Rangeland and pasture monitoring: an approach to interpretation of high-resolution imagery focused on observer calibration for repeatability

Michael C. Duniway · Jason W. Karl · Scott Schrader · Noemi Baquera · Jeffrey E. Herrick

Received: 1 March 2011 / Accepted: 30 June 2011 / Published online: 23 July 2011

Abstract Collection of standardized assessment and monitoring data is critically important for supporting policy and management at local to continental scales. Remote sensing techniques, including image interpretation, have shown promise for collecting plant community composition and ground cover data efficiently. More work needs to be done, however, evaluating whether these techniques are sufficiently feasible, cost-effective, and repeatable to be applied in large programs. The goal of this study was to design and test an image-interpretation approach for collecting plant community composition and ground cover data appropriate for local and continental-scale assessment and monitoring of grassland, shrubland, savanna, and pasture ecosystems. We developed a geographic information system image-interpretation tool that uses points classified by experts to calibrate observers, including point-by-point training and quantitative quality control limits. To test this approach, field data and high-resolution imagery (~3 cm ground sampling distance) were collected concurrently at 54 plots located around the USA. Seven observers with little prior experience used the system to classify 300 points in each plot into ten cover types (grass, shrub, soil, etc.). Good agreement among observers was achieved, with little detectable bias and low variability among observers (coefficient of variation in most plots <0.5). There was a predictable relationship between field and image-interpreter data (R^2>0.9), suggesting regression-based adjustments can be used to relate image and field data. This approach could extend the utility of expensive-to-collect field data by allowing it to serve as a validation data source for data collected via image interpretation.

Keywords Remote sensing · Image interpretation · Aerial photography · Repeatability · Assessment and monitoring · Large-scale

Introduction

Collection of standardized assessment and monitoring data is critically important for supporting policy and management at local to continental scales (NRC 1994). Of particular need are programs that collect data for indicators closely linked to key ecosystem services and sensitive to changes in land use and climate (Feld et al. 2010; Parr et al. 2003). Some of the most fundamental indicators relevant to ecosystem services in grassland, shrubland, savanna, and pasture ecosystems include ground cover (vegetation, rock,

Electronic supplementary material The online version of this article (doi:10.1007/s10661-011-2224-2) contains supplementary material, which is available to authorized users.

M. C. Duniway (✉) · J. W. Karl · S. Schrader ·
N. Baquera · J. E. Herrick
Jornada Experimental Range, United States Department of Agriculture-Agricultural Research Service (USDA-ARS),
P.O. Box 30003, MSC 3JER,
Las Cruces, NM 88003-8003, USA
e-mail: mduniway@usgs.gov

and litter cover) and vegetation community composition (NRC 1994). However, broad-scale data on such indicators are generally sparse for such ecosystem types that are not intensively managed. This low sampling density occurs because measuring these indicators in the field is time-intensive and expensive, especially in remote areas (Elzinga et al. 1998, Holthausen et al. 2005), and broad-scale remote sensing approaches typically cannot yet produce consistent measurements with the required accuracy and precision for long-term monitoring (Marsett et al. 2006).

Data collection methods used for assessment and monitoring programs need to be feasible, cost-effective, and repeatable (House et al. 1998). Considerations for feasibility include availability of sufficient equipment and personnel with the required skills to complete the sampling in the required time. Using cost-effective methods increases the likelihood that data can be collected on a sufficient number of plots to answer relevant questions. Evaluation of cost-effectiveness should account for equipment expenses, travel (if necessary), hours of labor, and the level of skill required (which affects hourly labor costs). Consistent and repeatable methods need to be employed because (1) extensive programs necessarily rely on data collection by many different observers at one point in time and across time periods, and (2) method consistency and repeatability affect accuracy and precision of estimates.

National-scale programs in the USA that collect standardized plant community composition and ground cover data currently rely on field collection methods, including the National Resource Inventory (NRI) program (Herrick et al. 2010; USDA-NRC 2010) and planned monitoring initiatives by the Bureau of Land Management (BLM; Mackinnon et al. 2011). These agencies use field methods, primarily line-point-intercept (Herrick et al. 2005), because they are known to be feasible to complete at national levels and, with proper calibration, potentially repeatable by multiple observers (maximum difference in line-point indicators collected by multiple observers of 5% absolute cover; see Appendix E-Quality Assurance Calibration for Rangeland Quantitative Protocols in USDA-NRCS 2007). However, field visits are, by their nature, expensive due to travel costs.

