United States Department of Agriculture

Forest Service

Rocky Mountain Research Station

General Technical Report RMRS-GTR-138

December 2004

Repeatability of Riparian Vegetation Sampling Methods: How Useful Are These Techniques for Broad-Scale, Long-Term Monitoring?

Marc C. Coles-Ritchie
Richard C. Henderson
Eric K. Archer
Caroline Kennedy
Jeffrey L. Kershner

Abstract

Coles-Ritchie, Marc C.; Henderson, Richard C.; Archer, Eric K.; Kennedy, Caroline; Kershner, Jeffrey L. 2004. **Repeatability of riparian vegetation sampling methods: how useful are these techniques for broad-scale, long-term monitoring?** Gen. Tech. Rep. RMRS-GTR-138. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 18 p.

Tests were conducted to evaluate variability among observers for riparian vegetation data collection methods and data reduction techniques. The methods are used as part of a large-scale monitoring program designed to detect changes in riparian resource conditions on Federal lands. Methods were evaluated using agreement matrices, the Bray-Curtis dissimilarity metric, the coefficient of variation, the percentage of total variability attributed to observers, and estimates of the number of sites needed to detect change.

Community type (CT) cover data differed substantially among the six to seven observers that sampled the same sites. The mean within-site similarity in the vegetation data ranged from 40 to 65 percent. Converting CT data to ratings (bank stability, successional, and wetlands ratings) resulted in better repeatability, with coefficients of variation ranging from 6 to 13 percent and a percentage of variability attributed to observers of 16 to 44 percent. Sample size estimates for the ratings generated from CT cover data ranged from 56 to 224 sites to detect a change of 10 percent between two populations. The woody species regeneration method was imprecise. The effective ground cover method was quite precise with a coefficient of variation of two, but had so little variability among sites that statistically significant change in this attribute would not be expected. In general, reducing the CTs to ratings increased precision because of the elimination of differences among observers that were not important from the perspective of the rating.

Studies that seek to detect change at a single site would need to take into account the observer variability described here. Studies that seek to detect differences between populations of sites could detect relatively large changes with these methods and ratings. Small differences among populations would be difficult to detect with a high degree of confidence, unless hundreds of sites were sampled.

Key words: riparian, vegetation, ecology, monitoring, observer variability, community types, repeat sampling.

Acknowledgments

We express appreciation to Alma Winward, former Regional Ecologist for the USDA Forest Service, for his pioneering work in riparian vegetation monitoring. We acknowledge assistance and draft reviews from Forest Service colleagues including Wayne Padgett, Mary Manning, Elizabeth Crowe, Dave Weixelman, Tim Burton, Bud Kovalchik, Desi Zamudio, Brett Roper, Dave Turner, and Dale Bartos. We appreciate comments from other reviewers including Peter Alpert, Rex Crawford, and Dave Roberts. We acknowledge the many field technicians who have collected data and provided insights about the methods of data collection.

Cover Photo: The riparian area at the Big Creek measurement site on the Nez Perce National Forest. The measuring tape begins at the greenline and extends 10 m along the vegetation cross-section, which is where the measurement data were collected in 1-m increments.

Repeatability of Riparian Vegetation Sampling Methods: How Useful Are These Techniques for Broad-Scale, Long-Term Monitoring?

Contents

The Authors

Marc C. Coles-Ritchie is a Riparian Vegetation Ecologist for the U.S. Department of Agriculture, Forest Service, Fish and Aquatic Ecology Unit in Logan, UT. He received a B.A. degree in comparative literature from the University of Massachusetts and an M.S. degree in environmental studies from Bard College, which included research on the harvest of tropical forest materials in Ecuador. Marc is completing a Ph.D. degree in vegetation ecology at Utah State University, Logan.

Richard C. Henderson is a Fisheries Biologist for the U.S. Department of Agriculture, Forest Service, Fish and Aquatic Ecology Unit in Logan, UT. He earned his B.S. degree in fisheries biology from Colorado State University in Fort Collins in 1992. In 1998 he received his M.S. degree in fisheries biology from Utah State University, Logan. Rick joined the Forest Service in 1998 as the Project Manager for the PACFISH/INFISH Effectiveness Monitoring Project.

Eric K. Archer is a Fishery Biologist with the Fish and Aquatic Ecology Unit in Logan, UT. Eric received a B.S. degree in Fisheries Management and an M.S. degree in Aquatic Ecology in 1999 from Utah State University, Logan. Eric joined the Forest Service as a Fisheries Biologist in 1999 and has worked as a Supervisor of Field Operations and Data Analyst for the PACFISH/INFISH Effectiveness Monitoring Project.

Caroline Kennedy is an M.S. degree candidate in range science at Utah State University, Logan. She earned bachelors' degrees in range management and technical journalism at Colorado State University. Her professional interests include riparian ecology and the social implications of natural resource-based economies.

Jeffrey L. Kershner is a Washington Office Aquatic Ecologist with the U.S. Department of Agriculture, Forest Service, Fish and Aquatic Ecology Unit, in Logan UT. He completed a B.S. degree in 1976 at Humbolt State University, Humboldt, CA, in fisheries. He earned an M.S. degree in natural resources from Humboldt State University in 1982 and a Ph.D. degree in ecology from the University of California, Davis, in 1991.

Repeatability of Riparian Vegetation Sampling Methods: How Useful Are These Techniques for Broad-Scale, Long-Term Monitoring?

Marc C. Coles-Ritchie
Richard C. Henderson
Eric K. Archer
Caroline Kennedy
Jeffrey L. Kershner

Introduction

Legal requirements for protecting species and ecosystems necessitate monitoring efforts that can determine the level of anthropogenic influence on ecosystems across space and through time. Researchers and land managers often need to detect the effects of human influence such as livestock grazing, logging, and roads on the riparian ecosystem. In order for actual differences to be detected, methods must be objective and repeatable by different observers at the same locations, and over different time periods.

Observer variability is often overlooked in studies of vegetation, or it is assumed to be zero as noted by Elzinga and others (1998) and Gotfryd and Hansell (1985). This assumption is dangerous because differences due to observers can arise from a number of sources, including: methods that allow observer subjectivity to affect decisions, methods that do not permit consistent measurement, recording errors, and incorrect species identification (Elzinga and others 1998). Differences due to observers can be minimized with proper protocol development and training of observers. However, some observer difference is inherent in all sampling methods, so quality assurance testing must be conducted to determine how much observer variability exists. Understanding the degree of difference due to observer allows researchers to know the level of change detection that is possible.

Earlier researchers recognized observers as a source of variability in vegetation sampling data (Greig-Smith 1957; Hope-Simpson 1940). Imprecision due to observer difference has been documented for: frequency of species observance (Hope-Simpson 1940; Kirby and others 1986; Leps and Hadincova 1992; Nilsson and Nilsson 1985); species cover estimates (Gotfryd and Hansell 1985; Kennedy and Addison 1987; Leps and Hadincova 1992; Sykes and others 1983); and mapping of cover types (Cherrill and Mcclean 1999). Studies have also found that the same observer varied when repeated sampling was done at a site (Kennedy and Addison 1987; Smith 1944; Sykes and others 1983). As would be expected, within-observer variability (the same observer doing repeat sampling) has been found to be lower than between-observer (more than one observer sampling the same site) variability (Kirby and others 1986; Smith 1944; Sykes and others 1983).

Studies have also tested the repeatability of riparian vegetation sampling methods, although there are fewer studies than for upland vegetation (Elzinga and Evenden 1997). Some studies of repeatability in riparian vegetation methods were part of stream habitat studies that measured general categories of vegetation. One such study found "fairly high" variation among observers (95 percent confidence interval one to six times the field measurement precision) for a bank vegetation metric that categorized vegetation into broad categories (woodland, shrub, meadow, residential) (Wang and others 1996). Another study found coefficients of variation (CV) over 33 percent for estimation of general riparian vegetation categories in Oregon, such as amount of canopy, mid-layer, and ground-layer cover (Kaufmann and others 1999). A method of characterizing vegetation was relatively precise for the categories of herbaceous plants and shrubs (CV of 8 percent) but was less precise for a tree category (CV of 25 percent) (Barker and others 2002). The use of such broad categories may have created difficulty distinguishing the categories, especially in transition zones, of which there are many in riparian areas.

Observer variability has also been evaluated for methods that evaluate ungulate browsing on woody species. A study of twig length measurement found that 20 percent of the total variation was due to the observer, leading the researchers to "seriously question the feasibility of measuring riparian shrub utilization" (Hall and Max 1999). A study of precision in estimating willow canopy volume found an average CV of 25 percent, indicating large differences among observers, although variability was lower for a single observer (Thorne and others 2002).

The CV as an estimate of precision can be misleading without considering total variability. When determining the usefulness of a particular variable it is important to consider the level of observer variability relative to total variability, in addition to the CV, as is done in this study.

Few studies have incorporated the implications of observer variability into their study design (Gotfryd and Hansell 1985; Sykes and others 1983). Understanding the level of observer difference is essential if researchers are to make appropriate assessments of change over time or across space. Without data on observer-based variability, incorrect conclusions can be made about changes in vegetation and the role of management. Spurious results can include detection of changes that did not occur but were an artifact of observer difference (type I errors) and the failure to detect changes that did occur (type II errors). To minimize these errors, quality assurance testing must be done to identify and develop highly repeatable methods that generate monitoring data that will effectively detect change.

Study Question

The concern over the status of anadromous and resident salmonids has prompted greater interest in the condition of stream and riparian habitat in the upper Columbia River Basin. This has underscored the need for monitoring that can document the condition of riparian areas and changes over time relative to management. Some data exist for riparian vegetation in the Basin, but they come from a variety of sources and methods, making it difficult to combine information and evaluate differences. Even when data have been collected in a consistent manner, the lack of quality assurance testing inhibits our ability to use such data to detect differences.

In response to this need, the U.S. Department of Agriculture, Forest Service (USFS) and the U.S. Department of the Interior, Bureau of Land Management (BLM) designed a large-scale monitoring program in 1998, the PACFISH/INFISH Biological Opinion Effectiveness Monitoring Program for Streams and Riparian Areas of the Upper Columbia River Basin (PIBO-EMP). The objective of the PIBO-EMP is to determine whether stream habitat and riparian condition on Federal lands is changing over time (Kershner and others 2004). The variables studied include a wide range of stream channel, riparian vegetation, aquatic biota, and watershed descriptors.

The riparian vegetation component of the PIBO-EMP adopted the riparian vegetation monitoring methods described by Winward (2000), which use community types to evaluate the vegetation at a site and estimate woody plant density. A method of estimating effective ground cover was also included in the protocol. These methods were chosen because: (1) they generate quantitative site values; and (2) they are currently used within land management agencies. Monitoring with community types allows rapid assessment of vegetation, although there is subjectivity in determining the appropriate community type on the ground. This subjectivity, and the inherent variability of observers in vegetation sampling, supported the need for quality assurance testing.

