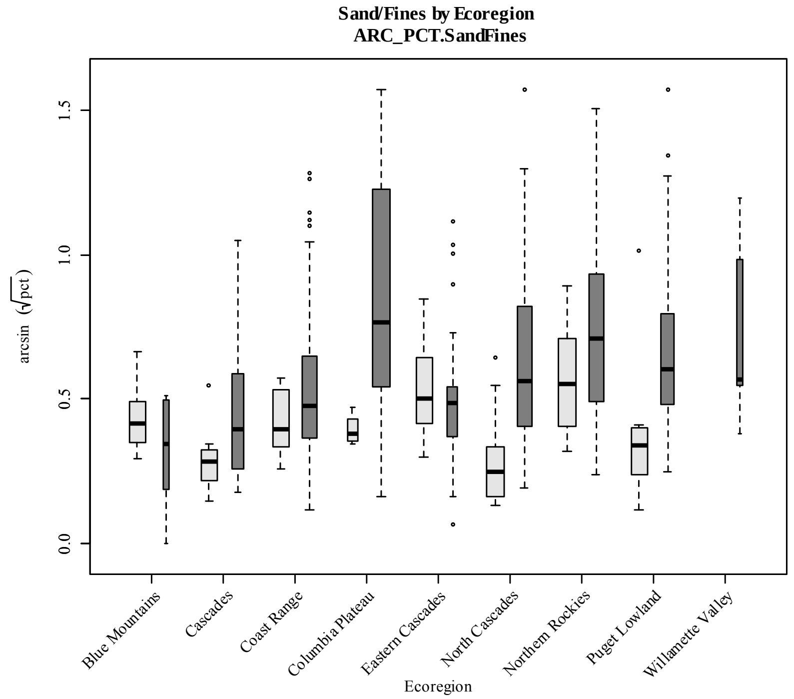
Relative Bed Stability by Region
LRBS

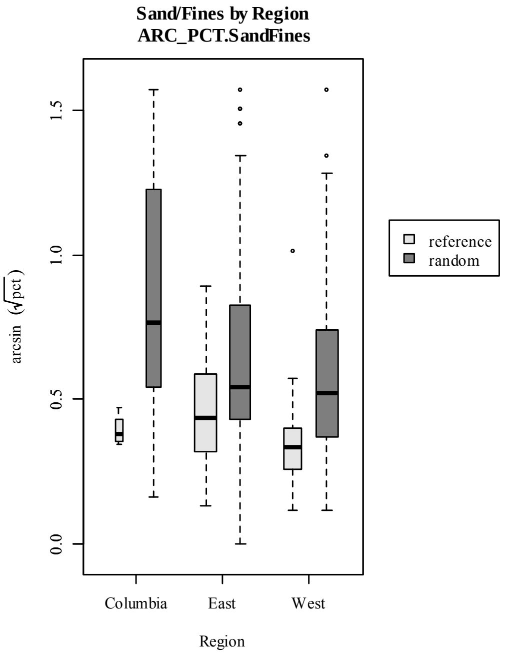

Bare Ground by Ecoregion
IDX.Ground
average %
0
20
40
60
80
100
120
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

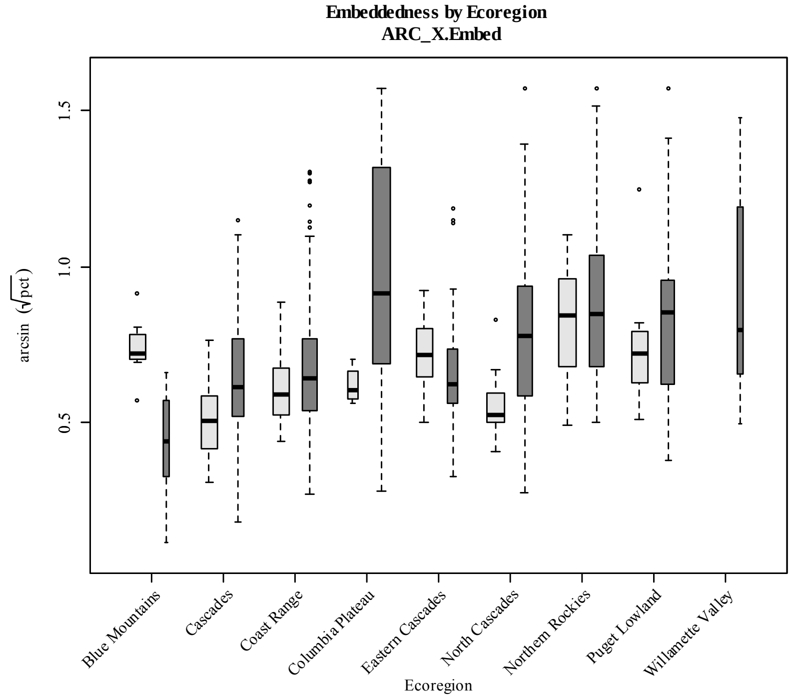

Bare Ground by Region
IDX.Ground
average %
0
20
40
60
80
100
120
Columbia
East
West
Region
reference
random

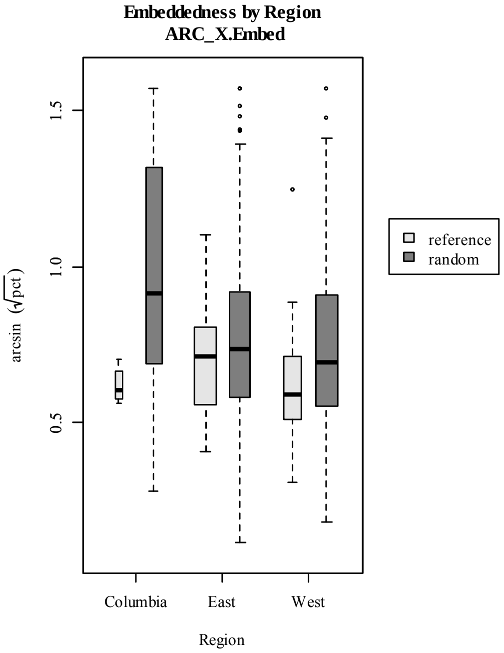

Mean Particle Size by Ecoregion
D50Log10
log10 [median particle size]
-1
0
1
2
3
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

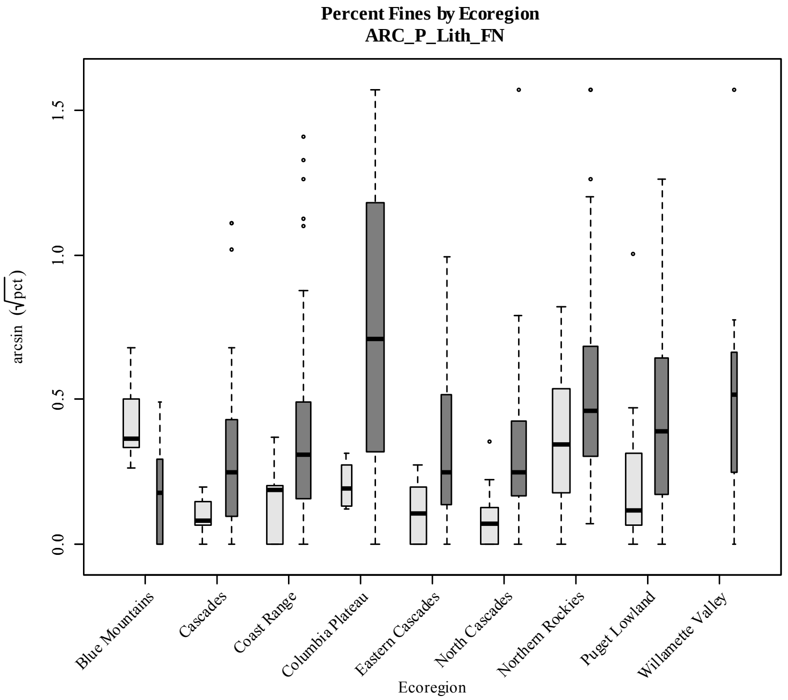

Mean Particle Size by Region
D50Log10
log10 [median particle size]
-1
0
1
2
3
Columbia
East
West
Region
reference
random

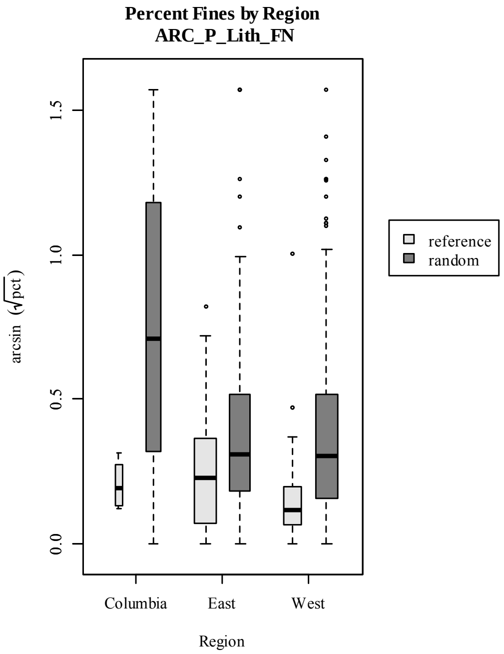

Sand/Fines by Ecoregion
ARC_PCT.SandFines
arcsin (√pct)
0.0
0.5
1.0
1.5
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

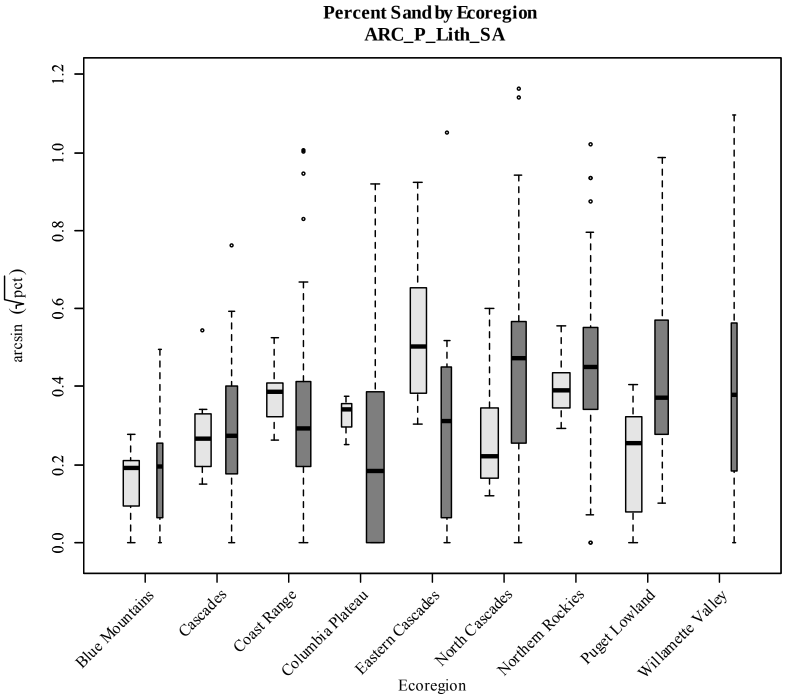

Sand/Fines by Region
ARC_PCT.SandFines
arcsin (√pct)
0.0
0.5
1.0
1.5
Columbia
East
West
Region
reference
random

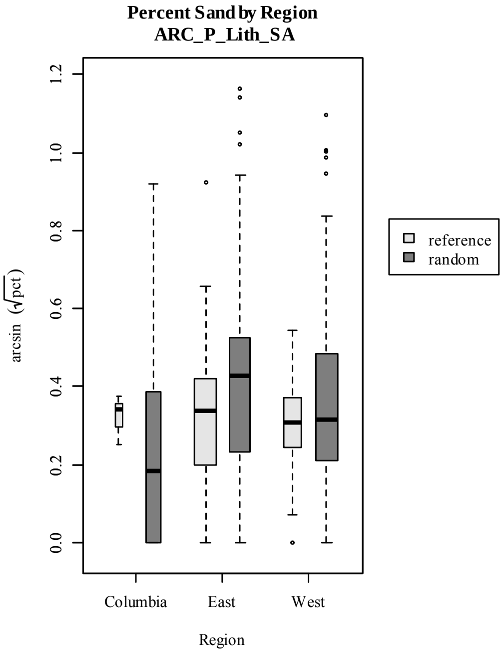

Embeddedness by Ecoregion
ARC_X.Embed
arcsin (√pct)
0.5
1.0
1.5
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

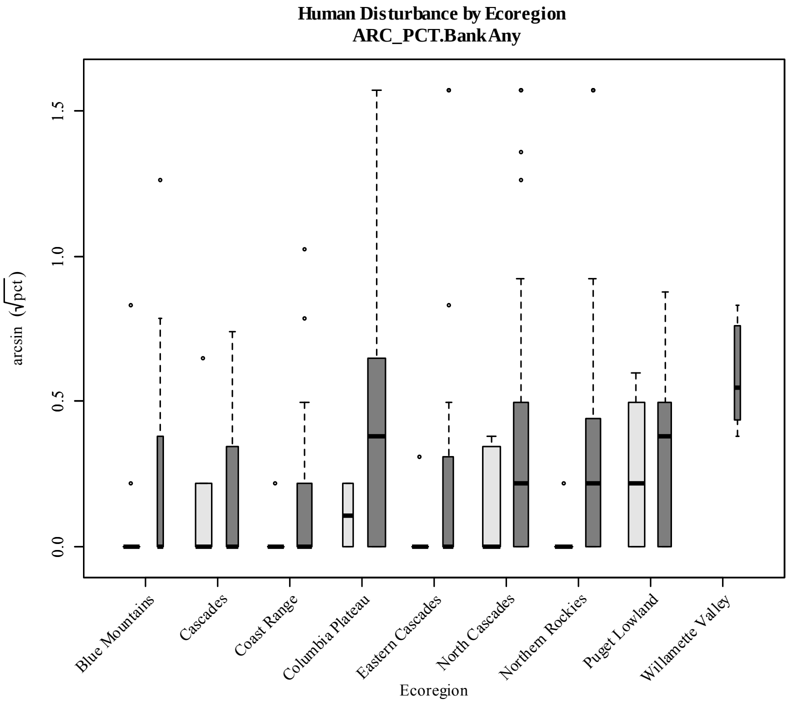

Embeddedness by Region
ARC_X.Embed
arcsin (√pct)
0.5
1.0
1.5
Columbia
East
West
Region
reference
random