Remote sensing techniques (i.e., biophysical or statistical models relating light reflectance to ecosystem attributes) have shown promise for measuring plant community composition and ground cover efficiently (Booth and Tueller 2003; Hunt et al. 2003; Karl 2010; Laliberte et al. 2010) but currently are not applied in existing national-scale surveys (e.g., NRI). For such techniques to be used in existing or future national-scale surveys, it is necessary that they meet the criteria outlined above (i.e., feasible, cost-effective, and repeatable). If a technique was developed that meets these criteria, it would also likely be appropriate for local to regional monitoring and assessment programs. Feasibility considerations include whether imagery of sufficiently high spatial resolution can be collected at the scale of the survey and whether there is sufficient expertise available to complete the required image analysis. For example, it has been demonstrated that vegetation cover data can be successfully collected using very-high-resolution imagery [~1 mm ground sampling distance (GSD); Booth and Cox 2009]. However, collection of this scale of imagery requires equipment (e.g., Booth and Cox 2009) not commonly available from aerial imagery contractors. Similarly, object-oriented image analysis techniques have shown much promise for collecting accurate vegetation cover data (Karl 2010; Laliberte et al. 2004; Laliberte et al. 2010) but many agencies that collect assessment and monitoring data currently lack the software and expertise to conduct such analyses. An analysis of cost-effectiveness for remotely sensed data should include evaluation of imagery costs and costs of collecting the indicator data from the images, either through image interpretation (manual classification by a person) or image analysis (automated classification by a computer; Laliberte et al. 2010). Interpretation can often be done by individuals with minimal training while analysis typically requires a remote sensing specialist. Thus, if the time and imagery requirements are similar, image interpretation can be more cost-effective since it can be completed by entry-level personnel with little training (Booth and Cox 2008). Evaluation of the repeatability of remote sensing techniques should consider whether the technique can produce similar measurements when applied multiple times to an area that has not changed. For image interpretation, this includes evaluating consistency of results from multiple observers at one point in time and through time. Making measurements of ecosystem indicators via remote sensing that are comparable to field-based measurements can help address many of these criteria.

Several studies have demonstrated that equivalent measurements to standard field collection methods can be made using remote sensing techniques, either through image analysis (e.g., Laliberte et al. 2010; Luscier et al. 2006) or image interpretation (e.g., Booth and Cox 2009). These approaches have the advantage of potentially allowing combination of remotely sensed data with current and legacy field data in analyses—a key requirement for incorporating remotely sensed data in existing (e.g., NRI) and planned (e.g., BLM) national surveys. Predictable relationships between indicators collected with remote sensing and those collected via field methods can partially address the repeatability criterion by allowing data collected using different sensors and techniques to be adjusted or calibrated based on field data. Furthermore, if relationships established through such calibration on a limited set of images can be reliably extended to other images without requiring additional field data, large gains in sampling efficiency may be possible.

There has been little research on the repeatability of image-interpretation methods for measuring plant community composition and ground cover by multiple observers. In this context, repeatability is a function of the degree of correspondence between indicator measurements from independent observers (i.e., different observers should produce similar results when using repeatable methods). Repeatability should be achievable through proper documentation of the method and observer training and calibration. For image interpretation, training includes teaching observers to look for features such as tone, color, texture, pattern, context, shape, and size (Morgan et al. 2010). The calibration process applies these concepts to a specific system to classify image features into predefined types in a consistent way. Previous work evaluating the repeatability of image interpretation with experienced but un-calibrated observers suggests that measurements made by each observer need to be adjusted with field data separately (Booth et al. 2005; Fensham and Fairfax 2007). To evaluate whether an image-interpretation observer is sufficiently calibrated, a standard or reference is needed for calibration evaluation. Field data could be used for this purpose, however, unless the imagery is in very-high-resolution (~1 mm ground sampling distance, e.g., Booth and Cox 2008), there will likely be a scale mismatch between the image resolution (pixel size) and the size of area measured by typical field methods for collecting cover (e.g., 1 mm diameter; Herrick et al. 2005). An alternative is to use image-interpretation classifications done by an expert (a person trained in image interpretation and familiar with the plant community) as the standard. This approach has an additional advantage of allowing for point-by-point training on correct classifications. If image interpretation is to be used in national-level surveys that collect fine-scale data on plant community composition and ground cover, more work is needed to understand the effects of observer training and calibration on the ability to derive repeatable measurements from high-resolution aerial imagery.