A pilot quality assurance test was conducted during the 1999 sampling season. Results were used to refine the PIBO-EMP protocols (see protocol in Kershner and others 2004). In 2000 and 2001 a more formal quality assurance program was undertaken to define the variability associated with the evaluation of stream (Archer and others 2004; Roper and others 2002) and riparian habitat.

This paper presents the riparian component of the quality assurance program, which was designed to quantify the variability in the measurement and summarization of attributes that describe riparian habitat. The sources of variability were quantified as: (1) differences among observers at specific points (small scale); (2) differences among observers for a site (larger scale); and (3) differences among sites. Quantification of these sources of variability allowed us to determine the sample sizes needed to detect differences between populations of sites (such as sites grazed by livestock versus ungrazed sites) for each method and summary technique. This information allows an evaluation of the usefulness of each method, and whether it should be retained, modified, or eliminated.

Study Area

The quality assurance study was conducted in central Idaho on lands managed by the Nez Perce and Payette National Forests (fig. 1). Data were collected at individual sites along six streams: Boulder and Little Goose Creeks within the Little Salmon River drainage; Lost Creek within the Weiser River drainage; and Big, Jack, and Meadow Creeks within the South Fork Clearwater River drainage (table 1). These stream sites are numbered 1 through 6 respectively in all figures.

An additional 44 randomly selected sites within the two forests were used to estimate site variability. All sites were associated with 2nd to 4th order streams, with stream gradients from 0.3 to 2.5 percent, and bankfull widths from 1.0 to 12.3 m. Riparian areas varied from narrow to wide valley bottoms, and from open meadows to forests. The quality assurance sites and the randomly selected sites represent a range of stream sizes, gradients, elevations, geology, and valley types that are sampled by the PIBO-EMP within these two National Forests.

Methods

The vegetation sampling methods of Winward (2000) and a Region 4 soils protocol (USDA 1989) were tested in this quality assurance study. The Winward (2000) protocol was designed to evaluate the condition of riparian vegetation with the following methods: greenline vegetative composition, vegetation cross-section composition, and woody species regeneration. The effective ground cover method of the Region 4 soils protocol (USDA 1989) was also tested. A brief description of the data collection methods and the different summary techniques is presented in the section entitled “Description, Results, and Discussion for each Method.” The complete protocol can be found in Kershner and others (2004).

Figure 1—Map of the study area of the PACFISH/INFISH Biological Opinion Effectiveness Monitoring Program for Streams and Riparian Areas of the Upper Columbia River Basin, and the locations of the quality assurance test sites in Idaho. "N" represents the three sites on the Nez Perce National Forest, and "P" represents the three sites on the Payette National Forest.

The greenline and vegetation cross-section methods require the use of riparian vegetation classifications. The riparian vegetation classification by Padgett and others (1989) was used for the quality assurance study on the Payette and Nez Perce National Forests and for many of the other 44 sites. The classification by Hansen and others (1995) was used for 12 of the sites located on the Nez Perce National Forest. The classifications describe 79 and 113 riparian community types, respectively, and assist observers in identifying those types in the field. In this report the term community type (CT) is used in a broad sense—regardless of successional status—to include plant community types, plant associations, and habitat types.

Vegetation data collectors (observers) received 12 days of training at the start of the field season, which is extensive for a seasonal field crew. The training included 8 days dedicated to learning the dominant riparian species, how to use CT classifications, and the sampling methods. The other 4 days were spent working in pairs to practice the methods at several training sites. The observers then worked independently in their respective geographical area (Idaho, Oregon, and Montana) before the quality assurance sampling.

For the quality assurance tests, three components of variability associated with the evaluation of riparian vegetation were assessed: (1) variability in application of a method at a marked point (measurement data); (2) variability in data summarized from an entire site (repeat data); and (3) variability among sites (site variability).

The "measurement" study was conducted in early August 2001, with seven observers collecting data at each of the six stream sites. This measurement study was designed to detect the causes of variability in each method at a small scale. We tested the variability associated with data collection for each method by having different observers collect data, or perform measurements, at marked points. Measurement data were collected for the four methods, described below. Instruction packets were distributed to observers before the study explaining the measurement data collection procedures. For greenline and vegetation cross-section data collection, observers recorded the CT for each meter along a measuring tape that was staked to the ground. This allowed an estimation of variability in using CTs to describe vegetation. For the woody species regeneration method, seven woody plants were flagged and numbered at the six streams, and the observers recorded the species, age-class, and height category. A mixture of woody species at that stream site was included. A random selection of plants was not attempted because the objective was to compare identification and age-classing of a number of woody species, especially willows. Rhizomatous species were included to test their knowledge of these species as well.

The three "repeat" sites on the Nez Perce National Forest were sampled in late June 2000 and August 2001. The three sites on the Payette National Forest were sampled in early

Table 1—General information for stream sites used in the measurement and repeat studies of riparian vegetation sampling methods. The sites are on the Payette and Nez Perce National Forests in central Idaho.

Site code	Stream	National Forest	Geology	Elevation	Gradient	Bankfull width
				feet	*Percent*	*m*
1	Boulder	Payette	Volcanic	4,750	0.34	7.7
2	Little Goose	Payette	Volcanic	5,000	1.57	3.5
3	Lost	Payette	Volcanic	4,855	0.59	5.4
4	Big	Nez Perce	Granitic	6,360	0.33	4.4
5	Jack	Nez Perce	Granitic	5,280	0.73	2.7
6	Meadow	Nez Perce	Granitic	3,200	0.37	8.3

August both years. Each of the six repeat sites was sampled by six observers in both years, except for two sites in 2000 that were sampled by seven observers. Observers were randomly assigned to a site each day. The observers sampled the same site based only on a fixed starting point. Each observer established the boundaries of the site and performed all methods using normal data collection procedures. These repeat sites were located adjacent to the area where measurement data were collected. The repeat study was designed to test not only measurement differences but also variability due to interpretation and application of sampling methods for an entire site.

The additional 44 sites were used to better characterize the "site variability" on the two forests. A random effects model (described below) was used to calculate the variance among all 50 sites, six of which were the repeat sites. Of the 50 sites, the six quality assurance repeat sites were sampled 12 to 13 times by 12 observers over 2 years, six sites were sampled by two observers in different years, and the remaining 38 were sampled once.

Wetland Rating System

An initial concern with these methods was how to evaluate the CT data in a way that would allow comparisons among observers and among sites. Winward (2000) describes two rating systems for greenline CT data, the greenline stability and greenline successional ratings, but no rating for vegetation cross-section CT data. In addition, the greenline ratings only apply to three of the eight vegetation classifications (only those in Region 4 of the USDA Forest Service) used within the PIBO-EMP study area. To have a uniform and quantitative rating system for our entire study area, we developed a "wetland rating" for the CT data collected along the greenline and vegetation cross-sections. Wetland ratings (between 0 and 100) were computed for the 725 vegetation types in eight classifications using data on the average species cover and constancy (percentage of plots with the species) from classifications, and the species wetland indicator status (Reed 1996).

The wetland rating is an index that quantifies the abundance of vegetation in relation to the wetland indicator status. A CT or site with a value close to 100 would indicate that the species were primarily obligate wetland species. A value close to 0 would indicate that the CT or site had mainly upland species, and no obligate wetland species. A high degree of moisture availability, and hence high wetland ratings, would be expected for the low-gradient reaches sampled in this study. The wetland rating should be considered in the context of the environmental conditions of each site because the environment greatly affects moisture availability. We are currently evaluating the usefulness of the wetland rating to describe the condition of riparian areas (Coles-Ritchie, in preparation).

The wetland rating provides indirect information about bank stability because plants that persist in the wettest part of the riparian environment (in other words, next to the stream) must withstand high shear stress from flowing water (Auble and others 1994; Bendix and Hupp 2000). As a result, a correlation between obligate wetland plants and strong rooting characteristics is expected in the riparian environment (Winward 2000). Communities dominated by obligate wetland species, such as *Carex nebrascensis* and *Juncus balticus*, often have greater very-fine root-length density, greater biomass, and a deeper distribution of roots in the soil profile than communities dominated by upland grasses (Dunaway and others 1994; Kleinfelder and others 1992; Manning and others 1989; Toledo and Kauffman 2001). The dense and deep root structure of these and many other obligate wetland species help them to maintain their position and decrease streambank erosion (Dunaway and others 1994). Therefore, the wetland rating may partially represent the bank stabilizing capacity of vegetation in riparian areas.

Data Analysis

A variety of graphical and statistical techniques were used to summarize and analyze the vegetation data. Data were initially examined for the presence of obvious data entry problems, which were minimal, and they were corrected. All variables used in the statistical models were normally distributed. The woody species regeneration data were not normally distributed, and therefore they were not analyzed with the random effects model.

Measurement Data—Descriptive statistics were computed from individual measurements of each attribute and then averaged by method, site, and observer. The mean, standard deviation (SD), and CV were computed for continuous data.

For the greenline and cross-section CT data, an agreement matrix was used to calculate the average between-observer agreement for all observer pairs. For each 1-m unit, a pair of observers recorded either the same CT or a different CT. The percentage of meters where pairs of observers agreed is the average between-observer agreement.

A fuzzy agreement matrix was also calculated by using the similarity of the two CTs rather than "agreement." At 1-m units where a pair of observers recorded the same CT, the similarity was 100 percent. At 1-m units where observers recorded different CTs, the agreement was not zero, but rather the similarity (from 0 to 100 percent) of the two CTs that each recorded. Those similarities were calculated for all pairs of CTs with a Bray-Curtis dissimilarity matrix. For example, the *Agrostis stolonifera* CT and the *Poa pratensis* CT (from Hansen and others 1995) are 27 percent similar. When two observers used those two CTs to describe the same 1-m area, the fuzzy agreement was 27 percent rather than 0. An average fuzzy agreement was calculated for all pairs of observers, based on the similarity of the CTs they recorded, for all the 1-m units.

Greenline and vegetation cross-section CT data were also converted to numeric values using the stability, successional, and wetland ratings. The SD and CV were used to compare these ratings.

Woody species regeneration data were summarized as the percentage of observers that agreed on the genus, species, age-class, and height of each willow plant.

Repeat Data: Measure of Vegetation Similarity—CT data were compared among observers and sites to determine observer agreement. The percent cover of CTs recorded by an observer at each site was used to generate a dissimilarity matrix, which was used to calculate within-group similarities and to generate ordinations. The 2 years were analyzed separately, even though the sites were the same, because annual variability (the year effect) could not be eliminated as it can be in the random effects model (discussed below). This eliminated any confounding sources of variability, such as differences in observers, training, or vegetation between years.

Data from six to seven observers at the same site were compared with multiresponse permutation procedures (MRPP) in PC-ORD (McCune and Mefford 1999). MRPP is a tool "to detect concentration within a priori groups" (Zimmerman and others 1985). MRPP calculated the mean within-group similarity to indicate the agreement among observers at the same site.