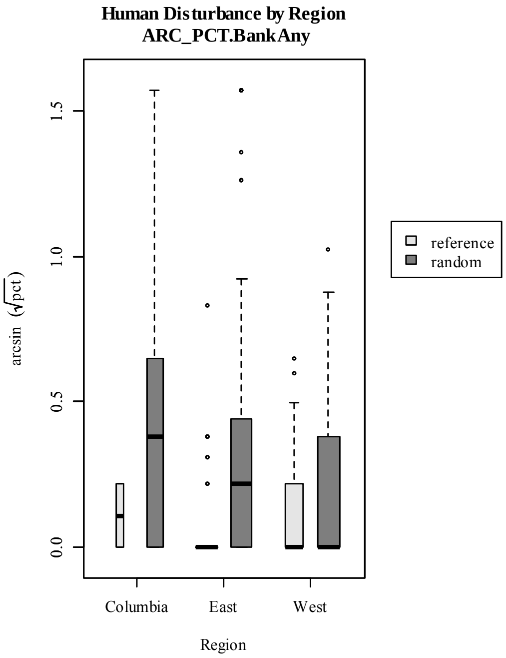

Percent Fines by Ecoregion
ARC_P_Lith_FN

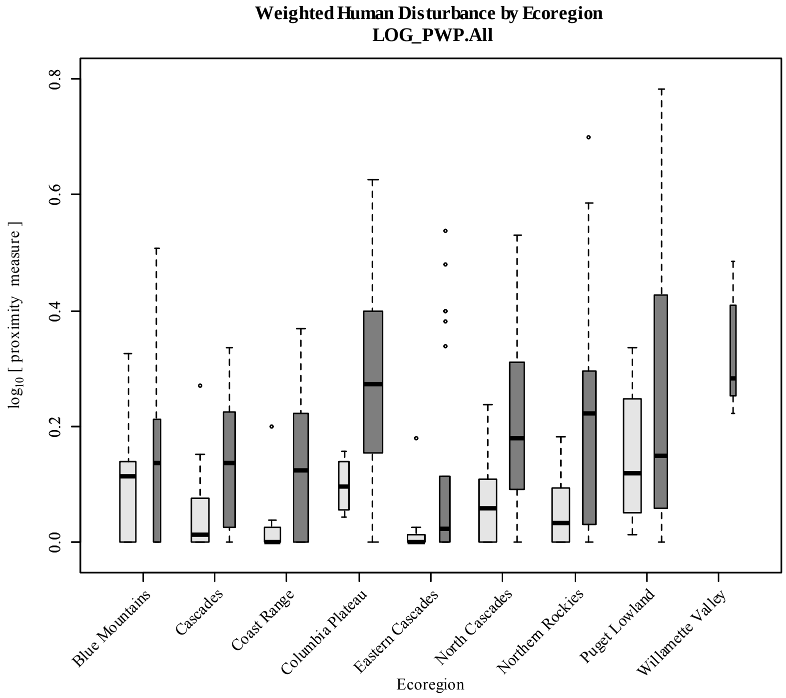

Percent Fines by Region
ARC_P_Lith_FN

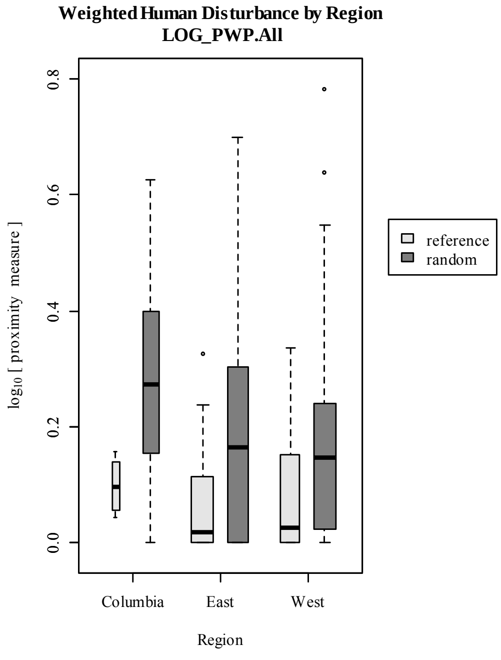

Percent Sand by Ecoregion
ARC_P_Lith_SA

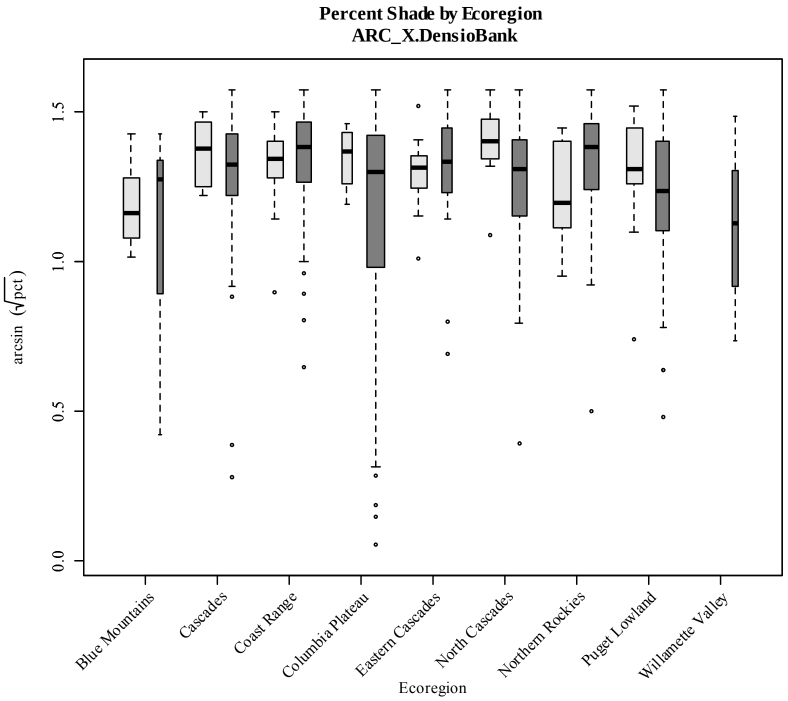

Percent Sand by Region
ARC_P_Lith_SA

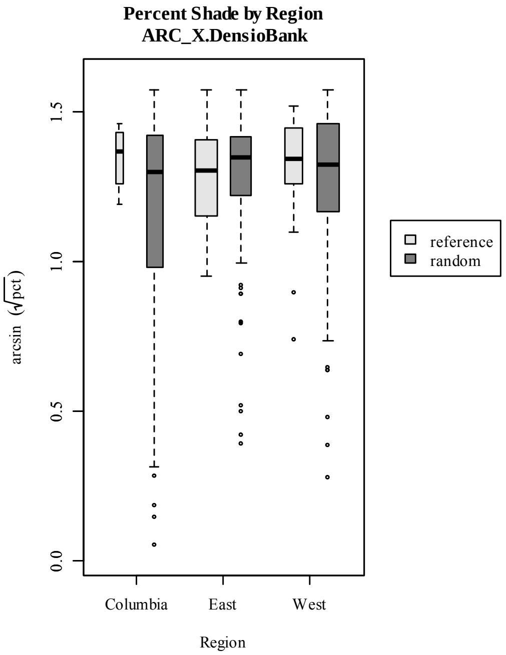

Human Disturbance by Ecoregion
ARC_PCT.BankAny

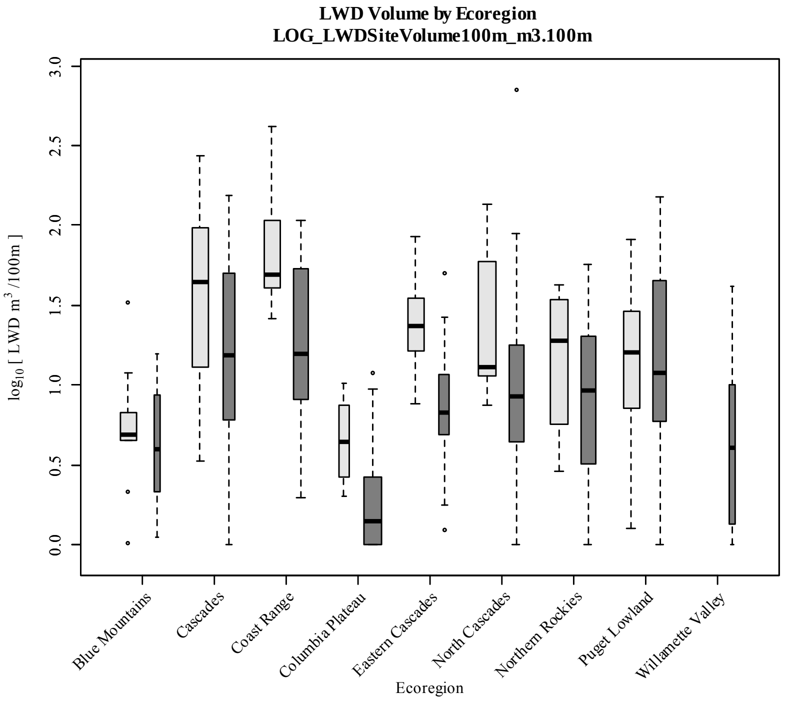

Human Disturbance by Region
ARC_PCT.BankAny

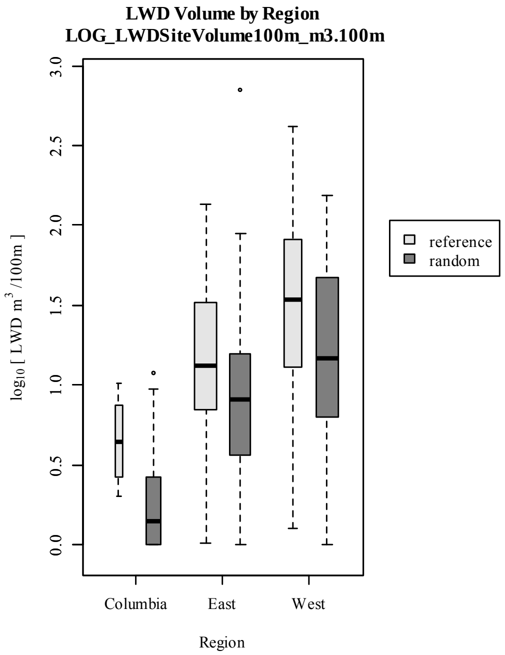

Weighted Human Disturbance by Ecoregion
LOG_PWP.All

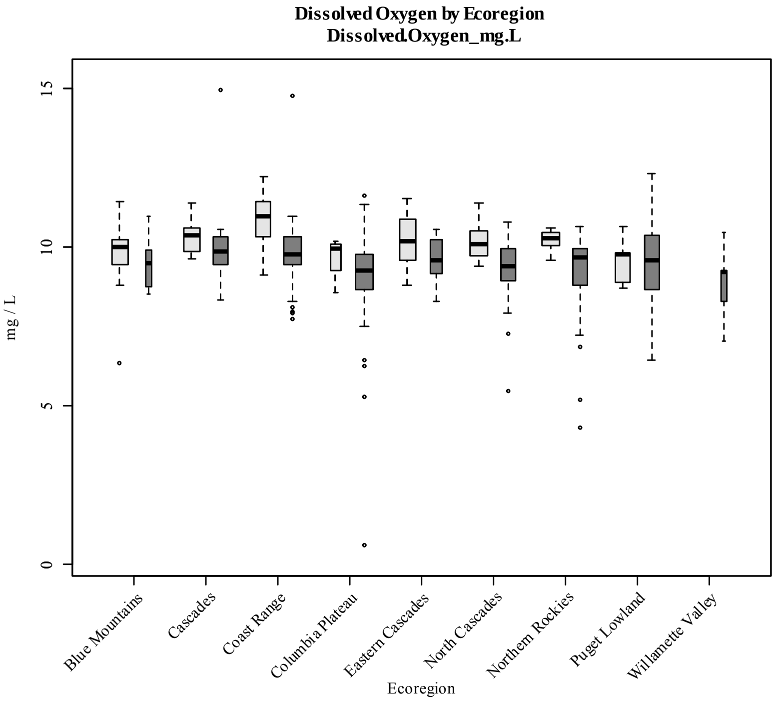

Weighted Human Disturbance by Region
LOG_PWP.All

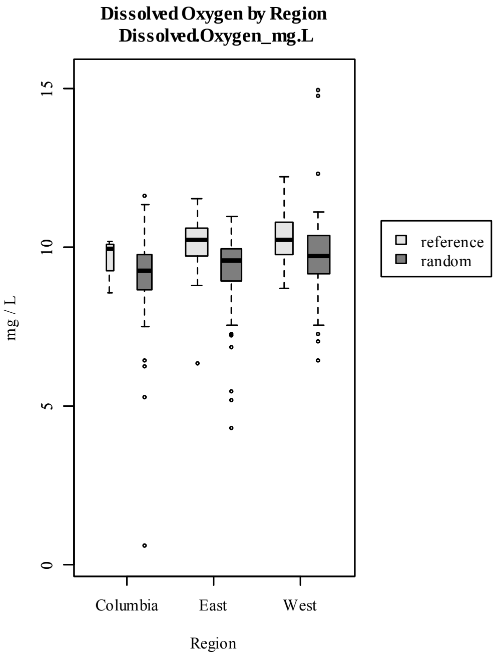

Percent Shade by Ecoregion
ARC_X.DensioBank

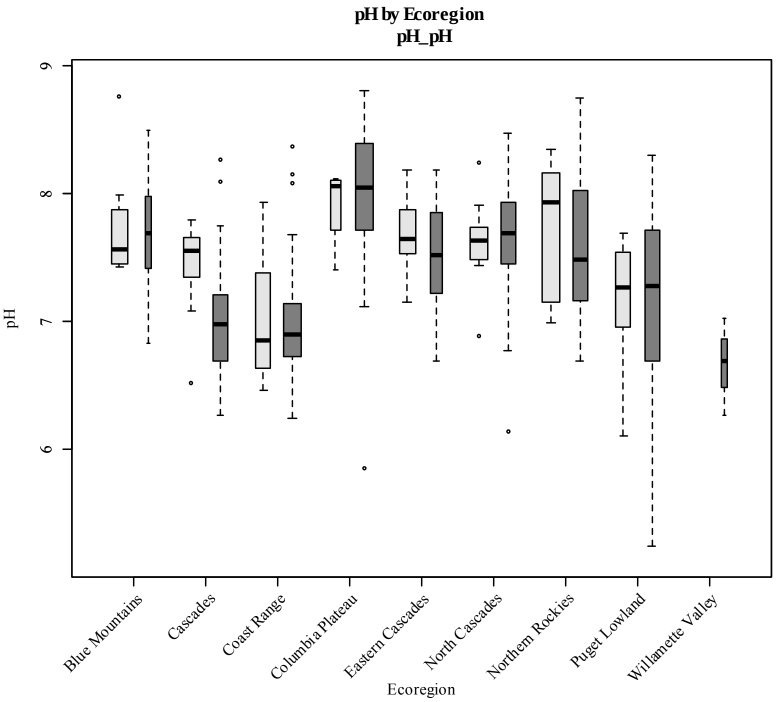

Percent Shade by Region
ARC_X.DensioBank

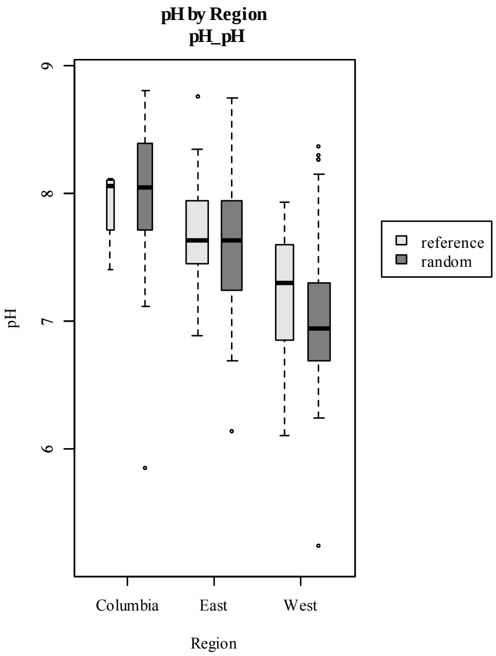

LWD Volume by Ecoregion
LOG_LWDSiteVolume100m_m3.100m

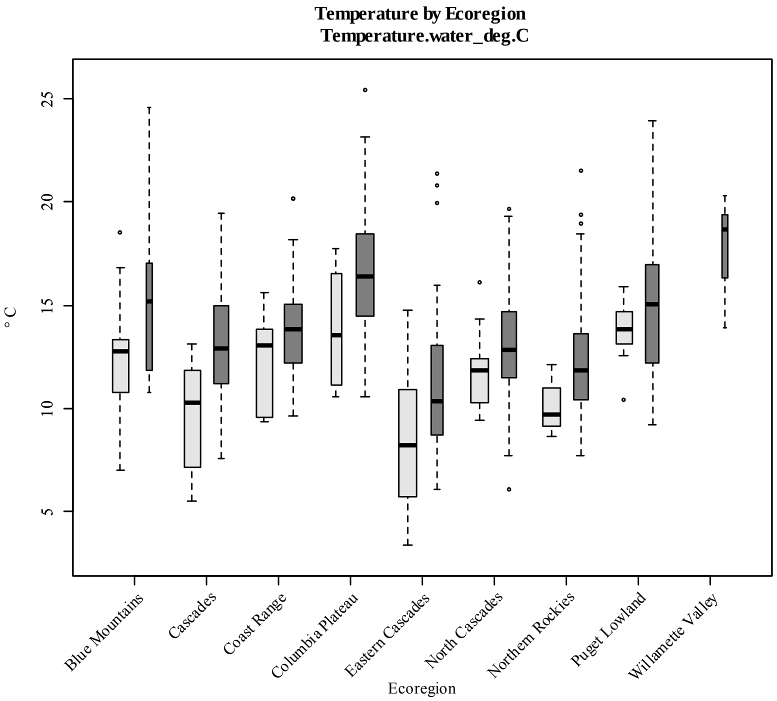

LWD Volume by Region
LOG_LWDSiteVolume100m_m3.100m

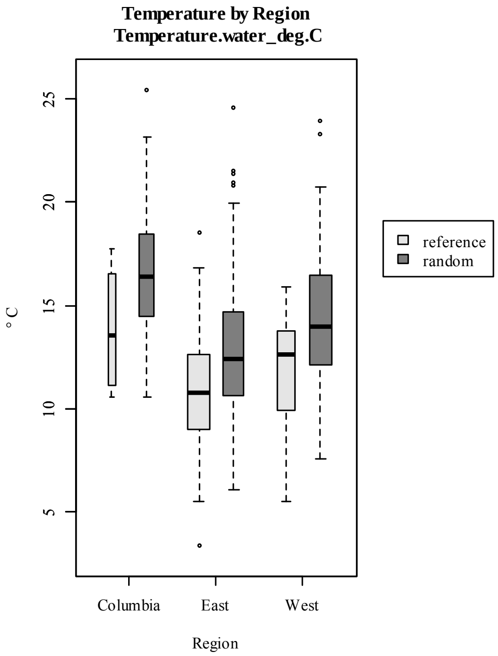

Dissolved Oxygen by Ecoregion
Dissolved.Oxygen_mg.L

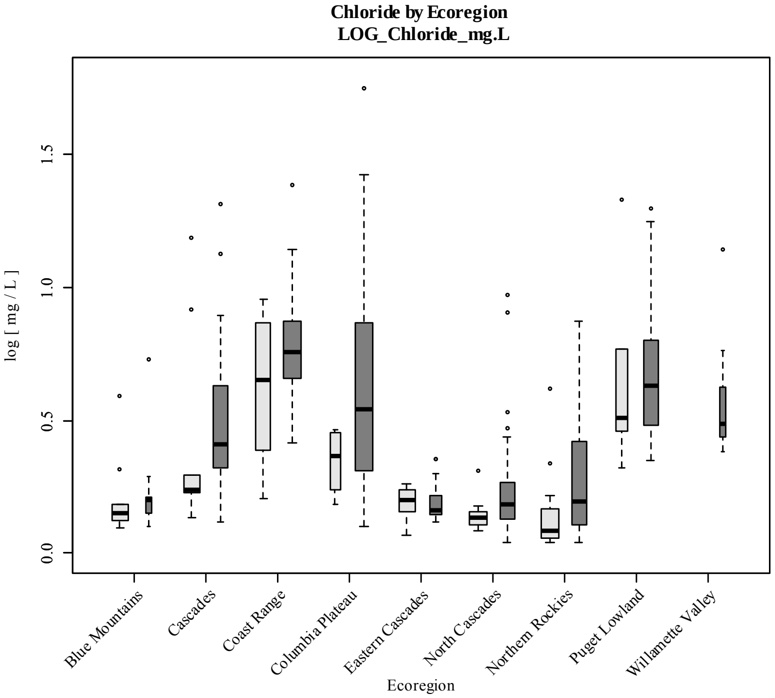

Dissolved Oxygen by Region
Dissolved.Oxygen_mg.L

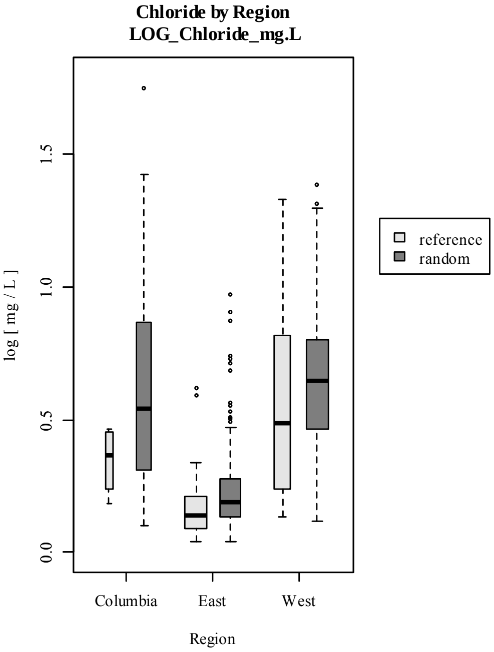

pH by Ecoregion
pH_pH

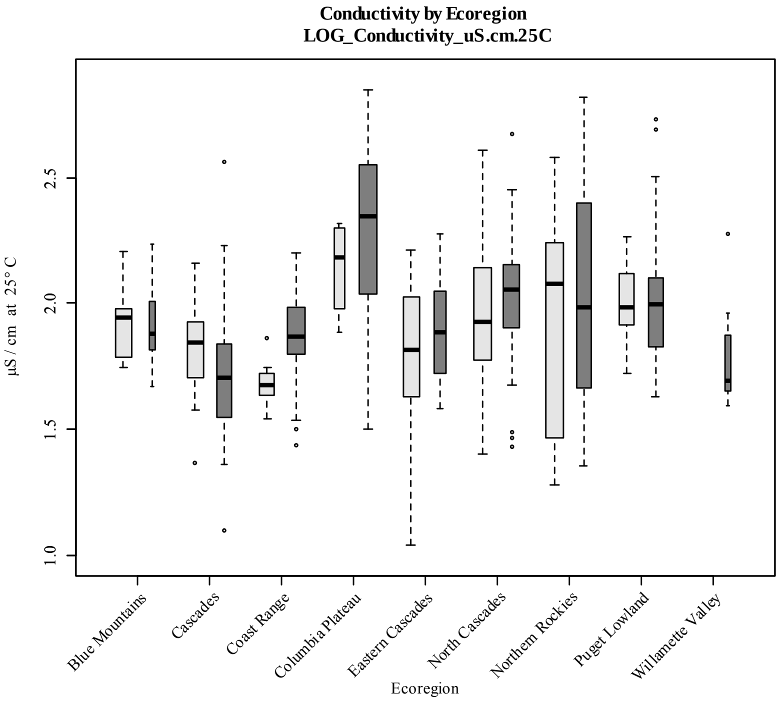

pH by Region
pH_pH