The goal of this study was to test a system for measuring plant community composition and ground cover from image interpretation that is applicable to national-scale surveys of grassland, savanna, and pasture ecosystems (hereafter referred to as grazing lands). As discussed previously, for the approach to be feasible for such large surveys, the imagery needs to be obtained using aerial imagery equipment available from vendors around the USA. Therefore, we did not use the highest resolution achievable (such as used by Booth and Cox 2009) but sought to use the highest-resolution imagery attainable with standard digital mapping cameras. The specific objectives of this study were to (1) develop a method for calibrating image-interpreter observers that is feasible for national-scale surveys, and (2) test if cover measurements collected through interpretation of imagery obtained with standard digital mapping cameras are repeatable by numerous observers calibrated using the methods developed. An additional key criterion for adoption of an image-interpretation approach by existing surveys (e.g., NRI) is that plot-level estimates derived from an image-interpretation approach need to be highly correlated with plot-level estimates from standard field collection methods. Therefore, an another objective was to (3) test if there is a predictable relationship between plot-level estimates derived using image interpretation and those obtained using field methods employed by such surveys. Finally, we discuss considerations for feasibility, repeatability, and, though not directly addressed by the study, cost-effectiveness of the proposed approach.

Materials and methods

Study locations

To test the applicability of this approach nationally, we selected six study sites from across the USA that represented a broad cross-section of grazing lands and where local collaborators were available to assist with plot selection and field data collection (Fig. 1a; Online Resource Table 1). Within the study areas, we used a nested approach both to maximize efficiency of field and aerial data collection and to simulate the NRI framework described by Nusser and Goebel (1997). Within each study site, we selected three 805×805-m areas (160 acres, referred to as "segments" in NRI; see Nusser and Goebel 1997) and three 50×50-m plots within each segment for field and image-interpretation data collection (i.e., data plots; Fig. 1b). An effort was made to select plots that captured the range in variability of plant community composition and ground cover in the region and to maximize the variability among segments and plots (Online Resource Fig. 1). Segment and plot locations were selected using available imagery and through consultation with local collaborators. For image-interpretation calibration purposes, an additional plot was selected adjacent to each data plot (i.e., calibration plots; Fig. 1b). A total of 54 data and 54 calibration plots were included in this study (i.e., six study areas, three segments per study area, and three data and three calibration plots per segment). The goal of this study was to assess repeatability and accuracy of the developed image-interpretation approach; therefore the sample design was necessarily different than a design intended for resource monitoring or assessment.

Image acquisition

For each study area, we acquired color-infrared aerial imagery (16-bit-depth; red, green, blue, and near-infrared spectral bands) at the highest resolution possible using a large-format, digital mapping camera. Images were collected using an UltraCamX (Vexcel Imaging; Graz, Austria) from approximately 305 m (1,000 feet) above ground level. Flying at this altitude with this sensor yielded a GSD of 2 to 3 cm and an image with a ground field of view of approximately 210 by 330 m. Imagery was collected within 2 h of solar noon to minimize shadows. Dates of image acquisition were timed to correspond to the growing season in each study area (Online Resource Table 2). All image acquisition, georeferencing, and orthorectification was completed by aerial imagery and photogrammetry contractors. The stated horizontal accuracy of the delivered products was less than 3 m (less than 2 m for most).

Two approaches were used to orthorectify and georeference the imagery. In both approaches, orthorectification and georeferencing of the imagery was accomplished by collecting imagery with a minimum of 60% forward overlap, using image center-points [based on the airborne global positioning system (GPS) and inertial measurement unit data] and auto-generated stereo-pair tie points (no surveyed ground control points). In New Mexico (NM), Idaho (ID), Pennsylvania (PA), and Florida (FL), only single high-resolution images (2–3 cm GSD) were collected over the plots. On the same day that the 2–3 cm GSD imagery was collected, coarser-resolution (9–15 cm GSD) images were collected with the necessary overlap to develop a stereo model. These coarser images were georeferenced and orthorectified as described above and then used to georeference the high-resolution images. In study areas flown later in the project (California (CA) and Nevada (NV)), three overlapping high-resolution images were collected allowing for direct orthorectification and georeferencing of the 2–3 cm GSD images.