Within-group dissimilarities were calculated with the Bray-Curtis dissimilarity metric. A p-value (P) indicated the probability of obtaining the observed weighted mean within-group dissimilarity, relative to the distribution of possible values (McCune and Grace 2002). With this repeat data, it was expected that the p-values would be significant, as they were in each case. That was not informative, because the data were from the same sites and therefore should be more similar than expected by chance. Therefore p-values were not reported.

More informative was the level of similarity of data among observers. This was represented by the agreement statistic (A), which is the "chance-corrected within-group agreement" (McCune and Mefford 1999). The A statistic would be 1 if all items in the group were identical, 0 if the items in the group had the same heterogeneity that would be expected by chance, or negative if there were more heterogeneity than expected by chance (McCune and Grace 2002). The A statistic provides an indication of observer agreement, and hence the repeatability of the method.

The similarity of CT data for different observers at the same site was represented in ordination diagrams. Ordination is a method of arranging sites based on their similarity of vegetation and/or their environmental conditions (Kent and Coker 1992). The ordination method used was nonmetric multidimensional scaling (NMDS), which Minchin (1987) and Clymo (1980) found to be the most robust of the ordination techniques that they evaluated. McCune and Grace (2002) concluded that NMDS was "the most generally effective ordination method for ecological community data and should be the method of choice, unless a specific analytical goal demands another method."

We used a Bray-Curtis dissimilarity metric for the NMDS ordinations as did Minchin (1987) where he found that NMDS to be robust. McCune and Mefford (1999) recommend the Bray-Curtis dissimilarity metric.

Preliminary ordinations were calculated to determine the appropriate number of dimensions and the best starting configuration, and to perform a Monte Carlo test with randomized runs. The Monte Carlo test was done to evaluate the probability that a similar final stress could be achieved by chance. "Stress" represents the degree to which the dissimilarity in the ordination (with the number of dimensions selected) differs from the dissimilarity in the original dissimilarity matrix (McCune and Mefford 1999). For preliminary runs, a randomized starting configuration was used, the maximum number of iterations was 400, the instability criterion was 0.00001, the starting number of dimensions was six, the number of runs with real data was 40, and the number of runs with randomized data was 50 (table 2). "Instability is calculated as the standard deviation in stress over the preceding x iterations, where x is set by the user" (McCune and Mefford 1999).

For each final ordination, the number of dimensions (axes) was selected according to the following criteria: an additional dimension was added if it reduced the stress by five or more, and if that stress was lower than 95 percent of the randomized runs (a $p <= 0.05$ for the Monte Carlo test) as recommended by McCune and Mefford (1999).

The final run was done with the best starting configuration for the number of dimensions selected and no step-down in dimensionality (table 2). For final runs the following information was reported: the number of dimensions that were statistically different from 95 percent of the randomized runs (Monte Carlo test results), number of iterations, final instability, final stress, and the cumulative R^2 based on the correlation coefficient between the ordination distances and distances in the original n-dimensional space.

Repeat Data: Random Effects Model—The CT data were also converted to ratings; a wetland rating (described above), and stability and seral ratings (described below). A random effects model run with PROC MIXED (Littell and others 1996; SAS 2000) was used to estimate the mean and variance of each rating associated with observer and site.

$$Y_{site,year,observer} = \mu + \xi_{site} = \eta_{year} = \xi\eta_{site,year} + \gamma_{observer(year)} + \varepsilon_{site,year,observer}$$

Table 2—The settings used to perform the NMDS ordinations, using the Bray-Curtis dissimilarity index. The preliminary run included runs with randomized data for the Monte Carlo test to determine the appropriate number of dimensions (axes).

Run	Starting configuration	Maximum iterations	Instability criterion	Starting number of axes	Runs with real data	Runs with randomized data
Preliminary	Random number	400	0.00001	6	40	50
Final	Best from runs with real data	200	0.0001	1-6	1	none

where:

- $Y_{site,year,observer}$ is the value of a response variable for a given site, year, and observer
- μ is the overall mean of the variable
- ξ_{site} is a random effect for the **site** with mean zero and variance σ^2_{site}
- η_{year} is a random effect for the **year**, a repeated measurement with mean zero and variance σ^2_{year}
- $\xi\eta_{site,year}$ is a random effect for the interaction of **site with year** with mean zero and variance $\sigma^2_{site,year}$
- $\gamma_{observer(year)}$ is a random effect for the **observer** with mean zero and variance $\sigma^2_{observer(year)}$
- $\varepsilon_{site,observer,year}$ is the underlying **residual** with mean zero and variance σ^2

The objective of this study was to assess the precision with which riparian vegetation attributes could be measured within a sampling season where habitat condition is assumed to be constant. All five variance components were estimated from the combined 2000 and 2001 data, but only the estimates for site (σ^2_{site}), observer within site ($\sigma^2_{observer(year)}$), and residual ($\sigma^2$) estimates were used to evaluate observer variability. Data from both years were included in the model to incorporate the observer variability inherent in long-term monitoring programs, which includes differences in training and unknown variability among observers across years. However, the year effect (σ^2_{year}), if there was a consistent difference between years, was not included in the analysis of observer variability.

We defined observer variability as the sum of the observer and residual terms, which includes all error associated with observer within year ($\gamma_{observer(year)}$) and all unexplained error ($\varepsilon_{site,observer,year}$). Site variability was based on the term ξ_{site} that describes the variation among sites that is not attributed to observer or interannual variability (year effect). We evaluated observer precision by calculating the SD and CV for each variable. In addition, we calculated the intraclass correlation coefficient as the magnitude of the observer variability relative to the overall variability:

$$\sigma^2_{observer(year)} / (\sigma^2_{site} + \sigma^2_{observer(year)} + \sigma^2).$$

Calculation of Minimum Sample Sizes for Observer and Total Variation—Sample size estimates are a useful tool to evaluate monitoring methods because they indicate the amount of effort necessary to be confident differences in an attribute will be detected (Eckblad 1991). Minimum sample sizes were calculated using specified differences between the means of two groups (for example, grazed versus ungrazed sites). For this study, observer and site variability (based on estimated variance components described above) were both included, in estimating sample sizes. We evaluated differences between means that ranged from 5 to 50 percent. This range was chosen because differences of these magnitudes likely included changes in attributes that would result in a biological response. We limited our evaluation to a type I error rate of = 0.1, and a type II error rate of = 0.1.

Estimates of sample size were calculated following the iterative procedure outlined by Zar (1996, page 133, equation 8.22):

$$n \geq \frac{2S_p^2}{d^2}\left(t_{\alpha(2),v} + t_{\beta(1),v}\right)^2$$

where:

$S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$ is the pooled estimate of variance

v = $(n_1 - 1) + (n_2 - 1)$ is the degrees of freedom for S_p^2

$t_{\alpha(2),v}$ = 2-tailed t-value on v df for a type I error rate of α (also used for 1-α, two-sided confidence intervals)

$t_{\beta(1),v}$ = 1-tallied (upper) t-value on v df where β is the acceptable type-II error rate

d = minimum difference to be detected.

For these calculations, we used variance estimates from the repeat study as estimates of S_p^2. Total variance was calculated as the sum of site and observer variability (Clark and others 1996; Montgomery 1984; Ramsey and others 1992). This equation calculates the number of samples needed from each population and assumes equal sample sizes. If the number of samples from one population is constrained (for example, few ungrazed sites), it would be necessary to adjust the sample size of the unconstrained population. When n exceeded 30, values for infinite sample size were substituted because differences in results were minimal.

When taking a sizeable sample (more than 5 percent) without replacement from a finite population, each observation "carries" more information than when sampling with replacement or from an infinite population. This "extra information" results in a slight decrease in the variance, accomplished by multiplying the usual variance by the finite population correction factor, $(1-n/N)$ where N is the number of elements of the population and n is the sample size. The value n/N is known as the sampling fraction. Corrections for finite populations were not included in our sample size estimates, so our estimates are conservative.

For the woody species regeneration method, the data summary technique presented by Winward (2000) was the ratio of young to old individuals. The description of that method suggests a ratio well over 1 is needed to sustain a healthy woody plant population and a ratio under 1 would indicate that regeneration was not occurring. We did not use the ratio because of the difficulty of analyzing ratio data and a lack of research on the expected value for a given stream type. Instead, we summarized the data as the number of individuals for each of the three age-classes, as well as for all age-classes combined, for each observer at each site.

Effective ground cover data were summarized as the percentage of the points (one for each step) along the cross-section transects that had effective ground cover. SD and CV values were calculated with these data.

Description, Results, and Discussion for Each Method

Greenline Composition

Observers determined the CTs along the greenline, which is the first rooted line of perennial vegetation adjacent to the stream (Winward 2000). In this paper, the term "greenline" refers to the sampling method as well as the location of the sampling area. Observers recorded the CT for each step along 110 m of both streambanks. These greenline CT data were summarized as a percentage of steps for the reach, and as ratings using the following three data reduction techniques.

- Greenline Stability Rating (or stability rating)—The CT stability classes of Winward (2000), ranging from 1 (low stability) to 10 (high stability), were used to assess the vegetation's ability to withstand the force of moving water. The percent of each CT at a site was multiplied by the CT stability class, and the resulting values were added to obtain the greenline stability rating for the site. There are no defined units for this rating, so the term "units" is used.
- Greenline Successional Rating (or successional rating)—The categorization of CTs as "early" or "late" successional by Winward (2000) was used to calculate the percent of late successional vegetation at a site.
- Greenline Wetland Rating—CT wetland ratings (described above) were used to indicate the abundance of species that grow in wetland conditions (Coles-Ritchie, in preparation). The percent of each CT along the greenline was multiplied by the CT wetland rating, and the resulting values were added to determine a greenline wetland rating for the site. There are no defined units for this rating, so the term "units" is used.

Greenline Measurement Results—The measurement data had a mean CT agreement for all observers of 38 percent for the 120 1-m units of vegetation along the greenline (table 3). The maximum agreement of CTs at a measurement site was 49 percent, and the minimum agreement was 29 percent. All seven observers recorded identical CTs at 5 percent of the 1-m units on the greenline. The fuzzy agreement calculation increased the average agreement to 48 percent because of the similarity in species composition between CTs.

The measurement CT data were also converted to the corresponding value for the three ratings and then compared among observers for every 1-m unit (table 3). The precision estimates for greenline stability and wetland ratings resulted in a CV of 9.1 and 10.7 percent, respectively. The greenline successional rating had 83 percent average agreement between observers, although with only two possible categories, random agreement would be 50 percent.

Greenline Repeat Results—For repeat sites, observers recorded an average of 10 CTs per site on the greenline. The maximum number of CTs recorded by an observer at one site was 18, and the minimum was three. The average within-site (all observers at one site) similarity in greenline CT data was 65 percent in 2000 and 51 percent in 2001 (table 4; fig. 2). The final solutions for the NMDS ordination of the greenline data are presented in table 5.