Temperature by Ecoregion
Temperature.water_deg.C
° C
5
10
15
20
25
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

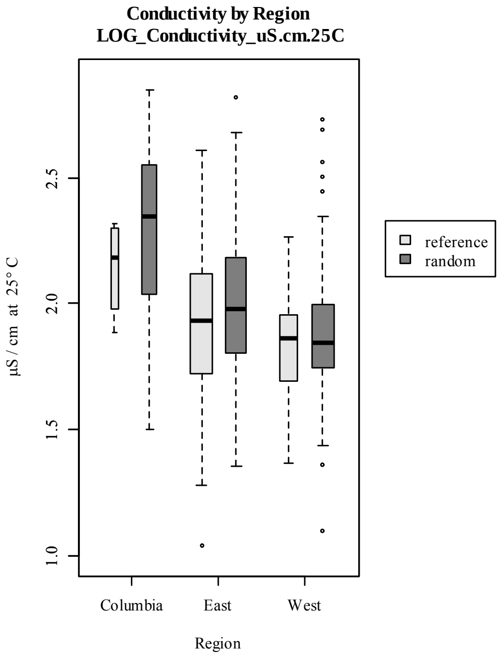

Temperature by Region
Temperature.water_deg.C
° C
5
10
15
20
25
Columbia
East
West
Region
reference
random


Chloride by Ecoregion
LOG_Chloride_mg.L
log [mg / L]
0.0
0.5
1.0
1.5
Blue Mountains
Cascades
Coast Range
Columbia Plateau
Eastern Cascades
North Cascades
Northern Rockies
Puget Lowland
Willamette Valley
Ecoregion

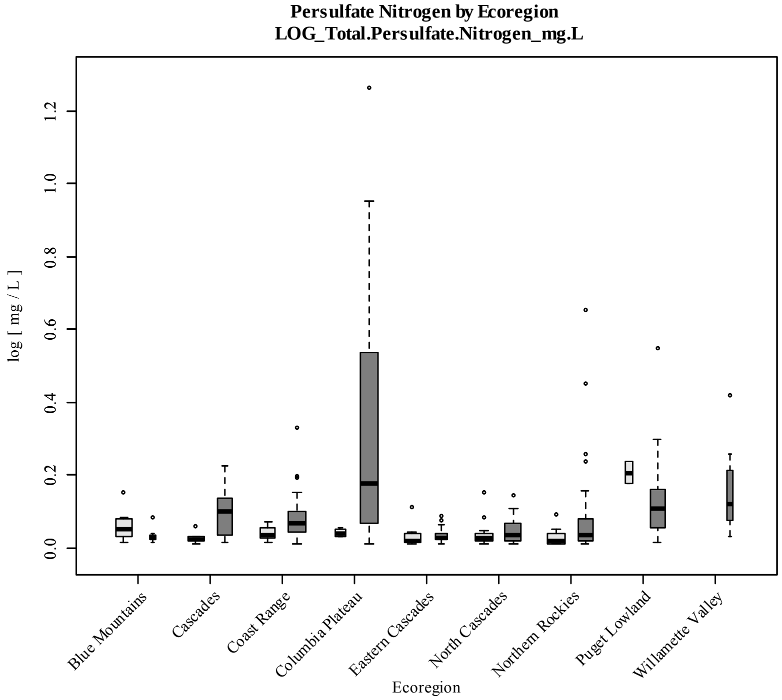

Chloride by Region
LOG_Chloride_mg.L
log [mg / L]
0.0
0.5
1.0
1.5
Columbia
East
West
Region
reference
random



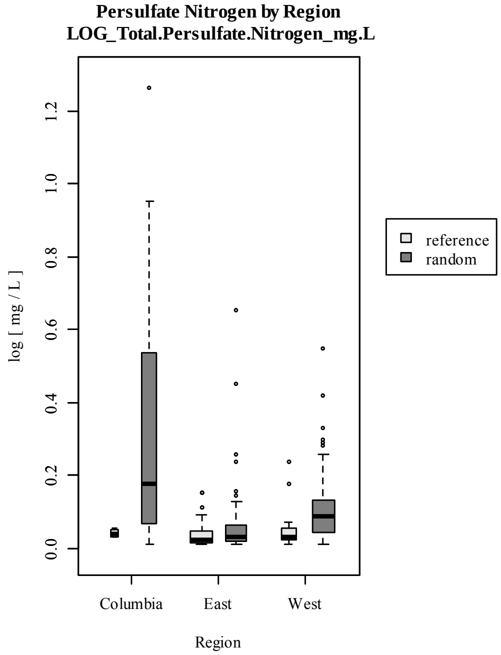
Conductivity by Ecoregion
LOG_Conductivity_uS.cm.25C



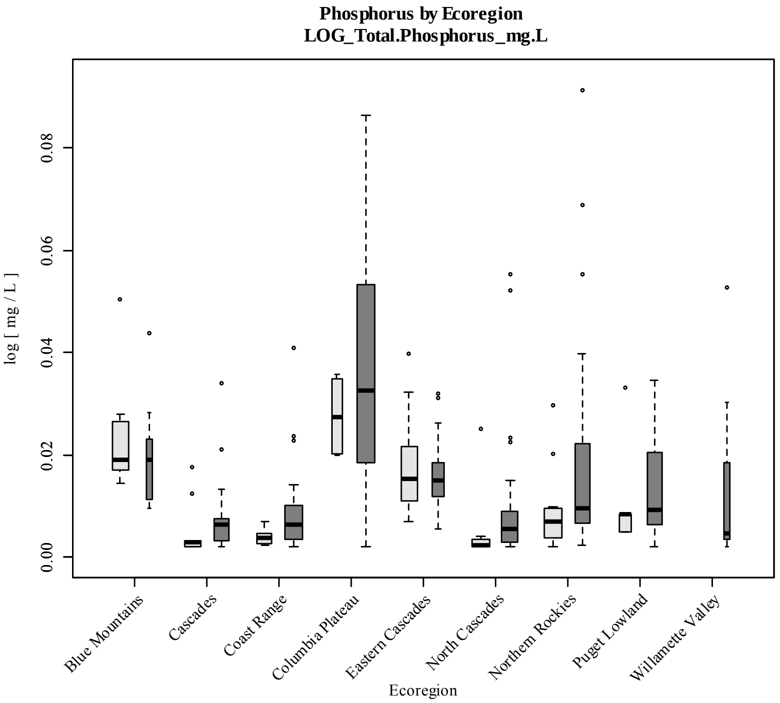
Conductivity by Region
LOG_Conductivity_uS.cm.25C



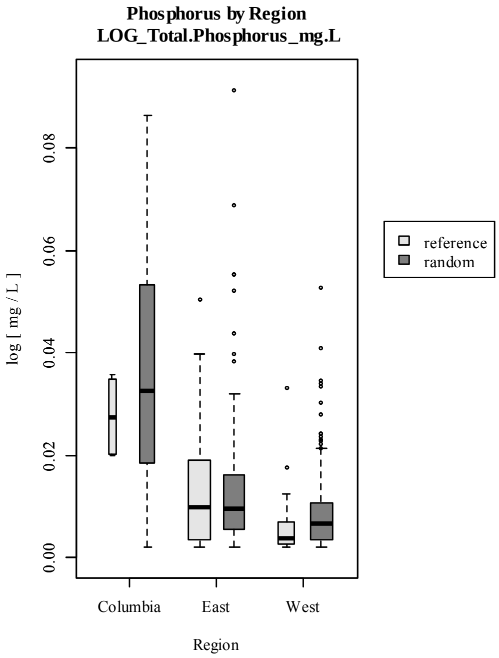
Persulfate Nitrogen by Ecoregion
LOG_Total.Persulfate.Nitrogen_mg.L



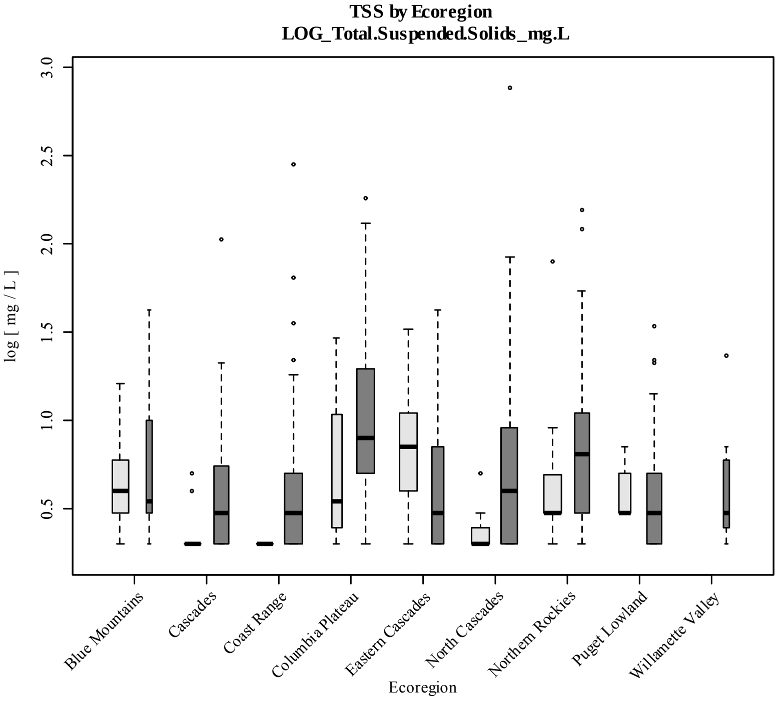
Persulfate Nitrogen by Region
LOG_Total.Persulfate.Nitrogen_mg.L



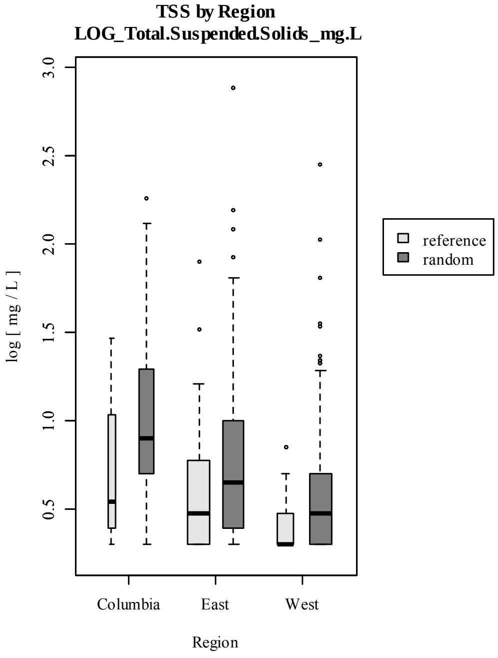
Phosphorus by Ecoregion
LOG_Total.Phosphorus_mg.L



Phosphorus by Region
LOG_Total.Phosphorus_mg.L



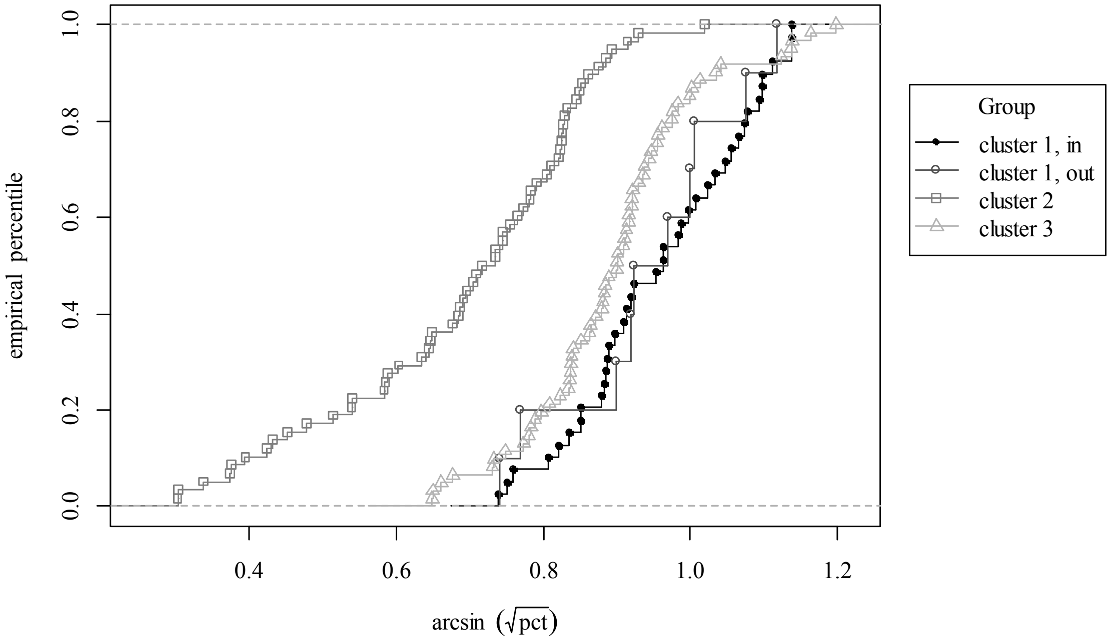
TSS by Ecoregion
LOG_Total.Suspended.Solids_mg.L



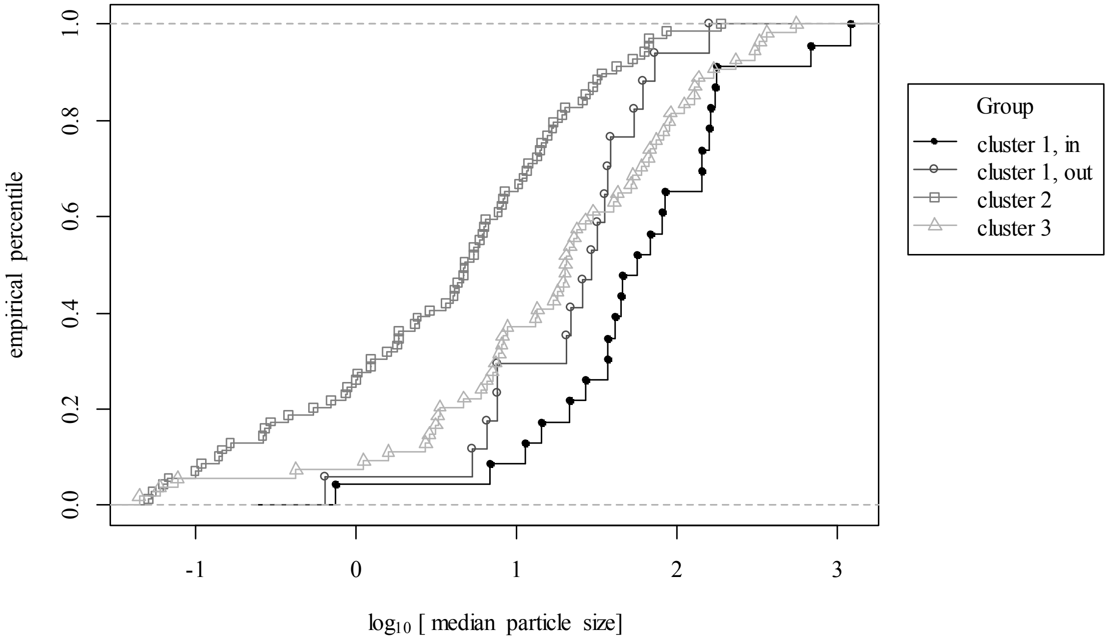
TSS by Region
LOG_Total.Suspended.Solids_mg.L