Field data collection

Data plots were preselected for each location and oriented with the cardinal directions. Data were collected along six 50-m transects oriented north to south and evenly spaced across the plot and entirely within the corresponding high-resolution image (Fig. 1b). Vegetation and ground cover data were collected at 1-m intervals on each transect (total of 300 points per plot) using the line-point-intercept (LPI) method (Herrick et al. 2005) with an approximately 1-mm diameter pin. All plant and ground cover intercepted by the pin lowered vertically were recorded, but only the top hit was used for analysis. Top hits were classified to match the types used in image interpretation (Table 1). Some plots were moved <300 m from the original position due to

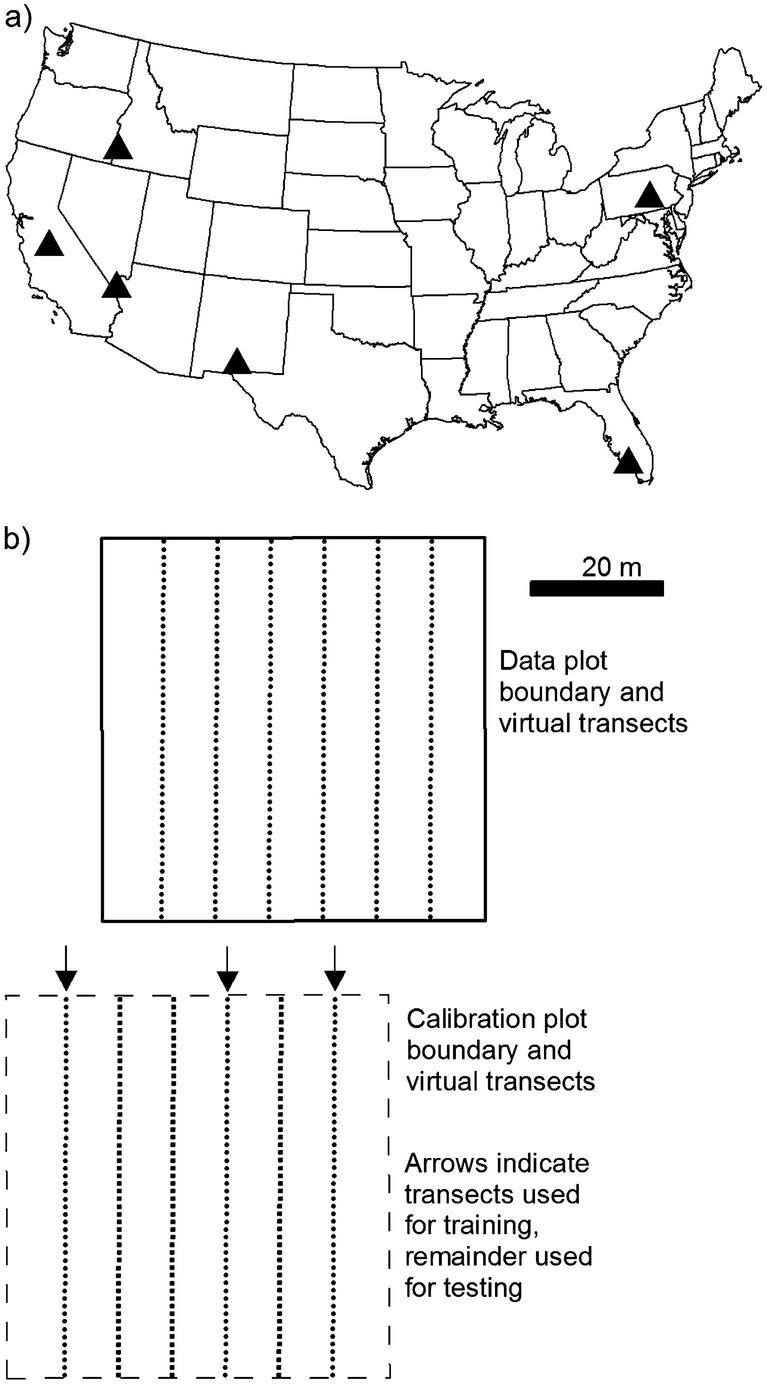
a)
b)
20 m
Data plot boundary and virtual transects
Calibration plot boundary and virtual transects
Arrows indicate transects used for training, remainder used for testing

Fig. 1 **a** Study area locations and **b** example arrangement of data and calibration plots, transects, and points

flooding (FL), obstructive presence of feeding troughs or livestock (PA), or when dense vegetation did not allow proper plot placement in the field (FL). Field data collection was completed within 2 weeks of image acquisition.