The CT data were converted to ratings and were then evaluated by site. The greenline stability rating had a CV of 7.4 percent, a SD of 0.6 units, and the largest deviation from the mean by an observer was 2.4 units (table 6; fig. 3). The greenline successional rating had a CV of 13.4 percent, a SD of 10.3 percent, and the largest difference between an observer and the grand mean was 29.7 percent (table 6; fig. 4). The greenline wetland rating had a CV of 5.9 percent, a SD of 4.4 units, and the largest difference from the grand mean of 11.9 units (table 6; fig. 5).

Table 3—Measurement site summary statistics describing observer agreement for riparian vegetation sampling methods. Values are based on the total number of sample units; 120 1-m units for the greenline and vegetation cross-section methods, and 26 plants for the willow regeneration method from all six measurement sites in 2001.

Method	Data summary technique	Mean	SD	CV	Mean agreement
					Percent
Greenline	CT agreement	*	*	*	38
	CT fuzzy agreement				48
	Stability rating	8.2	0.7	9.1	*
	Successional rating (percent)	*	*	*	83
	Wetland rating	81.1	8.1	10.7	*
Cross-section	CT agreement	*	*	*	39
	CT fuzzy agreement				50
	Wetland rating	81.7	10.1	13.0	*
Willow regeneration	Species ID	*	*	*	76
	Genus ID	*	*	*	95
	Age class	*	*	*	71
	Height class	*	*	*	94

* not applicable.

Table 4—Repeat site observer variability, based on CT cover data for six to seven observers who each collected data at six riparian sites. The same six sites were sampled in 2000 and 2001.

Data set	Within-site similarity	Agreement statistic
	Percent	
Greenline (2000)	64.9	0.436
Greenline (2001)	50.5	0.285
Cross-section (2000)	49.0	0.314
Cross-section (2001)	39.6	0.221

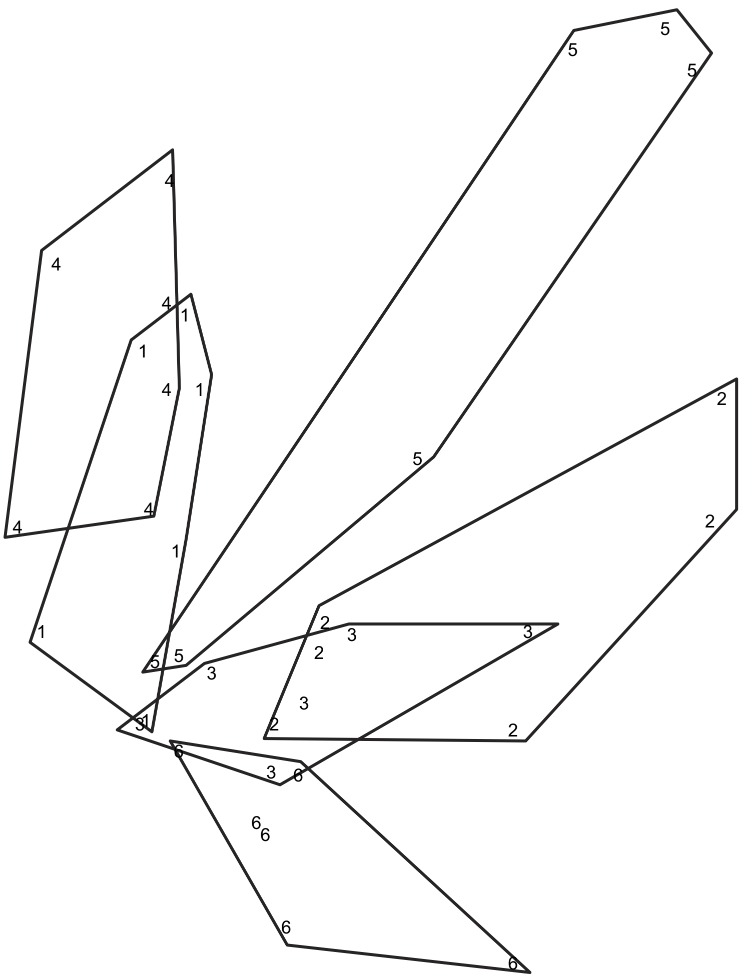

Figure 2—Nonmetric multidimensional scaling (NMDS) ordination of greenline community type data for the six observers in 2001 at the six repeat sites. The ordination indicates the similarity of data among sites (numbered 1 through 6), and for the six observers (same number) at each site. A convex hull surrounds the six observers at each site.

Table 5—The final solutions of the NMDS ordinations of riparian sites, based on cover of community types. Dimensions selected were those that were both statistically significant and that each decreased stress by at least five. The R^2 value is based on the correlation coefficient between the ordination distances and distances in the original n-dimensional space.

Data set	Dimensions selected	Iterations	Final instability	Final Stress	R^2
Greenline (2000)	2	44	0.00007	11.3	0.893
Greenline (2001)	3	200	0.03312	13.1	0.882
Cross-section (2000)	3	200	0.00268	12.8	NA
Cross-section (2001)	4	200	0.00483	10.2	0.593

Table 6—Repeat site observer variability statistics for quantitative summaries, or indices generated from CT cover data. The mean values are based on 50 sites to account for site variability. All other values are based on the six repeat sites. The stability and wetland ratings do not have units.

Method	Data summary technique	Mean	SD	CV	Variability attributable to observer
					Percent
Greenline	Stability rating	7.6	0.6	7.4	35.7
	Successional rating (percent)	76.3	10.3	13.4	27.8
	Wetland rating	75.6	4.4	5.9	15.8
Cross-section	Wetland rating	70.7	7.6	10.8	44.0
	Width (meters)	40.0	7.3	18.3	44.7
Effective ground cover	Cover of vegetation, litter, or rock (percent)	97.7	2.2	2.2	65.4

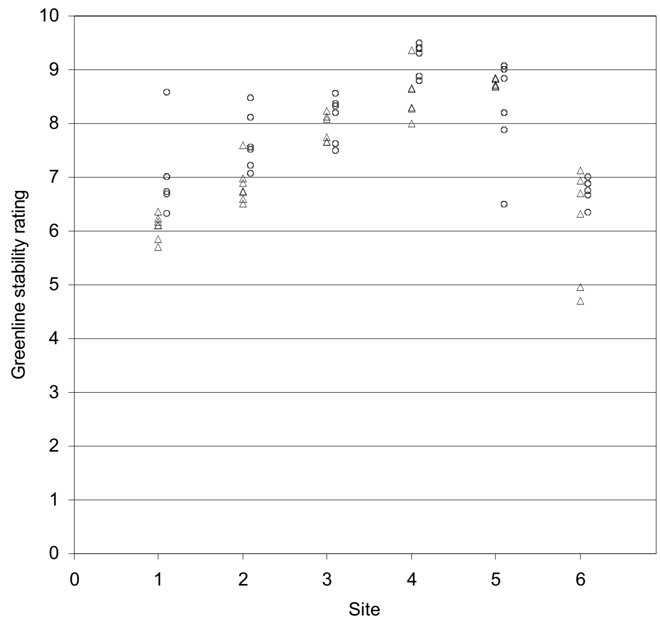

Figure 3—Greenline stability ratings for all observers at the six repeat sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

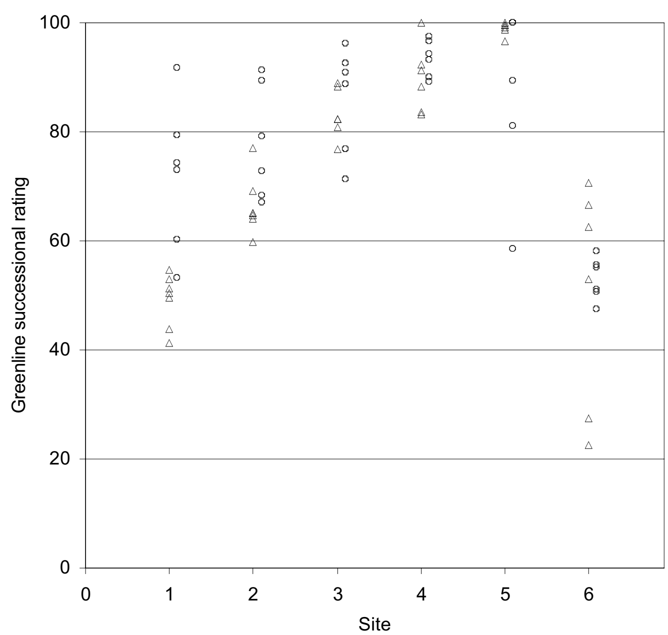

Figure 4—Greenline successional ratings for all observers at the six repeat sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

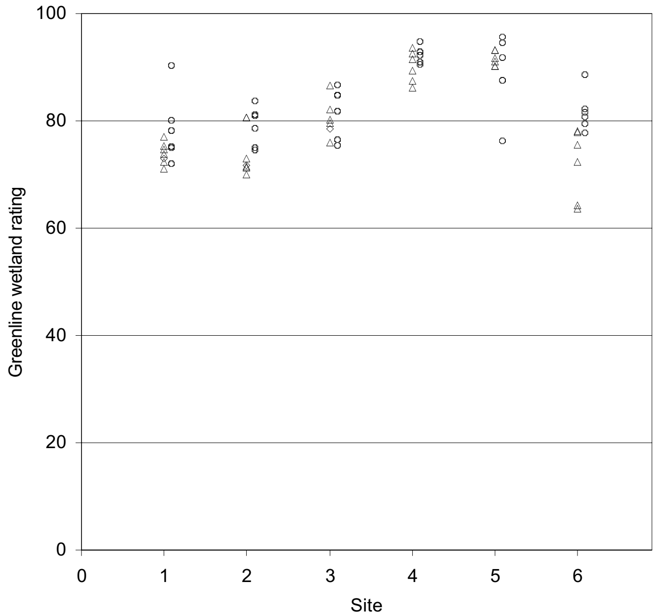

Figure 5—Greenline wetland ratings for all observers at the six repeat sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

The percent of the total variability attributed to observers was 35.7 percent for the stability rating, 27.8 percent the successional rating, and 15.8 percent for the greenline wetland rating (table 6). The sample sizes necessary to detect change were calculated to describe the implications of the total variability for monitoring efforts that seek to detect change at multiple sites. To detect a change of 10 percent (power was set to 0.9 for these calculations) between two populations, considering both observer and site variability, the following number of sites would be needed for each technique (double the value in the 10 percent column of table 7): 56 for stability, 224 for successional, and 78 for greenline wetland rating.

Greenline Discussion—Observer agreement for CT data was relatively low as indicated by the lack of clustering of sites in the ordination (fig. 2). There was also overlap of sites in the ordination (fig. 2) as indicated by the overlap in the convex hulls that surround all the observers at a given site. In some cases, observers differed almost as much at a site as sites differed from each other, which would make detection of change or differences among sites difficult.

Although the three rating techniques were derived from the same CT data, differences in observer variability were found. The low CVs (less than 8) for the greenline stability and greenline wetland ratings indicate that observer precision was relatively good. The successional rating, with a CV of 13.4, was less precise.