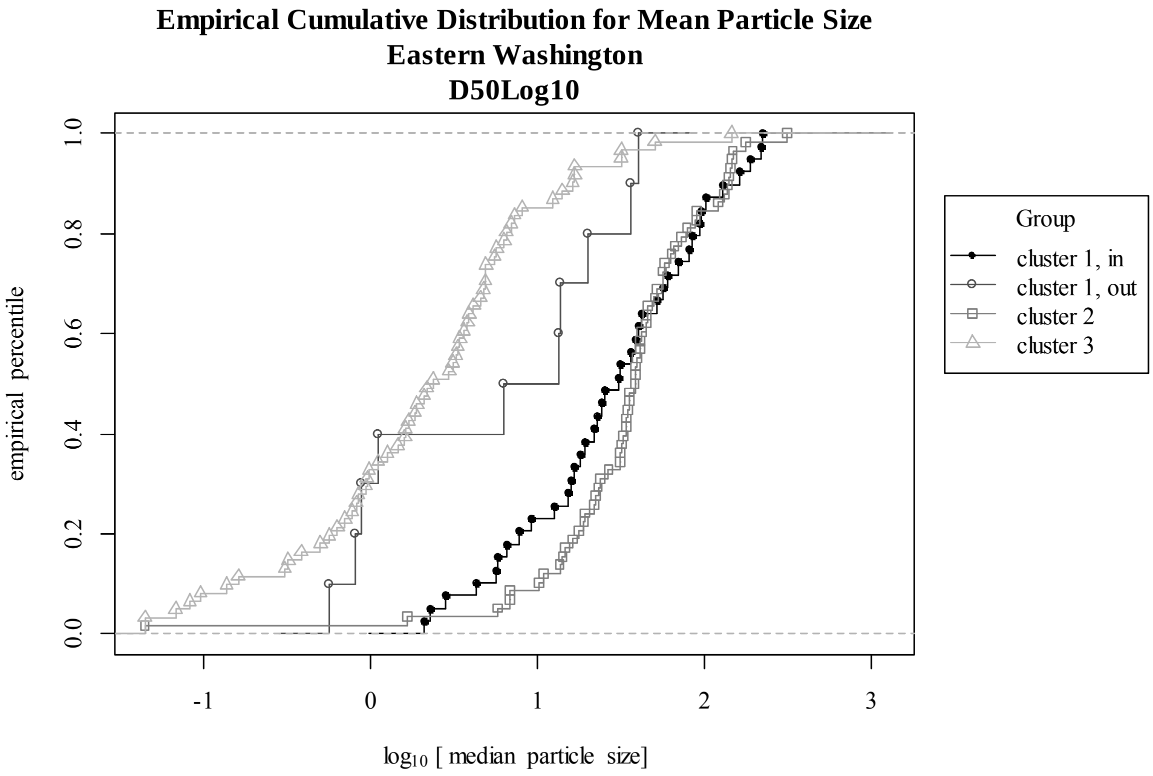
Appendix D. Empirical Cumulative Distribution Functions

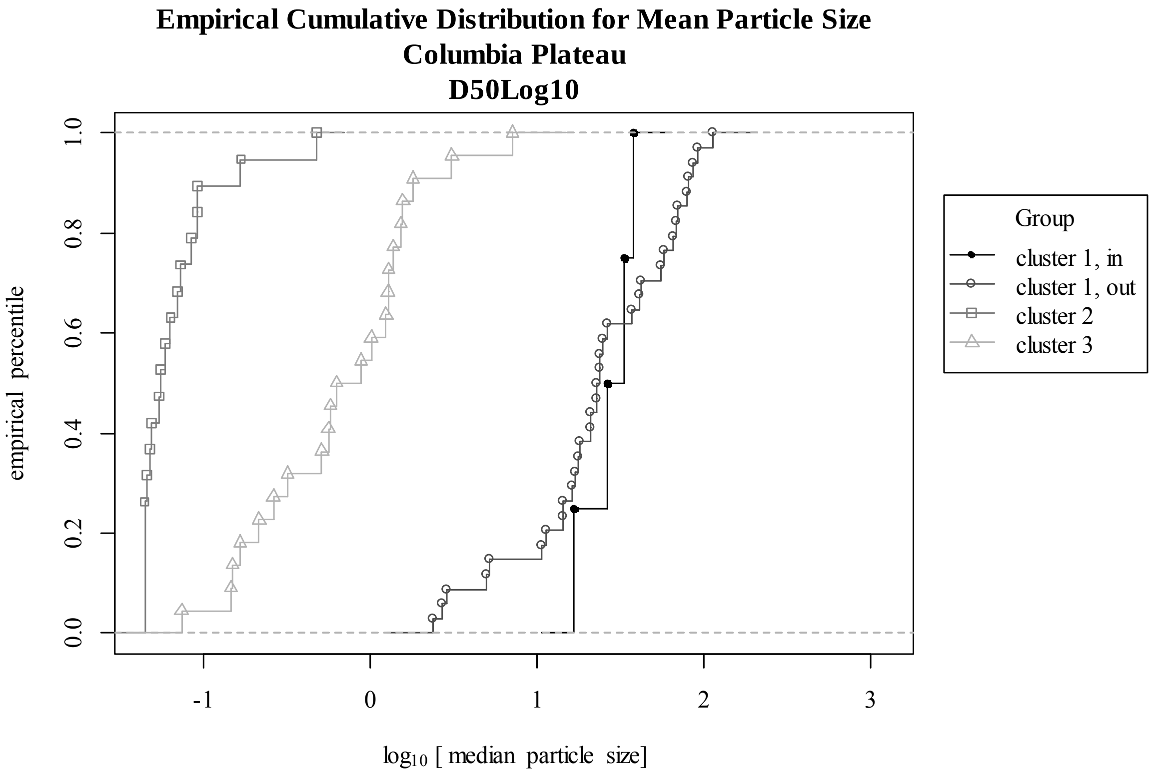
Empirical Cumulative Distribution for Percent Canopy
Eastern Washington
ARC_CANOPY_PCT

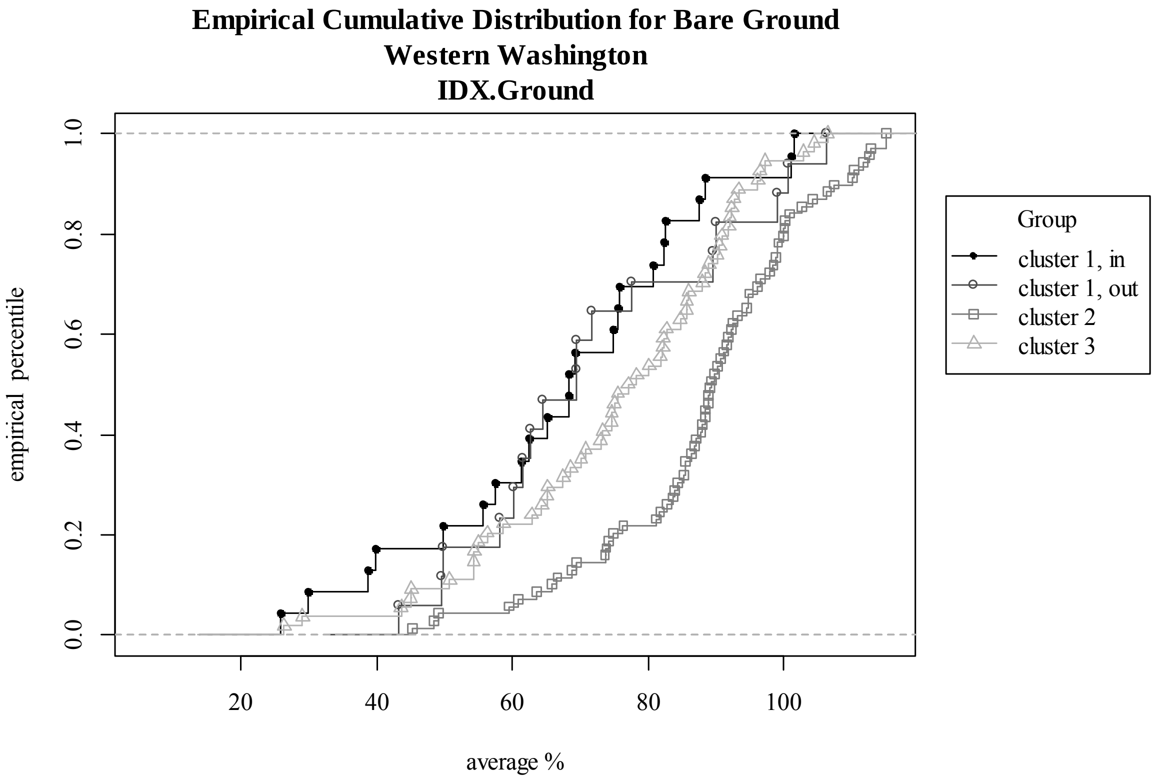


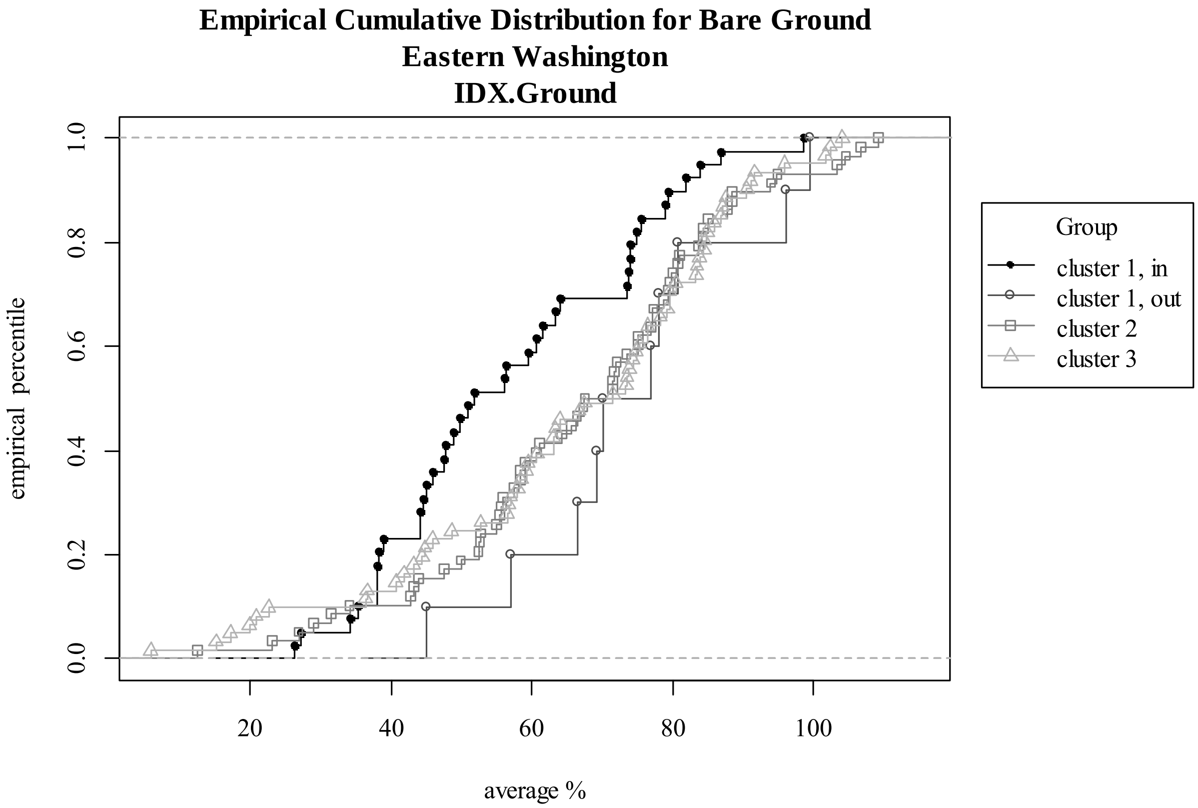
Empirical Cumulative Distribution for Mean Particle Size
Western Washington
D50Log10

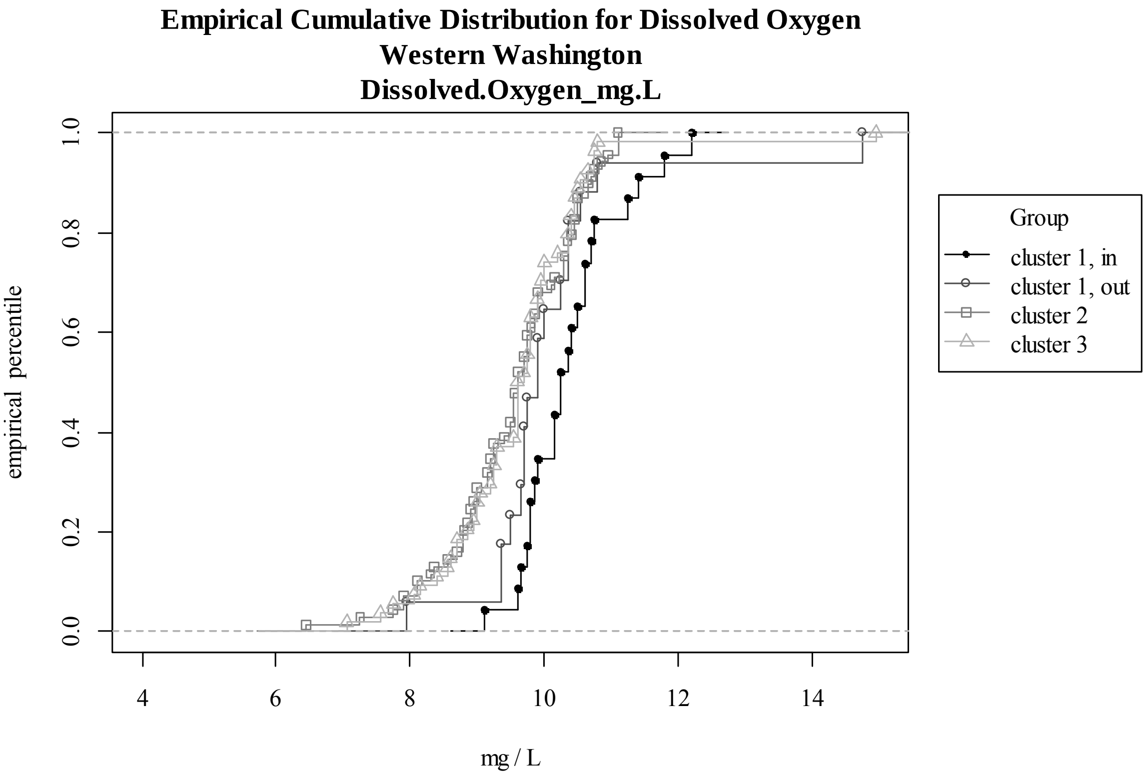


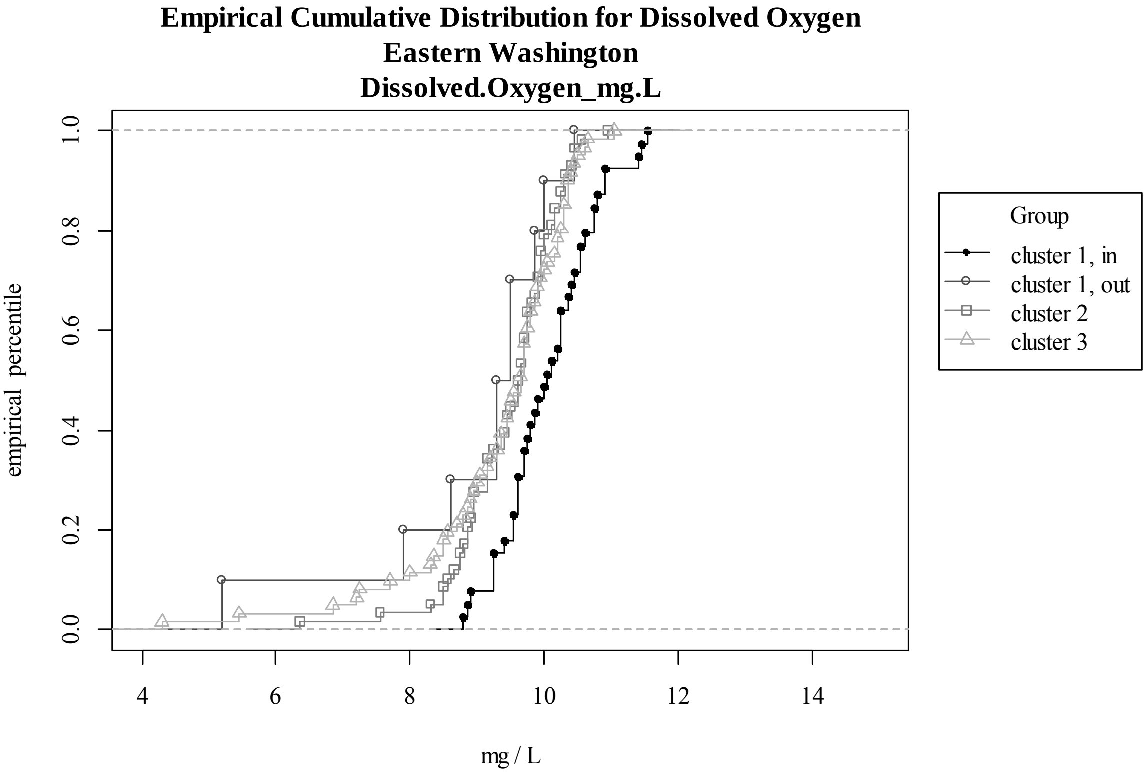

Empirical Cumulative Distribution for Mean Particle Size
Eastern Washington
D50Log10
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
-1
0
1
2
3
log10 [median particle size]
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3