Plot corner locations were recorded in the field using a differentially corrected GPS with sub-meter accuracy (GeoXT 2005, Trimble; Sunnyvale, CA). All comparisons of field and image data (see statistical analysis) were done at the plot-level,

Table 1 Image-interpreter classification types

Fine cover types	Aggregated general cover types
Soil, litter, rock, lichen	No-canopy
Grass, forb	Herbaceous
Sub-shrub, shrub, tree, succulent	Woody

never on a point-by-point basis. Thus, although the additive geo-location error was up to 4 m in total (maximum 3 m from the imagery and 1 m from the GPS) and much larger than the image GSD (2–3 cm), this level of accuracy was sufficient for co-registering the 50×50-m plots.

Image interpretation

The image-interpretation calibration process developed for this project provided the observers (i.e., individuals collecting the point classification data) with information and training necessary to differentiate cover types and then tested their ability to classify points in a calibration plot prior to interpreting points in a data plot. The process was designed to train individuals with little or no background in image interpretation or natural sciences how to collect vegetation community composition and ground cover data in a repeatable fashion. The processes entailed developing general background and training materials as well as calibration datasets for each calibration plot that could be used to train observers on a point-by-point basis and then test that observers were sufficiently calibrated prior to data collection (Fig. 2).

Image-interpretation points for the data plots were projected along transects in the same arrangement and location as the field transects, to the extent possible, using the differentially corrected GPS locations in the ArcMap geographic information system (GIS) software (version 9.3, ESRI; Redlands, CA; Fig. 1b). No further attempt was made to precisely match the image-interpretation transects with the field transects.

To create calibration data sets for each data plot, a paired calibration plot was selected in the available area (i.e., area not in the data plot) in the high-resolution aerial images (Fig. 1b). The selected calibration plots were as similar in vegetation structure and composition to the actual data plots as possible. Virtual transects (50 m long) and points (1-m spacing) were projected on the calibration plots in the same arrangement and density as in the data collection plots, except for one plot in FL. A comparable calibration plot was not achievable at this site, so a 25×25-m plot was used and the point and transects spacing reduced by half to obtain the same number of points (300) as the other plots. Field data were not collected in the calibration plots.

The two personnel who conducted the field surveys served as our “experts” and together classified each point on all 54 calibration plots using the Image-Interpretation Tool (discussed below; Table 1; Fig. 2; Online Resource Fig. 2). While classifying the points in the three calibration plots in a segment, the experts developed description keys with distinguishing characteristics for each cover type, including color in true color and color-infrared, shape, size, texture, and pattern. The description keys were designed to provide a logical path of deduction to determine cover

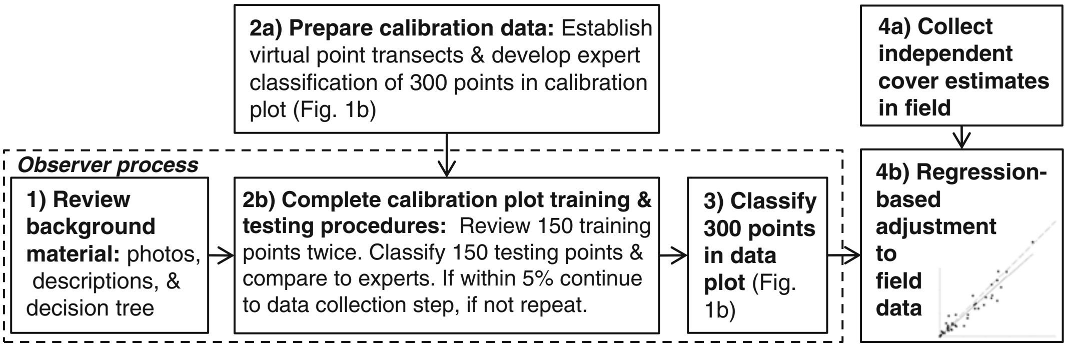
2a) Prepare calibration data: Establish virtual point transects & develop expert classification of 300 points in calibration plot (Fig. 1b)
Observer process
1) Review background material: photos, descriptions, & decision tree
2b) Complete calibration plot training & testing procedures: Review 150 training points twice. Classify 150 testing points & compare to experts. If within 5% continue to data collection step, if not repeat.
3) Classify 300 points in data plot (Fig. 1b)
4a) Collect independent cover estimates in field
4b) Regression-based adjustment to field data