Another way to assess the ability to detect change is to consider the percent of the total variability attributed to observers: observer variability / (observer + site variability). Values less than 20 percent (Clark and others 1996; Ramsey and others 1992) or less than 33 percent (Kaufmann and others 1999) are desirable. The wetland rating met both criteria. However, the successional rating exceeded the 20 percent criteria, and the stability rating exceeded the 33 percent criteria. This suggests that differences between observers are almost as great as differences between sites for these two variables. Therefore, it is unlikely that the stability and successional ratings would change enough (within these two National Forests) to be detectable given the differences between observers.

These results exemplify the difficulty of evaluating the usefulness of each method for monitoring. For example, the stability rating was precise (CV of 7.4) but had high variability attributed to observers (35.7 percent). In addition, the sample size to be able to detect a 10 percent change was 56 sites (28 for two different groups), which may be prohibitive for some studies. The wetland rating was precise (CV of 5.9) and had a relatively low variability attributed to observers (15.8 percent), but it still had a sample size of 78 sites (39 for two populations). Even though the wetland rating had a CV and percent variability attributed to observer that was lower than the stability rating, it had a higher sample size requirement because site variability (among all 50 sites) was higher than for the stability rating.

Table 7—Minimum sample sizes needed to detect change with riparian vegetation sampling methods when both observer variability and site variability are considered. Sample size estimates assume equal size samples, so values listed below indicate half the total sample needed. The value listed in each column is the sample size needed to detect the stated change with a type I error of 0.1 and a type II error of 0.1.

Method	Data summary technique	Sample sizes needed to detect a change of: 5 percent	10 percent	20 percent	30 percent	40 percent	50 percent
Greenline	Stability rating	107	28	8	4	3	3
	Successional rating	446	112	28	14	8	6
	Wetland rating	150	39	11	6	4	3
Cross-section	Wetland rating	183	46	13	6	4	3
Effective ground cover	Cover of vegetation, litter, or rock	6	4	1	1	1	1

Combining estimates of variability for both site and observer gives a useful view of the significance of the different sources of variability. When combined to calculate a sample size necessary to detect a change due to management, we believe that this represents an unambiguous and powerful way to display the consequences of variability to scientists and managers. All three of the greenline summary techniques had sample size estimates over 56 sites to detect a change of 10 percent, and over 200 sites to detect a change of 5 percent. Those large sample size requirements would limit the usefulness of these techniques for some types of monitoring studies.

Vegetation Cross-Section Composition

Observers recorded the CT for each step along five cross-sections perpendicular to the valley floor (as per Winward 2000). These cross-sections were within the sampling area established for greenline composition and extended to the edge of the riparian vegetation, up to a maximum of 27.5 m on both sides of the stream. The edge of the riparian area was defined as the point when nonriparian communities were encountered, or when riparian species no longer constituted at least 25 percent of the vegetation cover. The data were summarized using the following techniques:

- Cross-Section Wetland Rating—The percent of each CT in all five cross-sections was multiplied by the CT wetland rating (described above), and the resulting values were added to obtain a site cross-section wetland rating. There are no defined units for this rating, so the term "units" is used.
- Cross-Section Width—The distance across the riparian area (excluding the stream), up to 27.5 m on each side of the stream. This distance was calculated using known step lengths of the observers. These data were used to identify a potential source of observer variability, not as a monitoring tool.

Vegetation Cross-Section Measurement Results—The measurement results of the vegetation cross-section method were similar to the greenline method. The mean CT agreement for all observers was 39 percent for a given 1 m of vegetation along a cross-section (table 3). The maximum agreement at a site was 55 percent and the minimum was 29 percent. All seven observers recorded the same CT at 10 percent of the 1-m units at measurement sites. The fuzzy agreement calculation increased the CT agreement to 50 percent. Converting the CTs into wetland ratings for each step resulted in a CV of 13.0 units.

Vegetation Cross-Section Repeat Results—Observers recorded an average of 7.4 CTs per repeat site for the cross-sections. The maximum number of CTs recorded by an observer at one site was 12 and the minimum was four. The average within-site similarity in cross-section CT data was 49 percent in 2000 and 40 percent in 2001 (table 4; fig. 6).

The cross-section wetland rating had a CV of 10.8 percent and a SD of 7.6 units (table 6; fig. 7). The largest difference between an observer and the grand mean for a site was 18.2

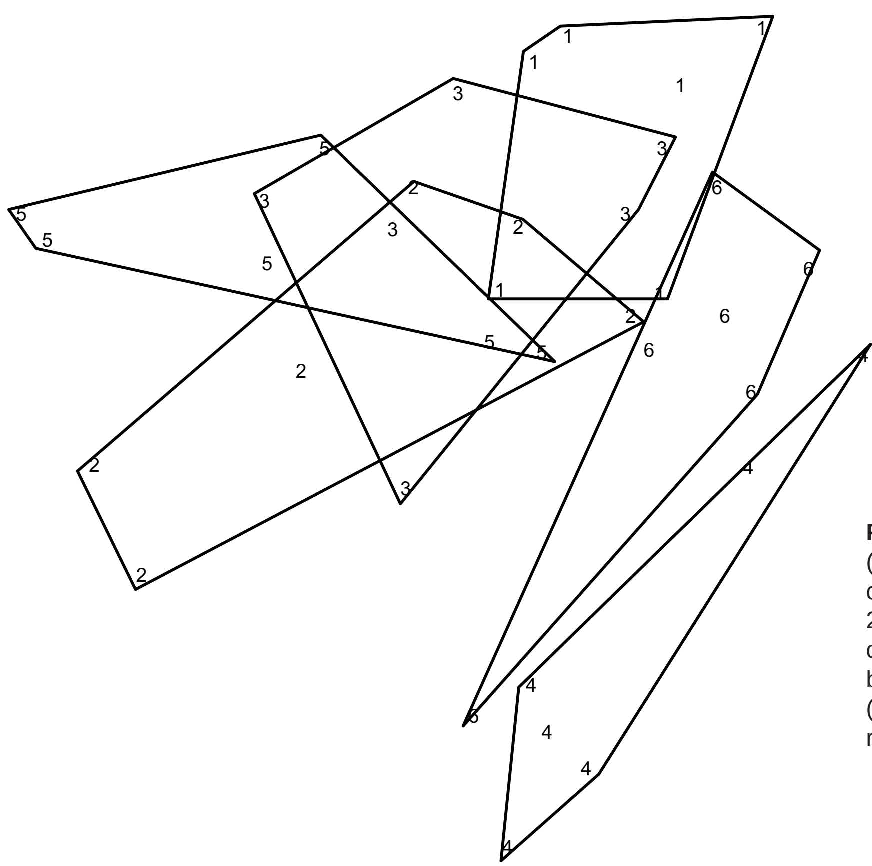

Figure 6—Nonmetric multidimensional scaling (NMDS) ordination of vegetation cross-section community type data for the six observers in 2001 at the six repeat sites. The ordination indicates the similarity of data among sites (numbered 1 through 6), and for the six observers (same number) at each site. A convex hull surrounds the six observers at each site.

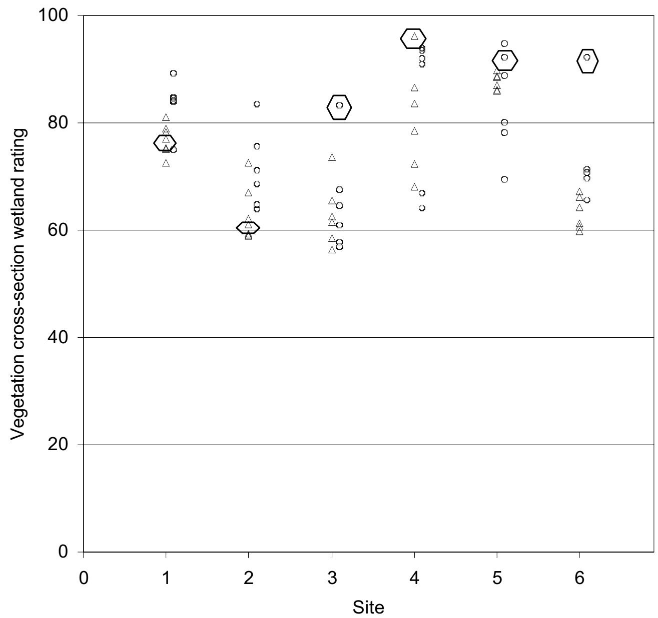

Figure 7—Vegetation cross-section wetland ratings for all observers at the six repeat sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right. The observers circled (with a hexagon) are the same observers circled in figure 8.

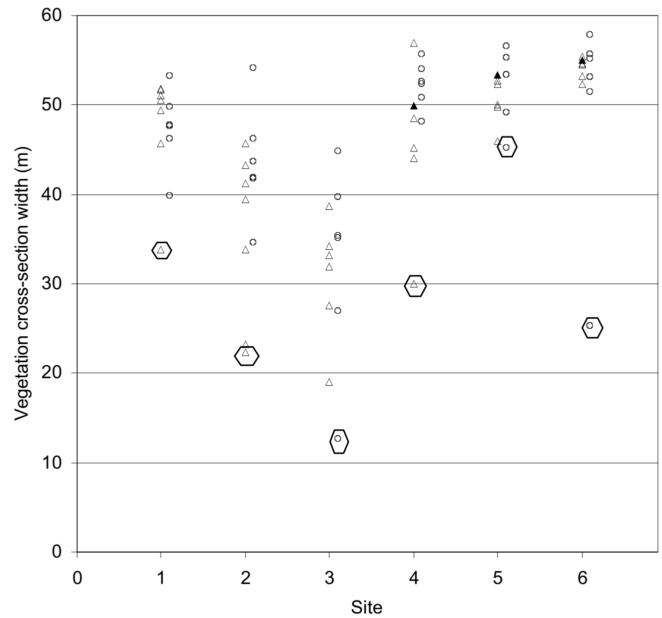

Figure 8—Average width of vegetation cross-sections for all observers at the six repeat sites. Observers in 2000 are represented by triangles and observers in 2001 are represented by circles, which are offset slightly to the right. The observers circled (with a hexagon) are the same observers circled in figure 7.

units. The cross-section width had a CV of 18.3 percent and a SD of 7.3 m (table 6; fig. 8). The largest difference between an observer and the grand mean for a site was 24.5 m.

The percent of the total variability attributed to observers was 44.0 percent for the vegetation cross-section wetland rating (table 6). The total number of sites needed to detect a 10 percent change (power was set to 0.9 for these calculations) would be 92 sites for the cross-section wetland rating (table 7).

Vegetation Cross-Section Discussion—Observer agreement of CT data was similar to the greenline method for the measurement data. The repeat data had an average similarity among observers that was 11 to 15 percent lower for this method than for the greenline method. This low CT agreement among observers is evident by the separation of observers at one site (same number) in the ordination (fig. 6). As with the greenline method, there is overlap among sites in the ordination, which means that using the cross-section data to detect differences in vegetation among sites would be difficult with multiple observers.