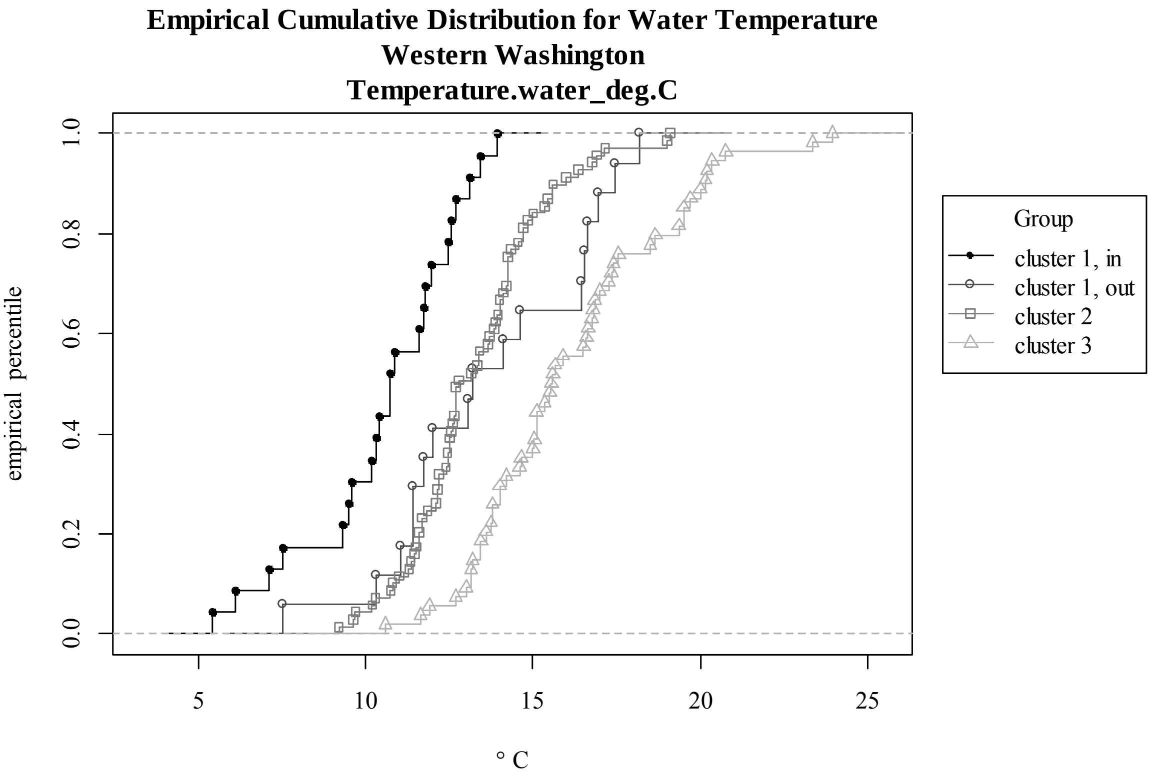

Empirical Cumulative Distribution for Mean Particle Size
Columbia Plateau
D50Log10
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
-1
0
1
2
3
log10 [median particle size]
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3


Empirical Cumulative Distribution for Bare Ground
Western Washington
IDX.Ground

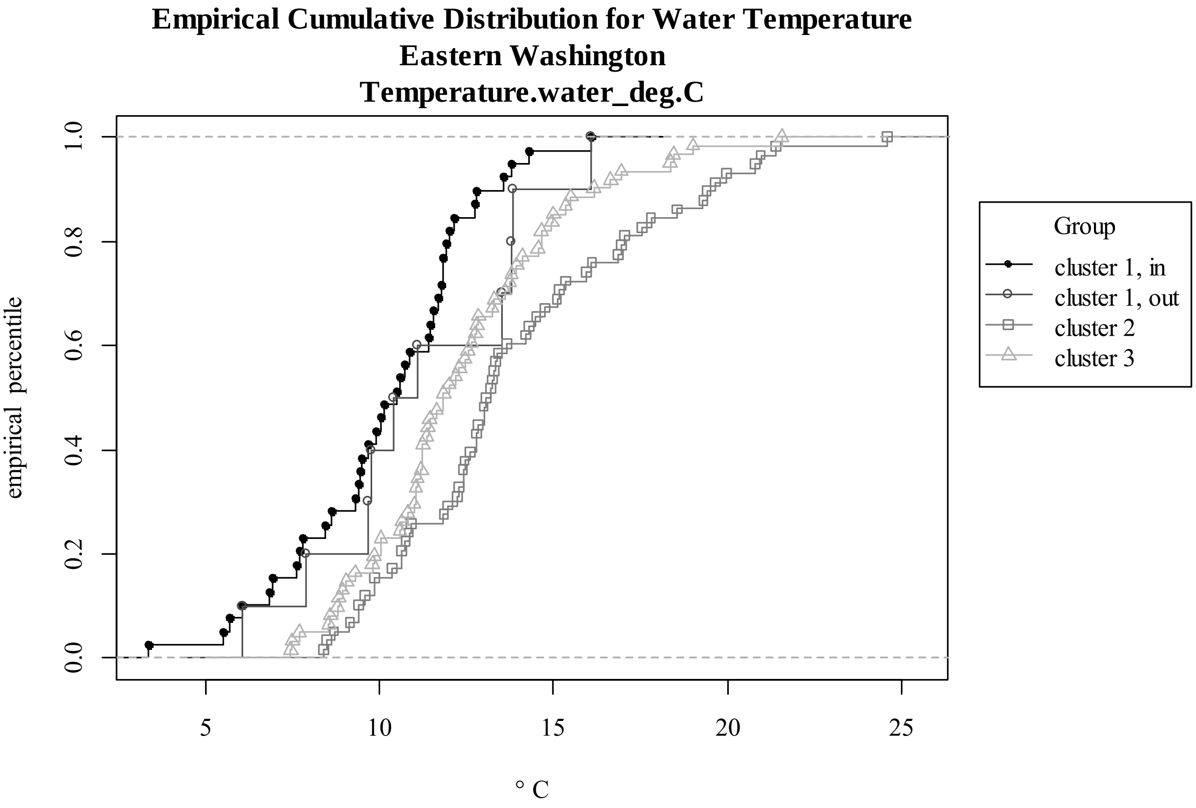

Empirical Cumulative Distribution for Bare Ground
Eastern Washington
IDX.Ground


Empirical Cumulative Distribution for Dissolved Oxygen
Western Washington
Dissolved.Oxygen_mg.L
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
mg / L
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3

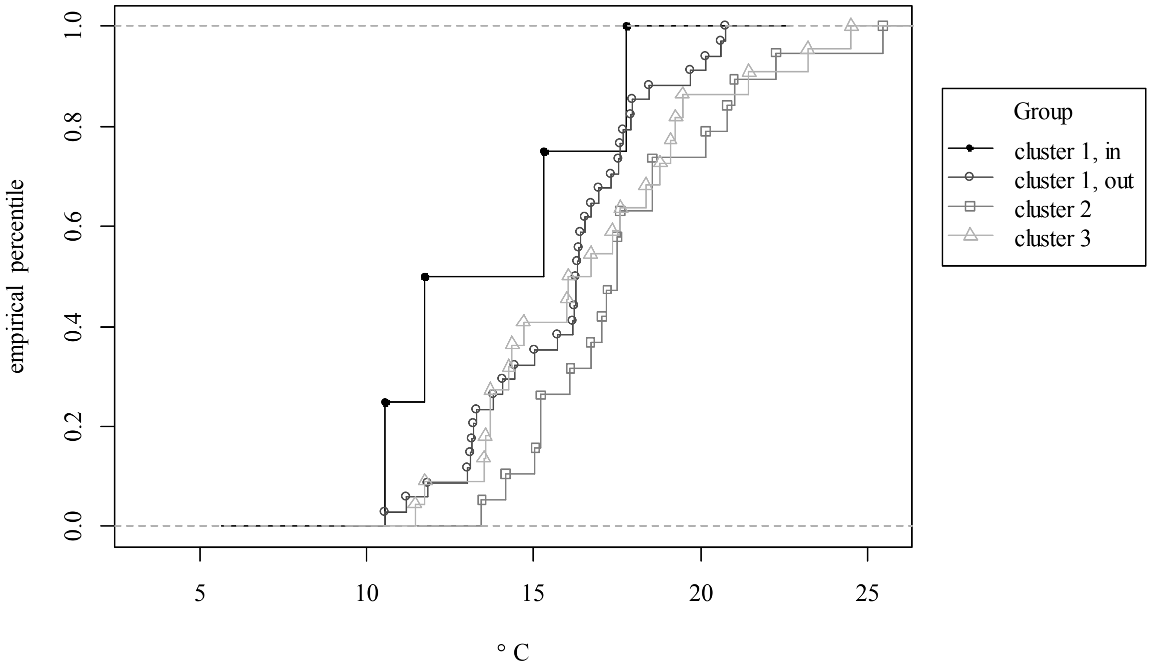

Empirical Cumulative Distribution for Dissolved Oxygen
Eastern Washington
Dissolved.Oxygen_mg.L
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
mg / L
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3

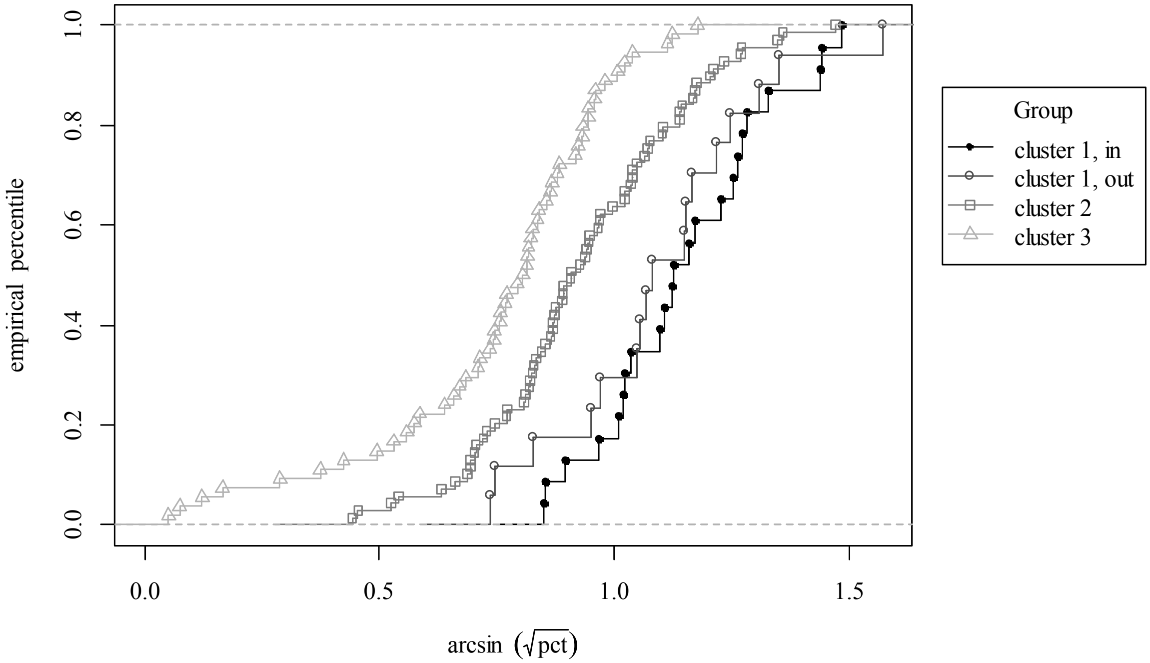


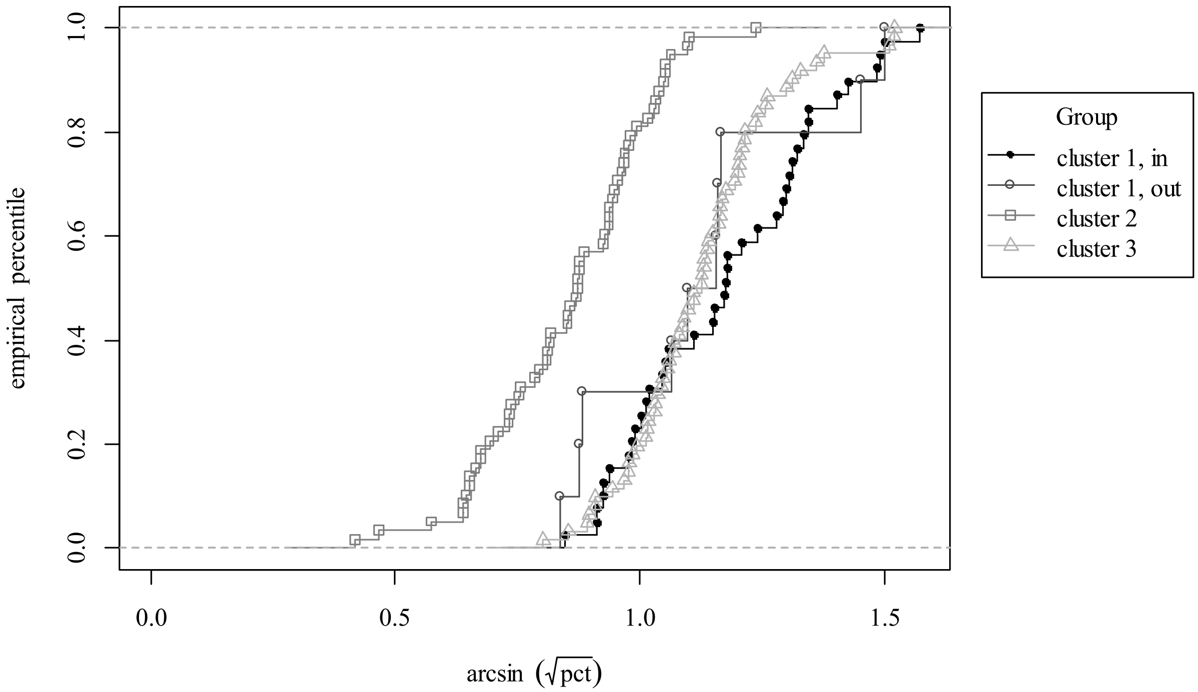
Empirical Cumulative Distribution for Water Temperature
Western Washington
Temperature.water_deg.C



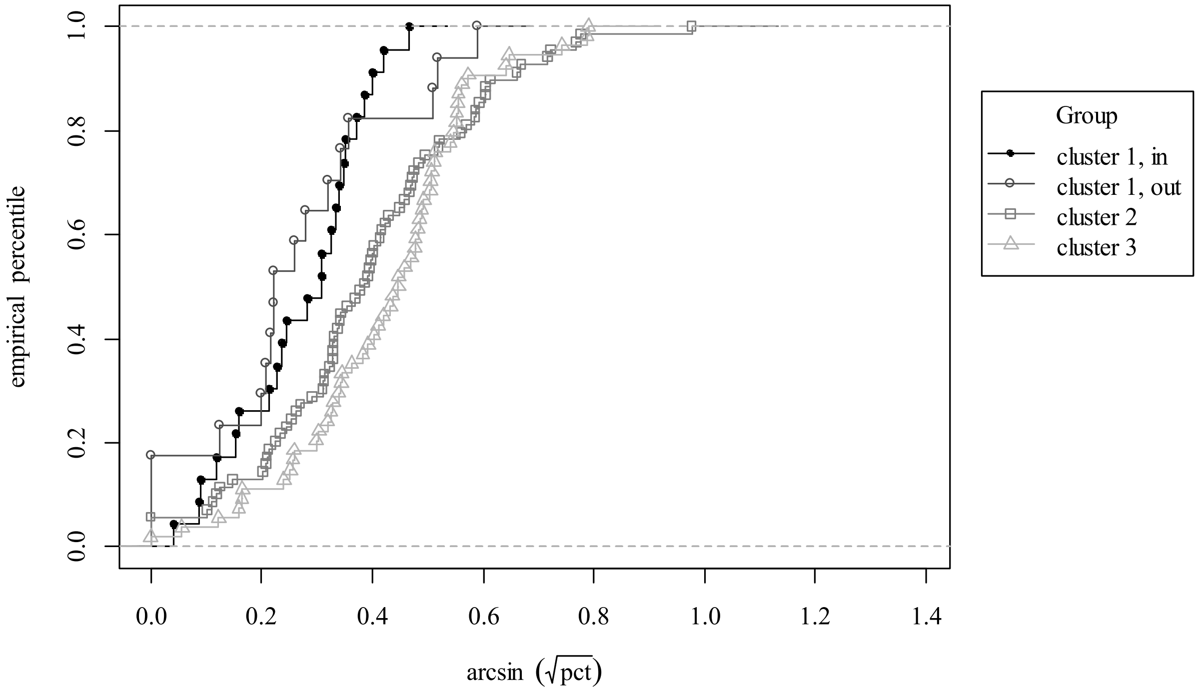
Empirical Cumulative Distribution for Water Temperature
Eastern Washington
Temperature.water_deg.C


Empirical Cumulative Distribution for Water Temperature
Columbia Plateau, Washington
Temperature.water_deg.C


Empirical Cumulative Distribution for Percent Evergreen Forest
Western Washington
ARC_evergreen.forest.prcnt