The cross-section wetland rating had a moderate level of precision relative to the other methods we tested. However, almost half of the total variability was due to observers, meaning that the variability among observers at a site was almost as large as the variability between the mean values for the six sites (fig. 7). The combination of observer and site variability resulted in a relatively large sample size needed to detect a 10 percent change. Therefore it would be difficult to determine whether a change in this rating reflected an actual change or was only an artifact of observer difference.

One source of variability for the vegetation cross-section method was differences in defining the edge of the riparian area. The location of the riparian edge was determined independently by each observer based on vegetation and landform, both of which are subjective judgments. Comparison of the vegetation cross-section widths and the wetland ratings for the same observers suggests that precision was sometimes affected by differences in defining the edge of the riparian area. Differences in the wetland rating would be expected because there is generally a decrease in obligate wetland plants moving away from the stream. To highlight this influence, the lowest value for the cross-section width for each site is circled (with a hexagon, fig. 8), and the value for the same observer is also circled (with a hexagon) on the cross-section wetland rating (fig. 7). For four of the six sites the lowest riparian width was associated with the highest or second highest wetland rating for that site. This suggests that defining a shorter riparian width can sometimes cause the wetland rating to be higher than it would be if a larger area were sampled. These data underscore the importance of using methods that result in a consistent sampling area. Winward (2000) suggests using permanent markers, which would alleviate the problem of variable width. In some situations, permanent markers are not feasible, in which case a more repeatable method of identifying the riparian edge, or techniques to always consider a comparable area, would be needed.

Causes of Observer Differences Common for Both Methods

Vegetation classifications have relatively few categories compared to species data, and therefore it has been presumed that they are easier to use. Alpert and Kagan (1998) noted that "easy detection" is one of the advantages of using CTs for management. Winward (2000) stated that one of the objectives of the greenline and cross-section methods was "to be efficient in both time and cost." However, while detection may appear easy, our results indicate that CT detection is not always easy or repeatable.

One problem with using CTs as the cover category is the difficulty in distinguishing CTs on the ground. In nature there are no distinct boundaries between communities, which makes it difficult for observers to consistently categorize the vegetation. There has been considerable debate in vegetation ecology about the appropriateness of CTs because they are not "real entities" (Alpert and Kagan 1998). Some CTs may seem obvious on the landscape, but as our results indicate, communities cannot always be consistently determined. It can be especially difficult to distinguish communities in riparian areas where disturbance from the stream creates small patches of vegetation and therefore more transition area between communities.

The overlap in species composition among CTs is another reason for the disagreement in determining the CT. This is evident by the similarity of the *Alnus incana*/mesic graminoid CT and the *Alnus incana*/*Equisetum arvense* CT, which are 65 percent similar (based on a Bray-Curtis dissimilarity index using species cover and constancy data from Padgett and others 1989). In addition, both of those CTs are about 60 percent similar to the *Alnus incana*/mesic forb CT. Many of the same species are found in those three CTs, so observers might justifiably use any of the three to describe the same vegetation.

The misidentification of species, or differences in estimating the species with the greatest cover, also contributed to variability in CT determination. However, converting the vegetation data to ratings often decreased the variability among observers due to misidentification or differing judgment of the dominant species. For example, at one repeat site, observers recorded three different sedge-dominated CTs (*Carex rostrata*, *Carex nebrascensis*, and *Carex aquatilis* CTs) along the greenline. This resulted in disagreement in CT data; however, it had little affect on the greenline ratings because all three CTs had the same successional (late) and stability (9) ratings (Winward 2000), and similar wetland ratings (96 to 97, for CTs of Padgett and others 1989). The dominant species in these CTs are probably ecological equivalents, in that they grow in similar habitats and perform similar ecosystem functions–such as bank stabilization and sediment trapping. In this and other cases, taxonomic differences, which were sometimes difficult to determine in the field, were not always important when trying to understand the functioning of the riparian ecosystem.

While the previous example showed that converting CTs into more general categories increased observer agreement, in some situations converting data to ratings did not improve agreement. This was evident at a site where some observers recorded more of the *Calamagrostis canadensis* CT (found in moist conditions) and others recorded more of the *Poa pratensis* CT (found in a drier environment). The differences in these CTs are reflected in the values of the stability ratings (7 versus 3), greenline successional ratings (late versus early), and the wetland ratings (85 versus 48). The fact that observers recorded these different grass CTs for the same area increased the variability in the rating for this site.

A related discrepancy occurred when some observers recorded more of a grass CT, and other observers recorded more of the wet sedge CTs, apparently because the grass flowers were taller than the sedges, which did not have flowers. Our data did not allow a determination of the sources for differences in CT determination among observers, but it was likely a difference in species identification, or a difference in estimating which species was dominant.

Other studies that have assessed the precision in estimating canopy cover found results ranging from low variability in a controlled experiment (Hatton and others 1986) to moderate to high variability using field sampling methods (Barker and others 2002; Pollard and Smith 1999; Smith 1944). When using CTs as the cover category, the dominant species is used to determine the CT, so variability in estimations of canopy cover can amplify differences among observers.

A problem encountered by this study was that a riparian vegetation classification had not been developed for central Idaho, where the study sites were located. The CTs described in the classifications by Hansen and others (1995) and Padgett and others (1989) covered much, but not all, of the vegetation encountered at these sites. Plant assemblages that were not documented in the classifications caused uncertainty and therefore variability among observers. Any large study that relies on CTs will encounter a similar problem, because classifications have not been done for all geographical areas, and those that are available have been developed in different, and sometimes incompatible, manners. For example, some classifications include sites that have been disturbed by human activities, while others do not.

We were also interested in whether the spatial area of observation affected observer variability when determining CTs. The area of observation was at different scales for the greenline and vegetation cross-section methods (Winward 2000). For the greenline the area of observation was only 0.3 m wide, and the CT was determined for each step (approximately 0.8 m). This area was sampled to assess the ability of vegetation in "buffering the forces of moving water" (Winward 2000). This was substantially smaller than the 4 m^2 to 400 m^2 area considered by most riparian vegetation classifications, including those used for this study (Hansen and others 1995; Padgett and others 1989). For the vegetation cross-section method, observers were instructed to consider an area of approximately 50 m^2, although they seemed to focus on a smaller area as they recorded the CT for each step. The area considered for the cross-sections was more similar to the large area used to define communities in the classifications. Given the different scales for the greenline and vegetation cross-section methods, it is interesting that observer agreement was similar, both from individual locations in the measurement study (38 percent versus 39 percent average agreement, respectively) and for the average site ratings in the repeat study. This suggests that determining CTs at different spatial scales, such as with greenline (small scale) and cross-sections (somewhat larger scale), is equally variable.

Woody Species Regeneration Method

This method was designed to assess the regeneration of woody plants for which regeneration is potentially inhibited by herbivory from ungulates (such as livestock, deer, elk, and moose). Because we found no comprehensive list on the palatability of woody plants, we chose to consider only willows, which are generally palatable to ungulates and are one of the more abundant woody plants along streams (Hansen and others 1995; Kovalchik and Elmore 1992). Therefore, we refer to this as the "willow regeneration" method.

This method involved the counting of willow plants (*Salix* spp.) within 1 m of the greenline along both streambanks (Winward 2000). Data were recorded on the species, age-class, and height-class of each willow encountered. Rhizomatous species were not counted because individual plants are difficult to distinguish and to age, as per Winward (2000). The observers counted the number of individual willow plants by species, in these categories: seedling/sprout, young, mature, and decadent. For this study we combined decadent with mature because most observers recorded no decadent individuals. Most willow species encountered were shrubs, or multistemmed species, for which the number of stems was used to determine an age-class: one stem was a seedling/sprout; two to10 stems was young; greater than 10 stems was mature. The counts were truncated at 50 for a given age-class for each species, if that many individuals were observed. The height of each willow was recorded as greater than 1 m or less than 1 m.

Willow Regeneration Measurement Results—Observer agreement in identifying the genus *Salix* was 95 percent, while agreement on the species, within the genus *Salix*, was 76 percent (table 3). Observers agreed on the age-class 71 percent of the time and height class of plants 94 percent of the time.

Willow Regeneration Repeat Results—For the 2001 data the smallest difference in the number of seedlings/sprouts among observers at a site was three and the largest difference was 69 (table 8; fig. 9). The smallest difference in the number of young willows was six and the largest difference was 50 (table 8; fig. 10). The smallest difference in the number of mature willows was two and the largest difference was 42 (table 8; fig. 11). The smallest difference in the number of total willow plants was seven and largest difference was 100 (table 8; fig. 12). Similar patterns for the 2000 data can be observed in figures 9 through 12.

Table 8—The range in the number of willow plants, by age-class, counted by six observers at six riparian sites in 2001.

Site	Seedling/ sprouts	Young	Mature/ decadent	Total plants
	- - - - - - - -	- - - - *Range*	- - - - - - - -	- - - - - - - -
1	0 - 6	0 - 43	0 - 26	0 - 53
2	0 - 10	5 - 51	4 - 26	9 - 70
3	0 - 3	0 - 6	0 - 2	0 - 7
4	0 - 50	0 - 50	0 - 5	0 - 100
5	1 - 6	0 - 7	0 - 8	5 - 18
6	0 - 69	7 - 53	0 - 42	50 - 91

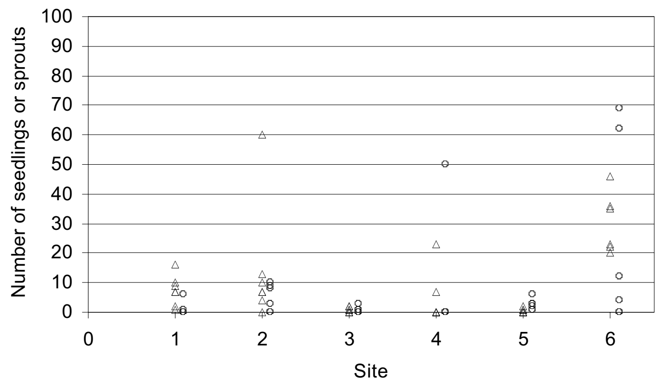

Figure 9—The number of willow seedlings and sprouts counted by observers at six sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

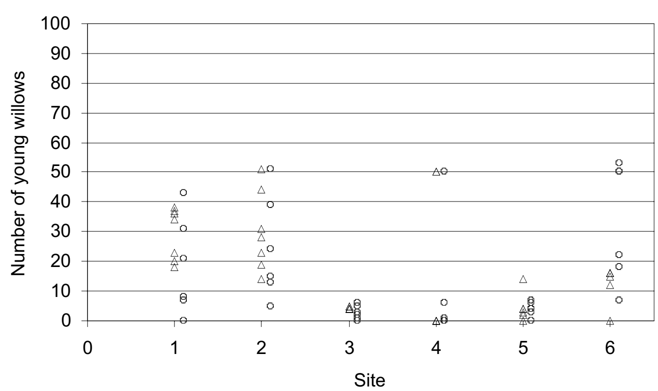

Figure 10—The number of young willow plants counted by observers at six sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

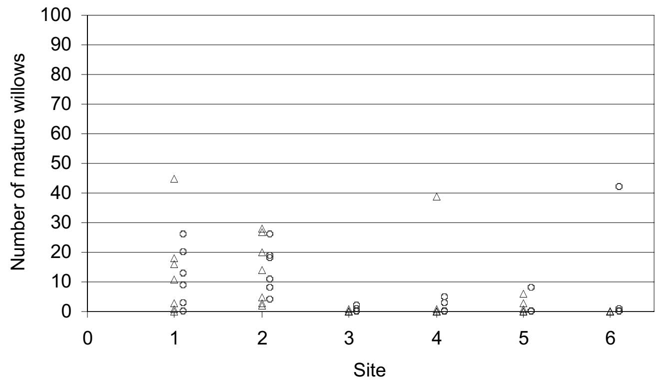

Figure 11—The number of mature or decadent willow plants counted by observers at six sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

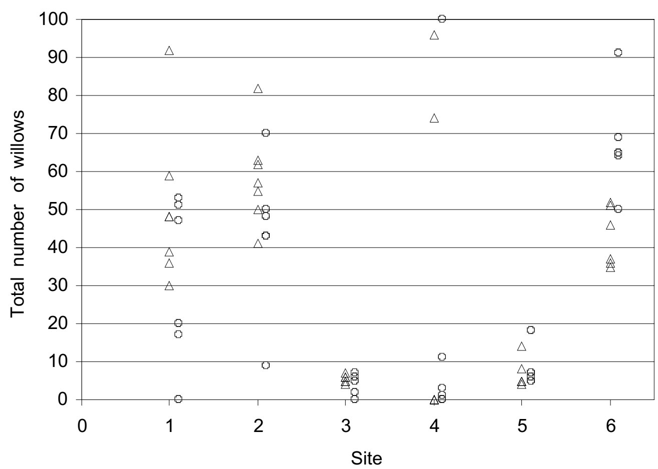

Figure 12—The total number of willow plants counted by observers at six sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

Willow Regeneration Discussion—The willow regeneration method had high observer variability as is evident in figures 9 through12. Given this variability and the diversity of sites sampled in this project, in terms of presence or absence of willows, we are unable to suggest modifications to improve the repeatability of this method so that it would be useful for monitoring riparian woody plant regeneration for a broad-scale study. However, we recognize that it may have utility with specific woody plant populations and at individual sites.

Our results identify two main factors that affected the precision of the willow regeneration method. First, observers had difficulty determining the age-class of individual plants, especially when deciding between the young and mature classes. This was due to the difficulty in distinguishing adjacent (often intermixed) individuals with many stems, which often grew in dense stands. Managers in central Idaho, who used this method to assess grazing management, also had difficulty with multiple-stemmed willow (rhizomatous or nonrhizomatous) species and concluded that counting individuals to assess woody species regeneration may not be possible (Burton and others 1998). Thorne and others (2002) also noted the difficulty of distinguishing individual willows in a test of canopy estimates.

Secondly, differences in plant identification, when rhizomatous and nonrhizomatous species were confused, led to observer variability. At two sites, observers disagreed on the species of willow with some identifying the plants as a rhizomatous species and others a nonrhizomatous species. This resulted in over 100 individuals being counted by one observer and none by another observer. When that type of error occurs it can lead to notably different data among observers.

Other issues that may have added to observer variability with this method involved sampling different areas and missing small plants. Determining the area of observation involved the subjective judgment of determining where the greenline was located. If observers considered the greenline to be at a different location, then they would have sampled different areas and therefore encountered different numbers of willows. In addition, observing seedlings less than 0.3 m tall was difficult because of their small size. Observers probably missed young individuals that were obscured by other plants.

Other studies have also found high observer variability in methods used to describe browsing impacts on woody species. Hall and Max (1999) studied the differences among 15 observers who assessed utilization by measuring twig length and found that observer variability was twice the size of variability among shrubs. Thorne and others (2002) found statistically significant differences among observers for a shrub canopy volume method, although estimates from a single observer were not statistically different. Winward (2000) describes a line-intercept method to estimate the cover of rhizomatous plants within the riparian area, which could potentially be used for all woody species, although it lacks information about age-classes.

This woody species regeneration method was not found to be repeatable, and we are not aware of any other repeatable methods to assess the impacts of grazing on woody species. This is unfortunate because utilization and recruitment of woody species, such as willows, would seem to be indicators of overgrazing or recovery.

Effective Ground Cover Method

Observers estimated the types of ground cover within the riparian area, based on the USDA Forest Service Region 4 soils protocol (USDA 1989). At each step along the five vegetation cross-sections, the observer looked at a 2-cm circle in front of his/her big toe and recorded the dominant cover as bare ground, live vegetation, litter, or rock. The step totals were converted to a percentage of steps with effective ground cover (live vegetation, litter, or rock) versus bare ground.

Effective Ground Cover Measurement Results—Measurement studies were not conducted for this method.

Effective Ground Cover Repeat Results—The variability associated with observer estimates of effective ground cover had a CV of 2.2 percent and an SD of 2.2 percent (table 6; fig. 13). The largest deviation from the grand mean by an observer was 4.5 percent.

Observer variability comprised 65.4 percent of the total variability for the effective ground cover data. A sample size of eight sites would be needed to detect a 10 percent difference in condition (power was set to 0.9 for these calculations; table 7).

Effective Ground Cover Discussion—Effective ground cover estimates had the highest precision and lowest sample size of the methods we tested. The data indicate that a change in effective ground cover of 5 percent could be detected with a sample size of only six sites. However, the variability attributable to observer was 65 percent of the total variability, the highest of any method. This is an indicator that this method is not as useful as the sample size and CV values would suggest. The limited variability between sites is the reason that the observer variability was such a large percentage of the total variability. The 50 sites in this study all had effective ground cover greater than 87 percent, and a mean of 98 percent. Therefore, it is unlikely that a 5 percent change would occur at the sites sampled by this project in the Payette and Nez Perce National Forests. This underscores the danger of considering only one statistic, such as a CV, to evaluate the usefulness of a method.

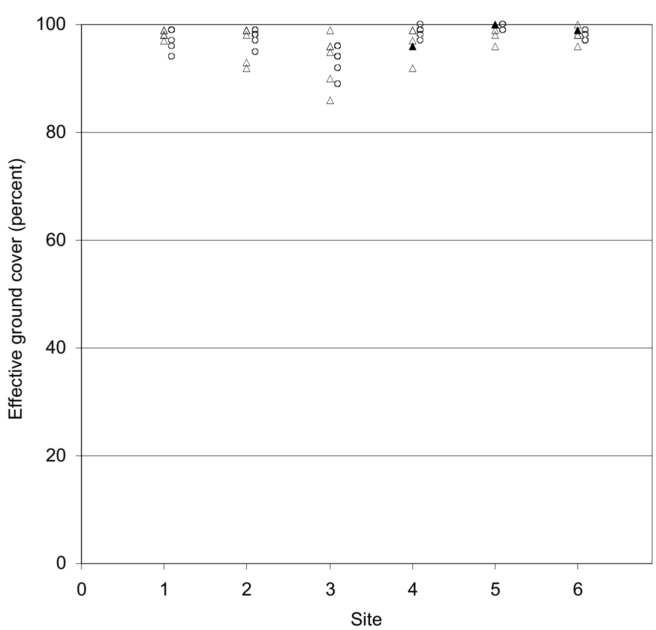

Figure 13—Effective ground cover ratings for all observers at the six repeat sites. Observers in 2000 are represented by triangles, and observers in 2001 are represented by circles, which are offset slightly to the right.

The effective ground cover method would only be useful in geographical areas where effective ground cover values have greater potential to vary, which may be the case in drier environments. However, the method may not have the same level of precision in areas where the effective ground cover is more variable because there would also be more potential for observer differences.

Conclusions

The objective of most riparian monitoring efforts, including the PIBO-EMP, is to detect changes in habitat characteristics that are caused by anthropogenic disturbances. Our ability to detect these changes is affected by the ability of observers to consistently characterize riparian vegetation (observer variability) and the heterogeneity of sites (site variability). Understanding the magnitude of these types of variability is essential in designing monitoring studies, which almost inevitably involve multiple observers.

Results from the greenline and vegetation cross-section methods indicate that CT data (as percent cover for a site) were not consistent enough among observers to be able to detect change in vegetation unless the change was dramatic. The variability in the CT data was due to the subjectivity of distinguishing CTs, especially at the scale of a step, which is what these methods require. CTs may not be definitive enough to generate consistent data among observers at the scale of a step. For studies involving multiple sites and multiple observers over time, CT cover data would be of limited use for detecting change.

When CT data were converted to ratings, which emphasize ecosystem attributes important to managers, observer agreement was better. Ratings have the advantage of minimizing the influence of differences in species or CTs that are unimportant, at least for a given rating. We found that the greenline stability, greenline wetland, and vegetation cross-section wetland ratings would be precise enough to detect large changes (greater than 20 percent) with feasible sample sizes (fewer than 13 sites in each of two populations). Detecting a smaller change (10 percent) with these three ratings would require larger sample sizes (between 28 to 46 sites in each of two populations), which may be impractical for many studies.

The greenline successional rating had much higher observer variability, suggesting that change detection would be difficult with that rating technique. The successional rating has only two categories (early and late), and therefore observer differences are always influential on the rating. The most useful ratings are probably those with many gradations that are based on scientific information about vegetation.

While the ratings seemed to have greater precision among observers than the CT cover data, ratings would still only permit detection of large changes in riparian vegetation. If a large change occurred quickly, then the wetland and stability ratings could prove useful. If change took a long time to occur, which is often the case, these ratings would be less useful. Negative change detection could take so long that resource damage may already have occurred before it was detected. Positive change detection could also take a long time, failing to provide prompt feedback about the effectiveness of management.

The woody species regeneration method had high variability among observers. Observers recorded different numbers of woody individuals (willows in this study) because of the difficulty distinguishing individual shrubs, variability in age-classing individuals, difficulty finding seedlings because of their small size, and errors in identification that mixed up a nonrhizomatous species with a rhizomatous (which cannot be age-classed with this method). This woody species regeneration method was found to be ineffective for monitoring streambank woody plant regeneration.

The effective ground cover method was precise among different observers; however, there was little difference in this attribute among sites. All sites in this study had high effective ground cover values, so little change would be expected in that geographical area. In geographical areas where there is more potential for change in effective ground cover, this method might be more useful if quality assurance tests for that area resulted in low observer variability.

The results from this study underscore the need for repeatable methods of monitoring riparian vegetation. Numerous researchers and land managers have collected data on riparian vegetation, but few protocols exist for systematic monitoring of riparian areas. The protocol of Winward (2000), evaluated here, was one of the first for monitoring riparian vegetation, and it has focused attention on the importance of riparian vegetation. The objectives of riparian monitoring protocols should be precision (repeatability with multiple observers), accuracy, and feasibility for summer field crews. We are evaluating other methods based on those criteria.