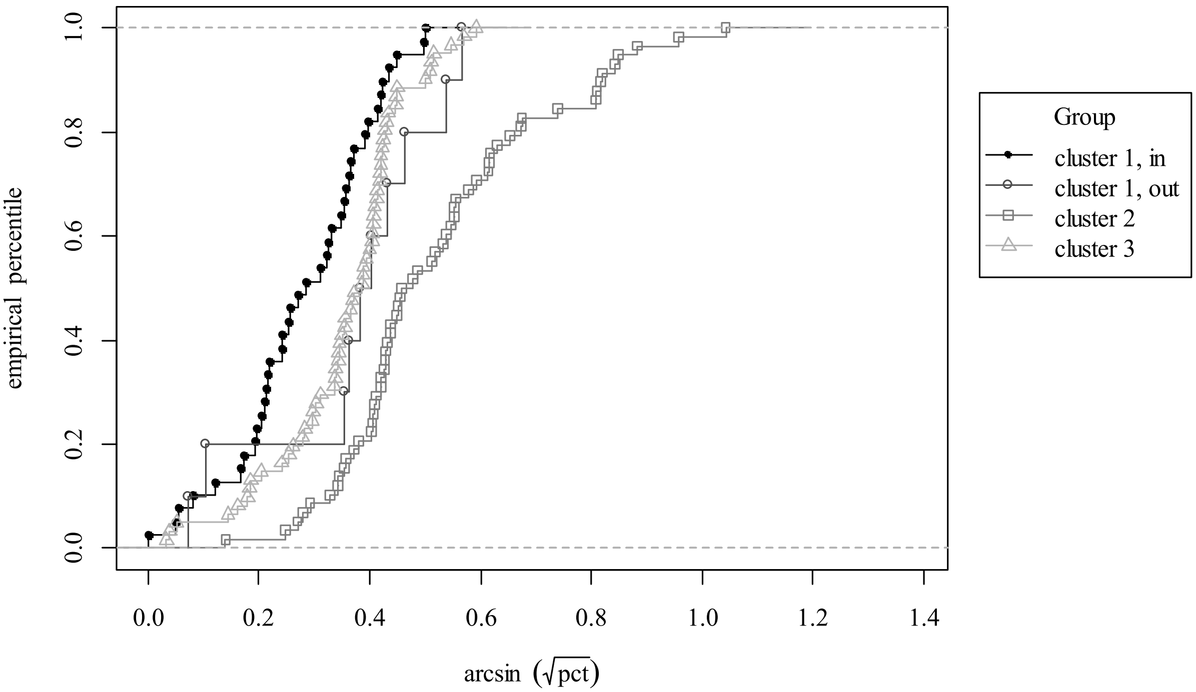


Empirical Cumulative Distribution for Percent Evergreen Forest
Eastern Washington
ARC_evergreen.forest.prcnt

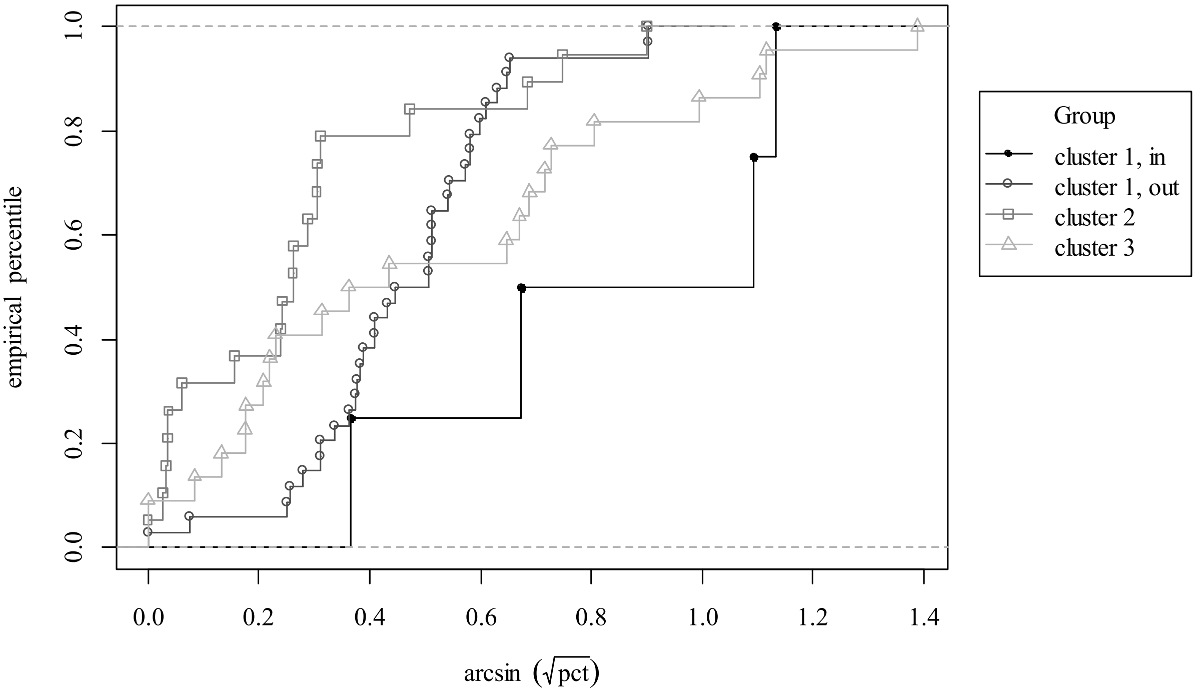


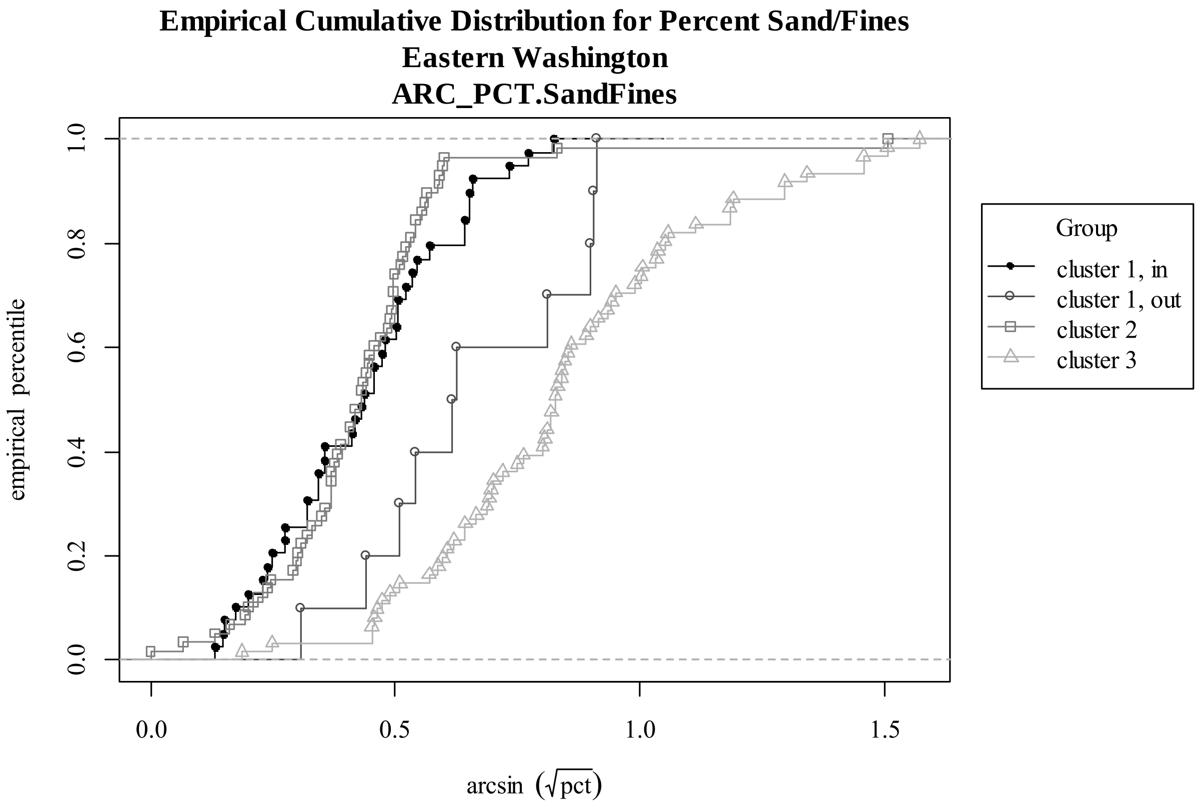
Empirical Cumulative Distribution for Percent Shrub/Scrub
Western Washington
ARC_shrub.scrub.prcnt

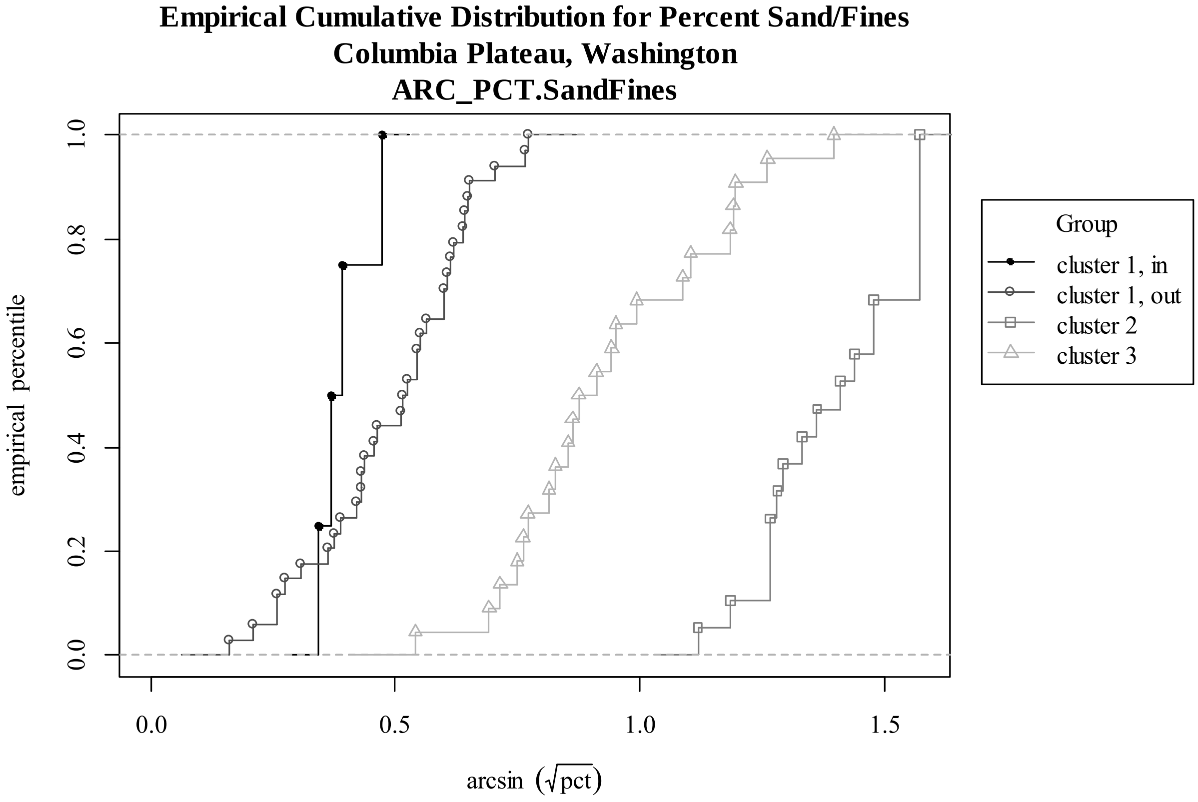
Empirical Cumulative Distribution for Percent Shrub/Scrub
Eastern Washington
ARC_shrub.scrub.prcnt

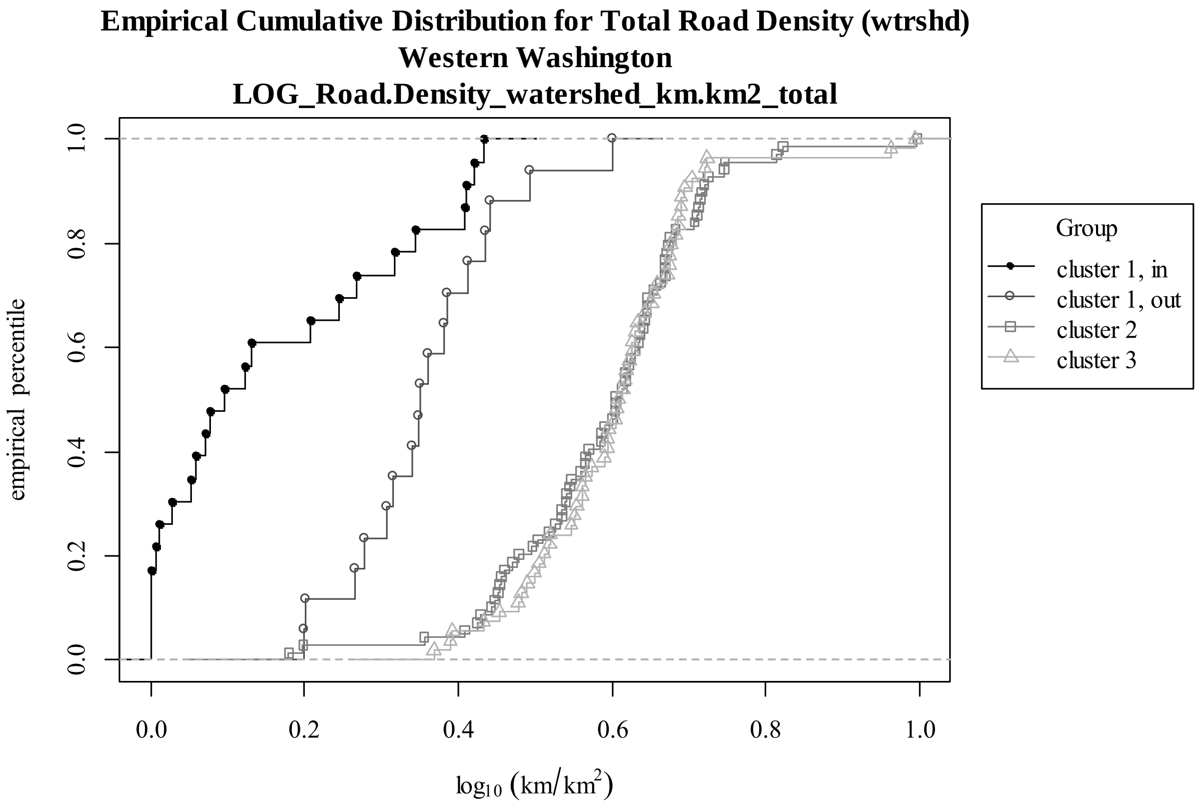


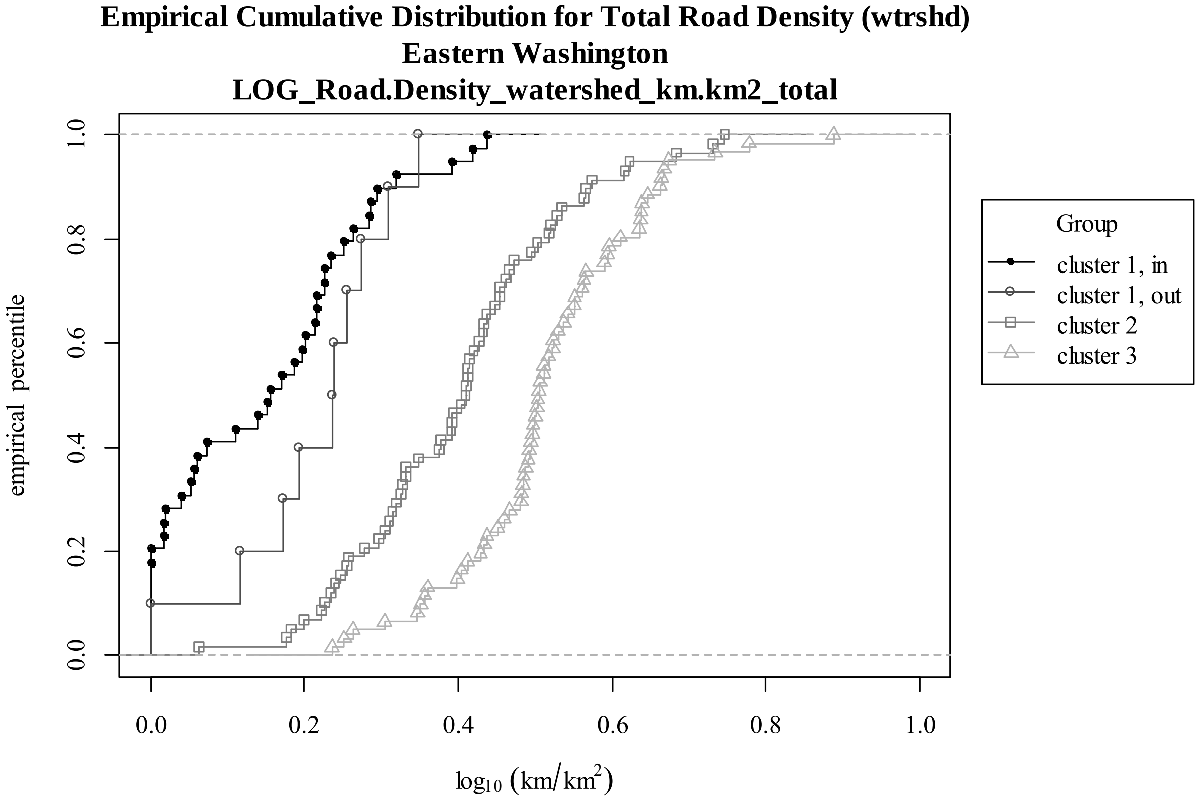
Empirical Cumulative Distribution for Percent Shrub/Scrub
Columbia Plateau, Washington
ARC_shrub.scrub.prcnt



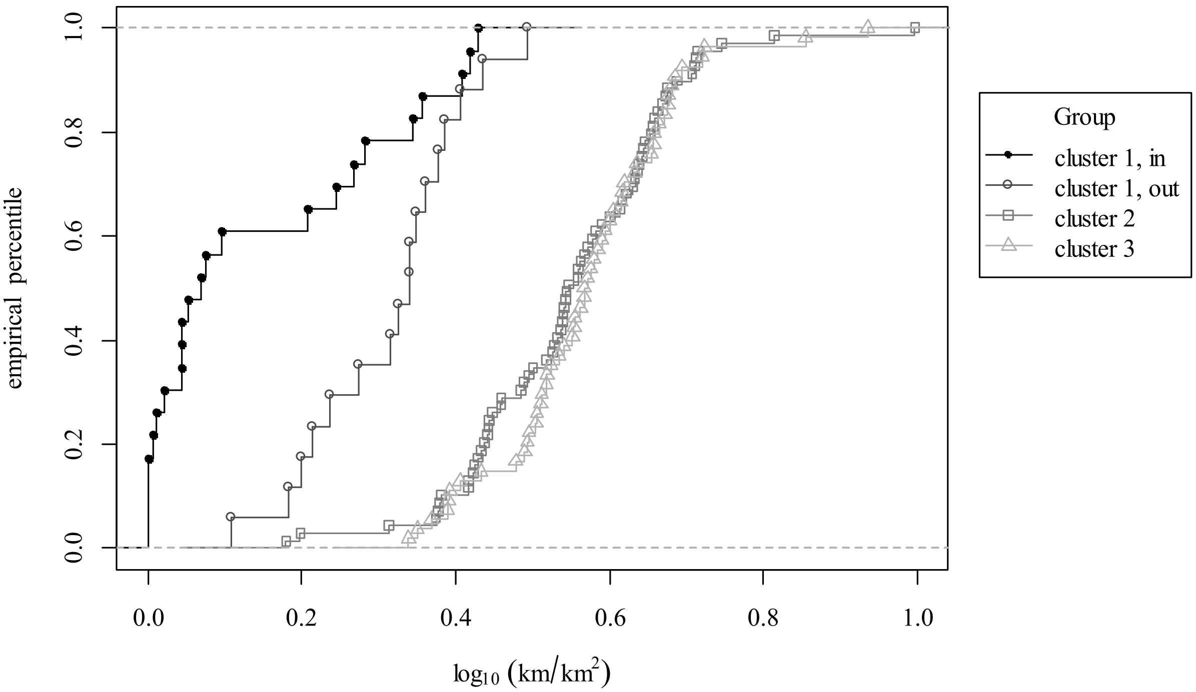

Empirical Cumulative Distribution for Percent Sand/Fines
Eastern Washington
ARC_PCT.SandFines


Empirical Cumulative Distribution for Percent Sand/Fines
Columbia Plateau, Washington
ARC_PCT.SandFines

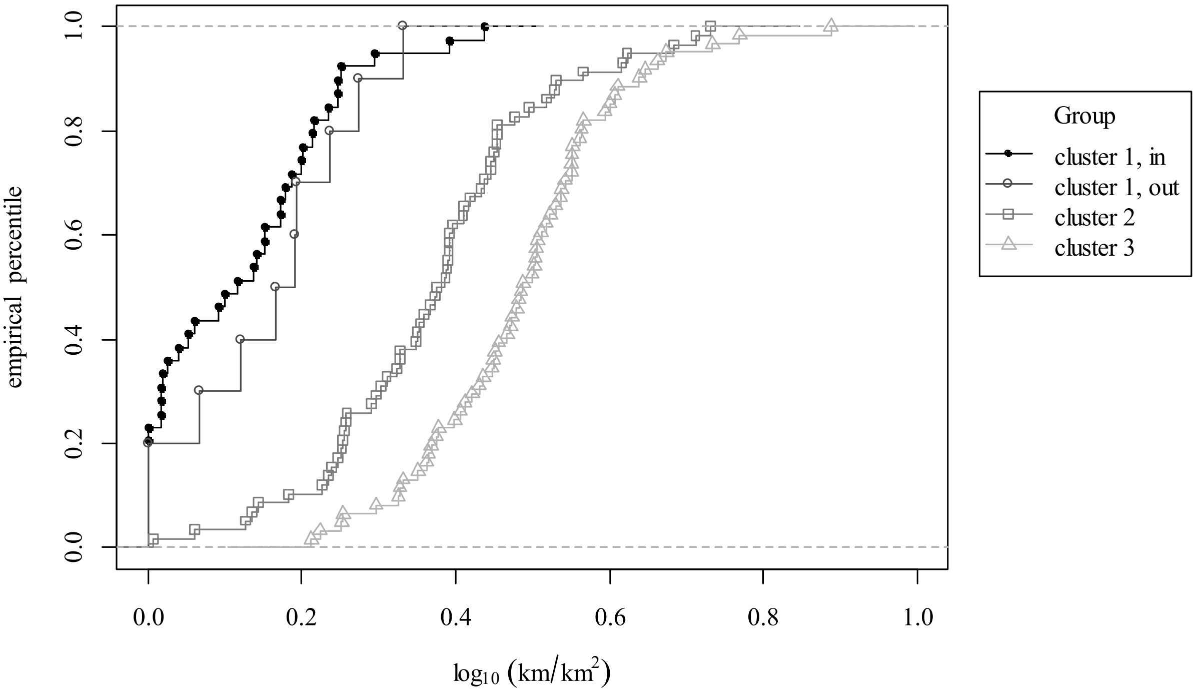

Empirical Cumulative Distribution for Total Road Density (wtrshd)
Western Washington
LOG_Road.Density_watershed_km.km2_total

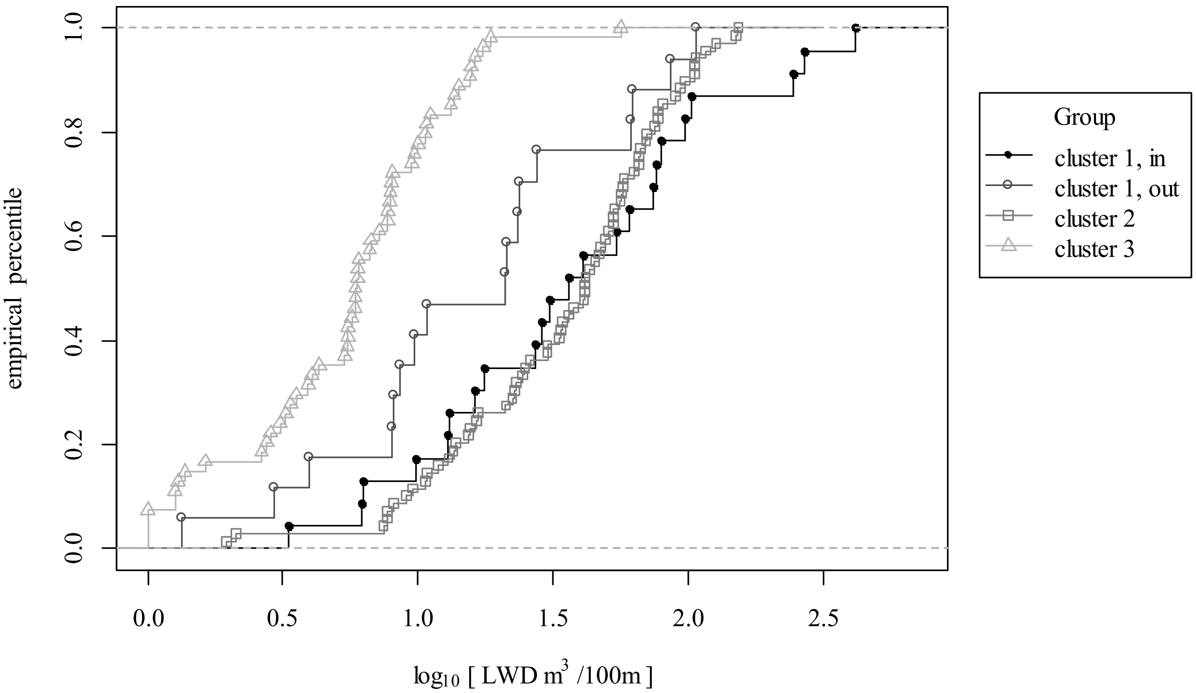

Empirical Cumulative Distribution for Total Road Density (wtrshd)
Eastern Washington
LOG_Road.Density_watershed_km.km2_total

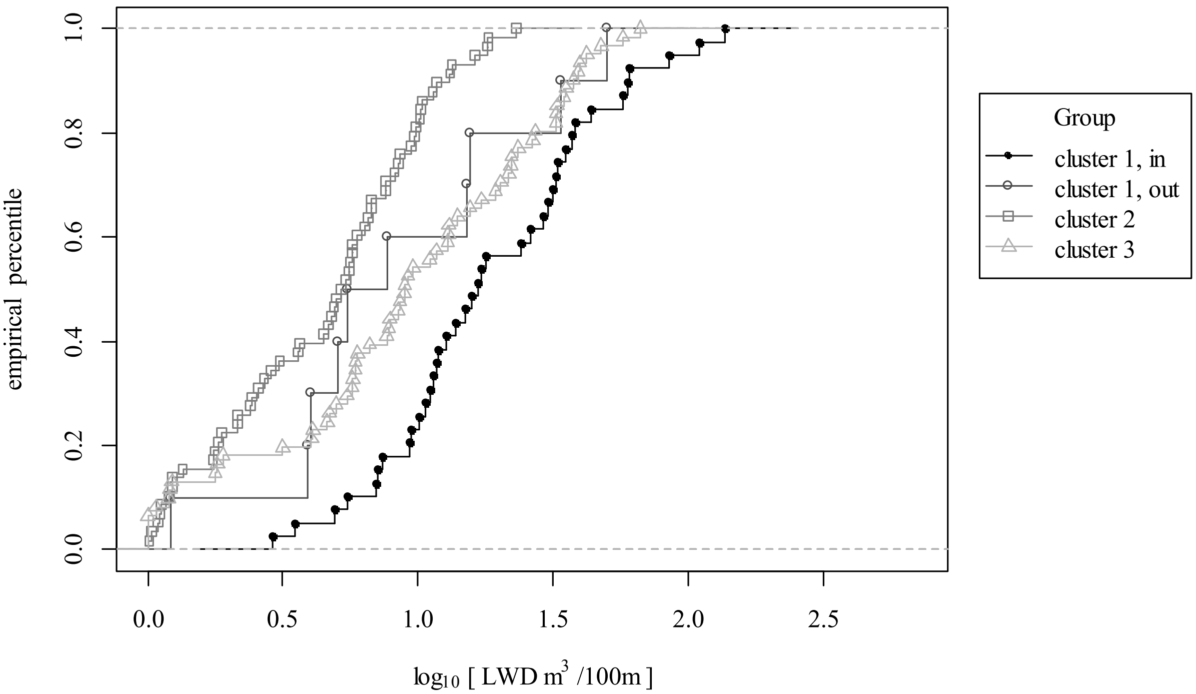
Empirical Cumulative Distribution for Unimpvd. Rd. Dens. (wtrshd)
Western Washington
LOG_Road.Density_watershed_km.km2_4

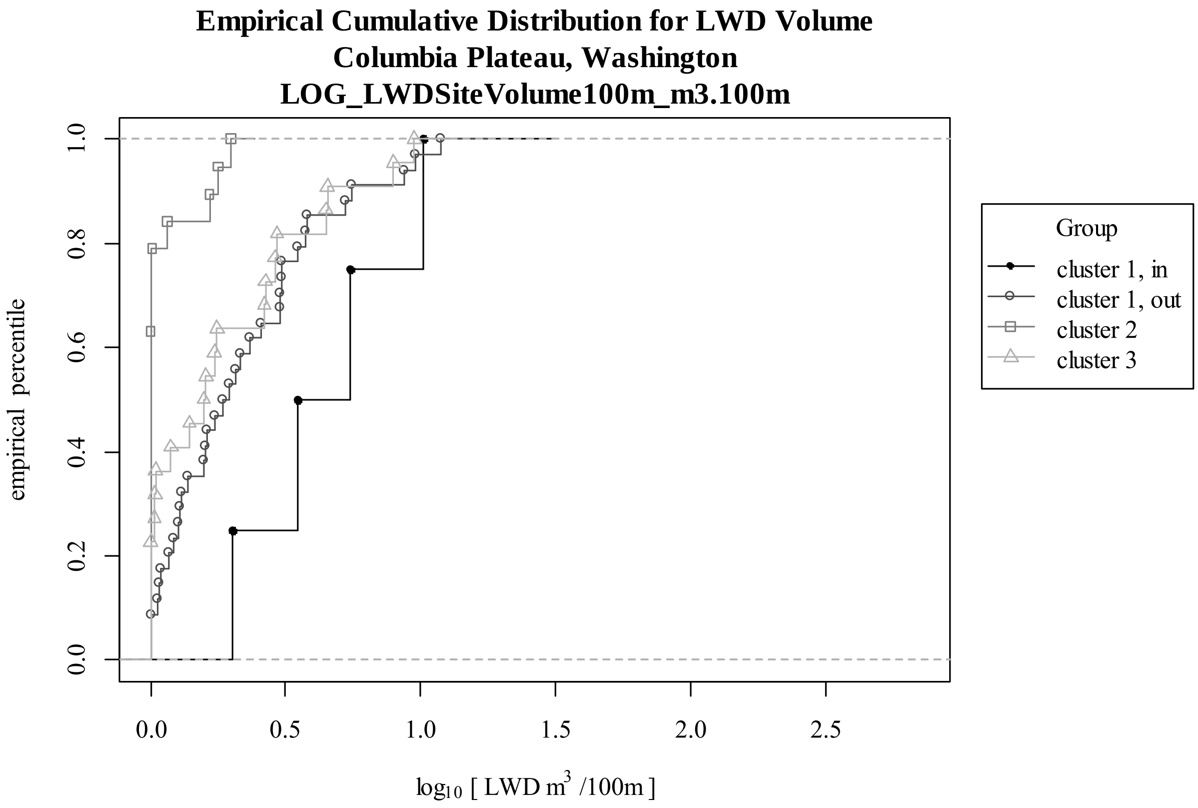


Empirical Cumulative Distribution for Unimpvd. Rd. Dens. (wtrshd)
Eastern Washington
LOG_Road.Density_watershed_km.km2_4

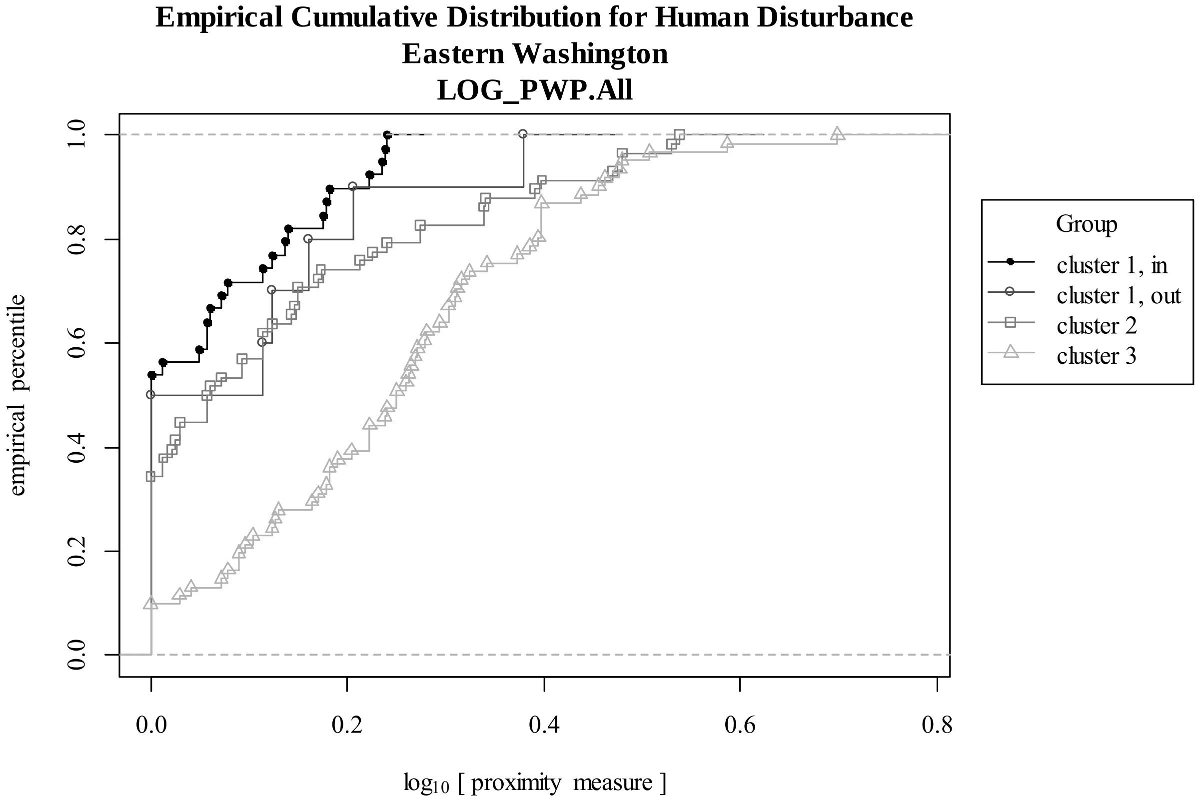



Empirical Cumulative Distribution for LWD Volume
Western Washington
LOG_LWDSiteVolume100m_m3.100m