We calculated sample sizes needed to detect changes between two populations,. However, land managers are often interested in documenting changes at a group of fixed sites. For a study of the same sites over time, the variance between sites would be greatly reduced (Roper and others 2003). Therefore, the sample size calculations would be primarily a function of observer variability. Therefore, studies that resample permanent sites would require smaller sample sizes to detect a change than studies that compare samples of two populations. In either case, the sites sampled would need to be randomly selected from the population of interest in order to make inferences from the data.

The methods evaluated here were developed to address a significant need, and we see the methods as an important step in the field of riparian monitoring. These methods have motivated many people to look more closely at riparian areas, the ecosystem functions they provide, and how these important functions can be altered by land management activities. However, the levels of repeatability for these methods limit their usefulness for many monitoring questions that seek to detect change. Relatively large changes could be detected with these sampling methods and data summary techniques. Smaller but ecologically important changes in riparian areas would be difficult to detect unless hundreds of sites were sampled. Methods that have greater repeatability among observers are desirable for monitoring because smaller sample sizes would be required to detect change, and because smaller changes could be detected. We are evaluating modifications to these methods, to determine how repeatability, as well as accuracy, can be improved.

References

Alpert, P.; Kagan, J. 1998. The utility of plant community types: a practical review of riparian vegetation classification in the Intermountain United States. Natural Areas Journal. 18: 124-137.

Archer, E. K.; Mellison, S. C.; Henderson, R.; Kershner, J. L.; Roper, B. B. 2004. Testing common stream sampling methods: how useful are these techniques for broad-scale, long-term monitoring? Gen.Tech. Rep. RMRS-GTR-122. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 15 p.

Auble, G. T.; Friedman, J. M.; Scott, M. L. 1994. Relating riparian vegetation to present and future streamflows. Ecological Applications. 4(3): 544-554.

Barker, J. R.; Bollman, M.; Ringold, P. L.; Sackinger, J.; Cline, S. P. 2002. Evaluation of metric precision for a riparian forest survey. Environmental Monitoring and Assessment. 75: 51-72.

Bendix, J.; Hupp, C. R. 2000. Hydrological and geomorphological impacts on riparian plant communities. Hydrological Processes. 14: 2977-2990.

Burton, T.; Ririe, W.; Erickson, J.; Miller, M.; Hardy, T. 1998. 5-year monitoring report for Bear Valley streams and riparian habitat. Boise, ID: U.S. Department of Agriculture, Forest Service, Intermountain Region, Boise National Forest.

Cherrill, A.; Mcclean, C. 1999. Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. Journal of Applied Ecology. 36(6): 989-1008.

Clark, M. J.; Laidlaw, M. C. A.; Ryneveld, S. C.; Ward, M. I. 1996. Estimating sampling variance and local environmental heterogeneity for both known and estimated analytical variance. Chemosphere. 32: 1133-1151.

Clymo, R. S. 1980. Preliminary survey of the peat-bog Hummell Knowe Moss using various numerical methods. Vegetatio. 42: 129-148.

Coles-Ritchie, M. C. [In preparation]. A wetland rating system for evaluating riparian vegetation. Logan: Utah State University. 218 p. Dissertation.

Dunaway, D.; Swanson, S.; Wendel, J.; Clary, W. 1994. The effect of herbaceous plant communities and soil textures on particle erosion of alluvial streambanks. Geomorphology. 9: 47-56.

Eckblad, J. W. 1991. How many samples should be taken? BioScience. 41: 346-348.

Elzinga, C. L.; Salzer, D. W.; Willoughby, J. W. 1998. Measuring and monitoring plant populations. Tech. Reference 1730-1/RS/ST-98/005+1730. Denver, CO: Bureau of Land Management.

Elzinga, C. L.; Evenden, A. G. 1997. Vegetation monitoring: an annotated bibliography. Ogden, Utah: Gen. Tech. Rep. INT-GTR-352. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 184 p.

Gotfryd, A.; Hansell, R. I. C. 1985. Impact of observer bias on multivariate analyses of vegetation structure. OIKOS. 45: 223-234.

Greig-Smith, P. 1957. Quantitative plant ecology. London: Butterworths.

Hall, F. C.; Max, T. 1999. Test of observer variability in measuring riparian shrub twig length. Journal of Range Management. 52: 633-636.

Hansen, P. L.; Pfister, R. D.; Boggs, K.; Cook, B. J.; Joy, J.; Hinckley, D. K. 1995. Classification and management of Montana's riparian and wetland sites. Missoula: University of Montana School of Forestry. 646 p.

Hatton, J. K.; West, N. E.; Johnson, P. S. 1986. Relationships of the error associated with ocular estimation and actual total cover. Journal of Range Management. 39: 1 91-92.

Hope-Simpson, J. F. 1940. On the errors in the ordinary use of subjective frequency estimations in grassland. Journal of Ecology. 28: 193-209.

Kaufmann, P. R.; Levine, P.; Robison, E. G.; Seeliger, C.; Peck, D. V. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. Washington, DC: U.S. Environmental Protection Agency. 102 p.

Kennedy, K. A.; Addison, P. A. 1987. Some considerations for the use of visual estimates of plant cover in biomonitoring. Journal of Ecology. 75: 151-157.

Kent, M.; Coker, P. 1992. Vegetation Description and Analysis: A Practical Approach. Boca Raton, FL: CRC Press. 364 p.

Kershner, J. L.; Coles-Ritchie, M.; Cowley, E.; Henderson, R. C.; Kratz, K.; Quimby, C.; Ulmer, L. C.; Vinson, M. R. 2004. Guide to effective monitoring of aquatic and riparian resources. Gen. Tech. Rep. RMRS-GTR-121. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 57 p.

Kleinfelder, D.; Swanson, S.; Norris, G.; Clary, W. 1992. Unconfined compressive strength of some streambank soils with herbaceous roots. Soil Science Society of America Journal. 56: 1920-1925.

Kirby, K. J.; Bines, T.; Burn, A.; Mackintosh, J.; Pitkin, P.; Smith, I. 1986. Seasonal and observer differences in vascular plant records from British woodlands. Journal of Ecology. 74: 123-131.

Kovalchik, B. L.; Elmore, W. 1992. Effects of cattle grazing systems on willow-dominated plant associations in central Oregon. In: Clary, Warren P. and others. Proceedings—Symposium on ecology and management of riparian shrub communities. Gen. Tech. Rep. INT-289. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station: 111-119.

Leps, J.; Hadincova, V. 1992. How reliable are our vegetation analyses? Journal of Vegetation Science. 3: 119-124.

Littell, R. C.; Miliken, G. A.; Stroup, W. W.; Wolfinger, R. D. 1996. SAS® system for mixed models. Cary, NC: SAS Institute, Inc. 633 p.

Manning, M. E.; Swanson, S. R.; Svejcar, T.; Trent, J. 1989. Rooting characteristics of four intermountain meadow community types. Journal of Range Management. 42(4): 309-312.

McCune, B.; Grace, J. B. 2002. Analysis of ecological communities. Gleneden Beach, OR: MJM Software Design. 300 p.

McCune, B.; Mefford, M. J. 1999. PC-ORD for Windows, Version 4.25. Gleneden Beach, OR: MJM Software Design.

Minchin, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio. 69: 89-107.

Montgomery, D. C. 1984. Design and analysis of experiments. New York: John Wiley and Sons. 538 p.

Nilsson, I. N.; Nilsson, S. G. 1985. Experimental estimates of census efficiency and pseudoturnover on islands: error trend and between-observer variation when recording vascular plants. Journal of Ecology. 73: 65-70.

Padgett, W. G.; Youngblood, A. P.; Winward, A. H. 1989. Riparian community type classification of Utah and southeastern Idaho. R4-ECOL-89-01. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Region. 191 p. plus insert.

Pollard, J. E.; Smith, W. 1999. Forest health monitoring 1998 plot component quality assurance report, Volume 1. Research Triangle Park, NC: U.S. Department of Agriculture, Forest Service, National Forest Health Monitoring Program. 21 p.

Ramsey, M. H.; Thompson, M.; Hale, M. 1992. Objective evaluation of precision requirements for geochemistry analysis using robust analysis of variance. Journal of Geochemical Exploration. 44: 23-36.

Reed, J.; Porter, B. 1996. Draft revision of 1988 national list of plant species that occur in wetlands: national summary. Washington, DC: U.S. Department of the Interior, Fish and Wildlife Service. 215 p.

Roper, B. B.; Kershner, J. L.; Archer, E. K.; Henderson, R.; Bouwes, N. 2002. An evaluation of physical stream habitat attributes used to monitor streams. Journal of the American Water Resources Association. 38: 1637-1646.

Roper, B. B.; Kershner, J. L; Henderson, R. 2003. The value of using permanent sites when evaluating stream attributes at the reach scale. Journal of Freshwater Ecology. 18(4): 585-592.

SAS Institute, Inc. 2000. SAS procedures guide. Release 8.0 ed. Cary, NC: SAS Institute. 441 p.

Smith, A. D. 1944. A study of the reliability of range vegetation estimates. Ecology. 25: 441-448.

Sykes, J. M.; Horrill, A. D.; Mountford, M. D. 1983. Use of visual cover assessments as quantitative estimators of some British woodland taxa. Journal of Ecology. 71: 437-450.

Thorne, M. S.; Skinner Q. D.; Smith, M. A.; Rodgers, J. D.; Laycock, W. A.; Cerekci, S. A. 2002. Evaluation of a technique for measuring canopy volume of shrubs. Journal of Range Management. 55: 235-241.

Toledo, Z. O.; Kauffman, J. B. 2001. Root biomass in relation to channel morphology of headwater streams. Journal of the American Water Resources Association. 37(6): 1653-1663.

U.S. Department of Agriculture, Forest Service. 1989. Soil Quality Monitoring Methods. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Region.

Wang, L.; Simonson, T. D.; Lyons, J. 1996. Accuracy and precision of selected stream habitat estimates. North American Journal of Fisheries Management. 16: 340-347.

Winward, A. H. 2000. Monitoring the vegetation resources in riparian areas. Gen. Tech. Rep. RMRS-GTR-47. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Zar, J. H. 1996. Biological analysis. Englewood Cliffs, NJ: Prentice Hall. 718 p.

Zimmerman, G. M.; Goetz, H.; Mielke, P. W., Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology. 66(2): 606-611.

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and number.

Telephone	(970) 498-1392
FAX	(970) 498-1396
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rm
Mailing Address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

Federal Recycling Program Printed on Recycled Paper

RMRS

ROCKY MOUNTAIN RESEARCH STATION

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona
Fort Collins, Colorado*
Boise, Idaho
Moscow, Idaho
Bozeman, Montana
Missoula, Montana
Lincoln, Nebraska
Reno, Nevada
Albuquerque, New Mexico
Rapid City, South Dakota
Logan, Utah
Ogden, Utah
Provo, Utah
Laramie, Wyoming

*Station Headquarters, Natural Resources Research Center,
2150 Centre Avenue, Building A, Fort Collins, CO 80526