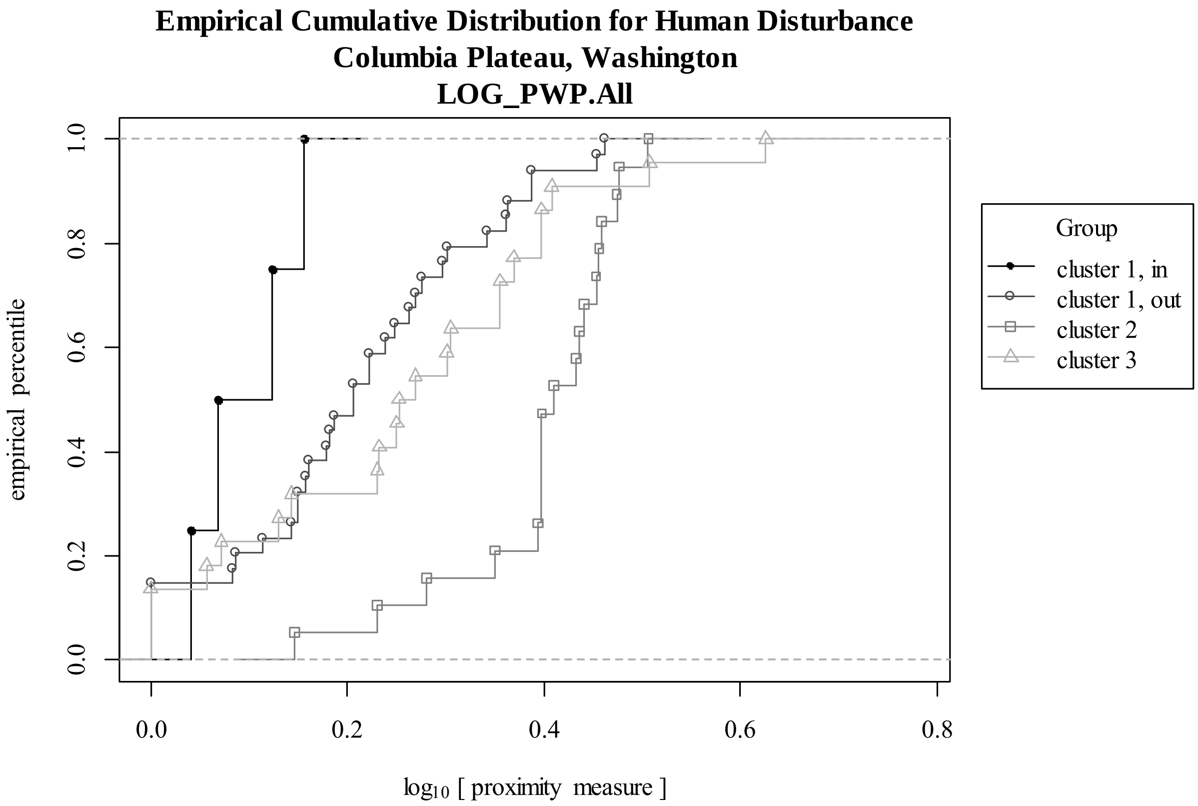

Empirical Cumulative Distribution for LWD Volume
Eastern Washington
LOG_LWDSiteVolume100m_m3.100m

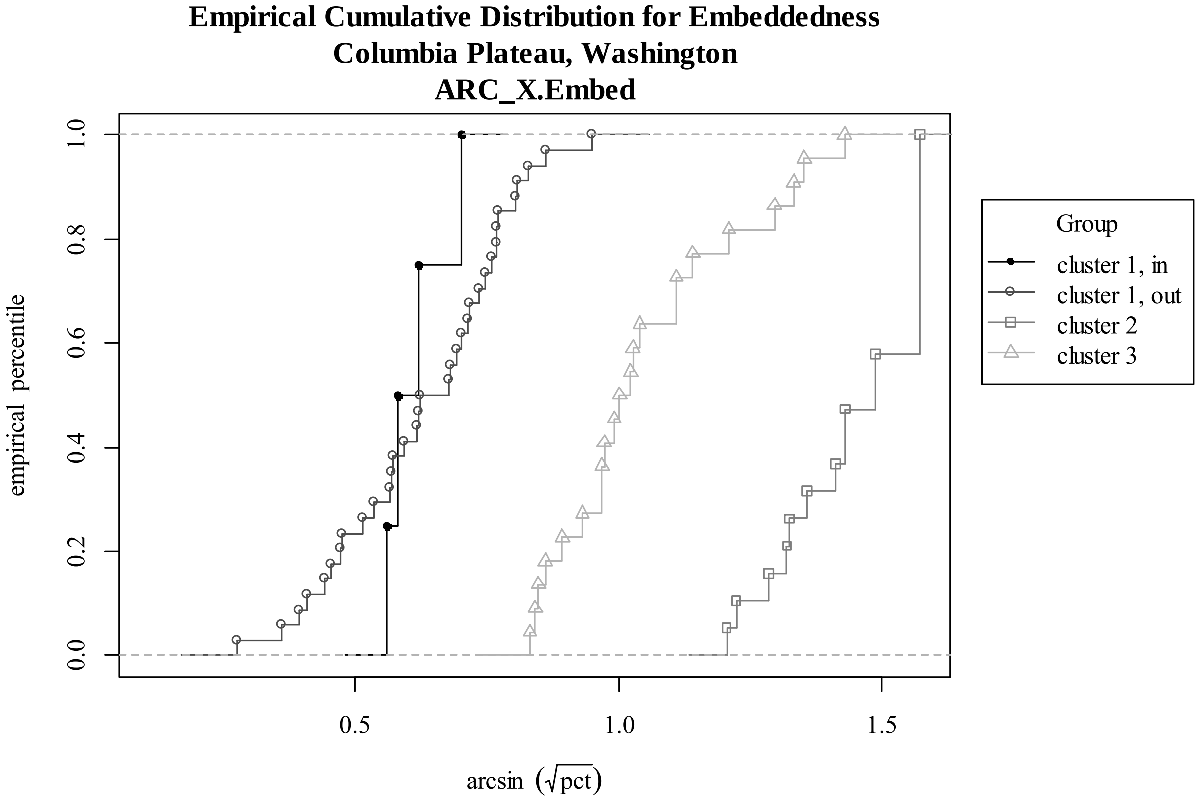


Empirical Cumulative Distribution for LWD Volume
Columbia Plateau, Washington
LOG_LWDSiteVolume100m_m3.100m



Empirical Cumulative Distribution for Human Disturbance
Eastern Washington
LOG_PWP.All


Empirical Cumulative Distribution for Human Disturbance
Columbia Plateau, Washington
LOG_PWP.All
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
$\log_{10}$ [proximity measure]
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3


Empirical Cumulative Distribution for Embeddedness
Columbia Plateau, Washington
ARC_X.Embed
empirical percentile
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
arcsin $(\sqrt{\text{pct}})$
Group
cluster 1, in
cluster 1, out
cluster 2
cluster 3

Appendix E. Percentiles of Least-Disturbed Sites as Thresholds

Table E-1. For Western Washington, quantiles for the group *cluster 1, in* for 9 GIS, habitat, and chemistry metrics. Quantiles represent values on the disturbance end of the continuum for each metric. For example, lower values of LWD occur at more disturbed sites, but higher values of bare ground cover (IDX.Ground) occur at more disturbed sites. For comparison, group means for BPJ reference sites and for all random sites are provided first. Means are on the transformed scale, where applicable.

Metric extent	**Metric**	**Western Washington**					
		BPJ ref.	**random**	**25%**	**10%**	**5%**	**1%**
		n=29	*n*=141				
Catchment	$\arcsin\sqrt{\%\ \text{evergreen}}$	1.07	0.88	1.02	0.91	0.86	0.85
Reach	D50Log10	1.57	0.91	1.50	1.08	0.86	0.09
Point	DO (mg/L)	10.27	9.67	9.83	9.67	9.61	9.21
Reach	$\log_{10}$[LWD vol.+1]	1.48	1.19	1.16	0.84	0.79	0.58
Reach	IDX.Ground	69.40	81.02	81.4	88.2	99.8	101.3
Point	Water Temp. $(^\circ C)$	11.75	14.32	12.20	13.02	13.42	13.84
Catchment	$\arcsin\sqrt{\%\ \text{shrub/scrub}}$	0.26	0.39	0.35	0.40	0.42	0.46
Catchment	$\log_{10}[\text{Total Rds.} + 1]$	0.31	0.55	0.29	0.41	0.42	0.43
Catchment	$\log_{10}[\text{Unimp. Rds.} + 1]$	0.29	0.51	0.28	0.40	0.42	0.43

Table E-2. For Eastern Washington, quantiles for the group *cluster 1, in* for 12 GIS, habitat, and chemistry metrics. Quantiles represent values on the disturbance end of the continuum for each metric. For example, lower values of LWD occur at more disturbed sites, but higher values of bare ground cover (IDX.Ground) occur at more disturbed sites. For comparison, group means for BPJ reference sites and for all random sites are provided first. Means are on the transformed scale, where applicable.

Metric extent	**Metric**	**Eastern Washington**					
		BPJ ref.	**random**	**25%**	**10%**	**5%**	**1%**
		n=42	*n*=133				
Catchment	$\arcsin\sqrt{\%\text{ canopy}}$	0.91	0.82	0.88	0.82	0.76	0.74
Catchment	$\arcsin\sqrt{\%\text{ evergreen}}$	1.12	1.01	1.01	0.92	0.91	0.87
Reach	D50Log10	1.37	0.94	1.15	0.73	0.45	0.34
Point	DO (mg/L)	10.08	9.36	9.60	9.25	8.90	8.82
Reach	$\log_{10}[\text{LWD vol.} + 1]$	1.17	0.87	1.02	0.83	0.68	0.49
Reach	IDX.Ground	51.48	68.29	73.75	79.88	84.09	94.02
Point	Water Temp. $(^\circ C)$	10.72	12.91	11.85	12.95	13.85	15.42
Catchment	$\arcsin\sqrt{\%\text{ shrub/scrub}}$	0.30	0.44	0.37	0.43	0.45	0.50
Catchment	$\arcsin\sqrt{\%\text{ sand/fines}}$	0.46	0.63	0.54	0.65	0.74	0.80
Catchment	$\log_{10}[\text{Total Rds.} + 1]$	0.24	0.42	0.23	0.30	0.39	0.43
Catchment	$\log_{10}[\text{Unimp. Rds.} + 1]$	0.20	0.40	0.20	0.25	0.30	0.42
Reach	$\log_{10}[\text{PWP. All} + 1]$	0.06	0.18	0.12	0.19	0.24	0.24

Table E-3. For the Columbia Plateau Ecoregion, quantiles for the group *cluster 1, in* for 7 GIS, habitat, and chemistry metrics. Quantiles represent values on the disturbance end of the continuum for each metric. For example, lower values of LWD occur at more disturbed sites, but higher water temperatures occur at more disturbed sites. For comparison, group means for BPJ reference sites and for all random sites (*n*=76) are provided first. Means are on the transformed scale, where applicable.

Metric extent	**Metric**	**Columbia Plateau**					
		BPJ ref.	**random**	**25%**	**10%**	**5%**	**1%**
		n=4	*n*=76				
Reach	D50Log10	1.44	0.26	1.38	1.28	1.25	1.23
Reach	$\log_{10}[\text{LWD vol.} + 1]$	0.65	0.25	0.48	0.37	0.34	0.31
Point	Water Temp. $(^\circ C)$	13.84	16.67	15.91	17.02	17.38	17.68
Catchment	$\arcsin\sqrt{\%\text{ shrub/scrub}}$	0.82	0.43	1.103	1.121	1.127	1.132
Catchment	$\arcsin\sqrt{\%\text{ sand/fines}}$	0.39	0.86	0.412	0.449	0.461	0.471
Reach	$\arcsin\sqrt{\%\text{ Embed}}$	0.62	0.96	0.643	0.679	0.691	0.701
Reach	$\log_{10}[\text{PWP.All} + 1]$	0.10	0.27	0.132	0.146	0.151	0.154

Appendix F. Least-Disturbed Random Sites

Site Name	Ecoregion
WAM06600_000497	Cascades
WAM06600_001405	Cascades
WAM06600_001422	Cascades
WAM06600_000510	Coast Range
WAM06600_000787	Coast Range
WAM06600_002878	Columbia Plateau
WAM06600-000091	Columbia Plateau
WAM06600-002485	Columbia Plateau
WAM06600-003730	Columbia Plateau
WAM06600_004334	Eastern Cascades Slopes and Foothills
WAM06600_000188	North Cascades
WAM06600_000384	North Cascades
WAM06600_000672	North Cascades
WAM06600_009724	North Cascades
WAM06600_018208	North Cascades
WAM06600_024284	North Cascades
WAM06600_004100	Northern Rockies
WAM06600_004240	Northern Rockies
WAM06600_009440	Northern Rockies
WAM06600_011760	Northern Rockies
WAM06600_018592	Northern Rockies
WAM06600_000831	Puget Lowland

Appendix G. Group Membership of BPJ Reference Sites

Western Washington		
Unique Site ID	**Ecoregion**	**Group**
BIO06600-HUCK04	Cascades	cluster 1, in
BIO06600-SUMM04	Cascades	cluster 1, in
BIO06600-RUSH04	Cascades	cluster 1, in
BIO06600-BEAR04	Cascades	cluster 1, in
BIO06600-UNIO04	Cascades	cluster 1, in
BIO06600-KALA04	Cascades	cluster 1, in
BIO06600-CAYA04	Cascades	cluster 1, in
BIO06600-BIGC04	Cascades	cluster 1, in
SEN06600-LAUG07	Cascades	cluster 1, in
SEN06600-TRAP08	Cascades	cluster 1, in
BIO06600-SLIT01	Coast Range	cluster 1, in
BIO06600-HYAS01	Coast Range	cluster 1, in
BIO06600-NSKO01	Coast Range	cluster 1, in
SEN06600-TWIN02	Coast Range	cluster 1, in
EPA06600-DUCK01	Puget Lowland	cluster 1, in
EPA06600-OYST01	Puget Lowland	cluster 1, in
SEN06600-HAMM03	Puget Lowland	cluster 1, in
BIO06600-CHRI01	Coast Range	cluster2
BIO06600-DAVI01	Coast Range	cluster2
BIO06600-BEAR01	Coast Range	cluster2
SEN06600-ELLS01	Coast Range	cluster2
EPA06600-BATT01	Puget Lowland	cluster2
EPA06600-DEWA01	Puget Lowland	cluster2
BIO06600-AUST02	Puget Lowland	cluster3
BIO06600-SEAB02	Puget Lowland	cluster3
EPA06600-TULA01	Puget Lowland	cluster3
SEN06600-GRIF09	Puget Lowland	cluster3

Eastern Washington		
Unique Site ID	**Ecoregion**	**Group**
BIO06600-TUCA11	Blue Mountains	cluster 1, in
BIO06600-SPAN11	Blue Mountains	cluster 1, in
SEN06600-PANJ11	Blue Mountains	cluster 1, in
BIO06600-NTAN09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-STAN09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-FIRS09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-MORR09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-BIRD09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-WILS09	Eastern Cascades Slopes and Foothills	cluster 1, in
BIO06600-DIOB77	North Cascades	cluster 1, in
SEN06600-TEAN04	North Cascades	cluster 1, in
BIO06600-BLAC77	North Cascades	cluster 1, in
BIO06600-STRA77	North Cascades	cluster 1, in
BIO06600-QUAR77	North Cascades	cluster 1, in
BIO06600-SKUL77	North Cascades	cluster 1, in
BIO06600-WARC77	North Cascades	cluster 1, in
SEN06600-CLEE12	North Cascades	cluster 1, in
SEN06600-MARB21	North Cascades	cluster 1, in
SEN06600-TWEN05	North Cascades	cluster 1, in
BIO06600-NROC01	Northern Rockies	cluster 1, in
BIO06600-BURP15	Northern Rockies	cluster 1, in
BIO06600-PRIE15	Northern Rockies	cluster 1, in
BIO06600-LOST15	Northern Rockies	cluster 1, in
BIO06600-MFLU15	Northern Rockies	cluster 1, in
BIO06600-NDEA15	Northern Rockies	cluster 1, in
SEN06600-DEAD19	Northern Rockies	cluster 1, in
SEN06600-SULL20	Northern Rockies	cluster 1, in
SEN06600-CUMM10	Blue Mountains	cluster2
BIO06600-SASO10	Blue Mountains	cluster2
BIO06600-COUG11	Blue Mountains	cluster2
BIO06600-MENA11	Blue Mountains	cluster2
BIO06600-JOSE11	Blue Mountains	cluster2
BIO06600-ELKC09	Eastern Cascades Slopes and Foothills	cluster2
BIO06600-ROCK09	Eastern Cascades Slopes and Foothills	cluster2
BIO06600-SHAD09	Eastern Cascades Slopes and Foothills	cluster2
BIO06600-CHAR11	Blue Mountains	cluster3

Eastern Washington		
Unique Site ID	**Ecoregion**	**Group**
BIO06600-BACO09	Eastern Cascades Slopes and Foothills	cluster3
SEN06600-POOR22	North Cascades	cluster3
BIO06600-NGOO15	Northern Rockies	cluster3
BIO06600-MMIL15	Northern Rockies	cluster3
BIO06600-TONA15	Northern Rockies	cluster3

Columbia Plateau Ecoregion, Washington		
Unique Site ID	**Ecoregion**	**Group**
BIO06600-OAKC10	Columbia Plateau	cluster 1, in
BIO06600-UMTA10	Columbia Plateau	cluster 1, in
SEN06600-ASOT13	Columbia Plateau	cluster 1, in
SEN06600-UMTA18	Columbia Plateau	cluster 1, in