Fig. 2 Image-interpretation approach developed for this study includes observer calibration (with point-by-point training and testing modules) (*2b*) and data collection (steps *1*, *2a*, and *3*), preparation of calibration data (*2a*), and regression-based adjustments of image-interpretation data to field data (*4a* and *b*)

type and to allow the observer to consistently classify each point in a plot based on observations and descriptive characteristics. Descriptive keys were developed for each plot and provided to the observers as part of the background material. After classifying each point in the three calibration plots within a segment, the experts then reviewed all their classifications in the segment to verify that they were consistent with the rules developed during the initial round of classifications.

To allow for training, testing, and data collection by multiple observers on 54 plots to be completed in as smooth and efficient manner as possible, we developed a tool for ArcMap using ESRI's ArcObjects. This Image-Interpreter Tool (IIT) has two main components: a calibration procedure (including training and testing modules) and a data collection module. The 300 points in each calibration plot are divided into two groups—training and testing points—by randomly assigning every other transect as either a training or testing transect (Fig. 1b). In calibration training mode, the tool cycles through each of the 150 training points at a set map scale (set to 1:40 for all steps) and presents the observer with an interface to select an appropriate cover type (Table 1; Online Resource Fig. 2). Incorrect decisions cause the selected button to be highlighted in red, providing immediate feedback to the observer. Correct answers cause the tool to cycle to the next sample point. In calibration testing mode, the tool performs the same tasks as in training mode without immediate feedback for incorrect classification of points. When the observer completes all calibration testing points, calibration test summary statistics (percent cover of each type) and comparisons of his or her results to the expert calibration data set are presented (Online Resource Fig. 2). The observers use the test results to determine if they are sufficiently calibrated. In this study, we required the observers to achieve less than 5% difference in cover to the experts' classification of each type before they could proceed to data collection.

IIT data were collected on each plot by seven observers working independently. All of the observers were undergraduate students at New Mexico State University (NMSU). None of the observers had any prior experience in image interpretation, and the majority had no background in GIS, remote sensing, natural sciences, or the plant communities in the study areas. We conducted an introductory session with all the observers to provide background information on the project, initial training on how to use IIT, and general guidelines for image interpretation. For each study area, we provided the observers with background material on the vegetation community including Major Land Resource Area descriptions (USDA-NRCS 2006) and oblique ground photos with important vegetation and ground cover features labeled (photos were from the field plots, but plot identification information was not provided to the observers). Each of the seven observers collected image-interpretation measurements of cover on all 54 data plots.

This original round of IIT data collection was designed to test the repeatability of this approach given the best possible circumstances (one calibration plot for each data plot, hereafter referred to as One-to-One). To test if one calibration plot could serve more than one data plot, we repeated the study with four observers (two new observers and two observers from the One-to-One approach) with the same calibration procedure as above except we used one of the data plots in each segment as a calibration plot for the other two data plots in the segment (hereafter referred to as Many-to-One). For each segment, we selected one of the plots for calibration that was the most representative of the other two plots in the segment. To generate calibration data, we simply used the majority decision from the first round of IIT data. Observers then completed the calibration procedure prior to collecting data on the remaining two data plots in the segment.

To account for the possibility that the observers' ability to accurately collect IIT data might improve with experience, the order that each observer collected data on study areas was randomized. Similarly, it was expected that within a study area and even within a segment, observers' ability to correctly classify points would increase as they became more familiar with the vegetation communities. Therefore, the order of segments within study areas and plots within segments was randomized for each observer. Additionally, image quality can vary depending on the quality of the computer monitor as well as the computer's graphics card. To control for this additional source of error, each observer was required to use the same computer for the entire project.

Statistical analysis

Because our primary interest was in the observers' ability to differentiate among general cover types, we aggregated the fine cover type data collected by the observers into three general cover types (Non-Canopy, Woody, and Herbaceous; Table 1). Analyses of the fine classes were conducted primarily to help explain the trends observed in the general cover types.

To meet our objective of assessing the ability of calibrated observers to collect IIT data in a repeatable fashion using the One-to-One approach, we conducted tests evaluating observer bias and variability between observers. For both of these tests, cover class frequencies in each plot as classified by each observer was calculated. Data was arcsine square-root-transformed to meet normality assumptions. First, we tested for observer bias using a mixed-effects analysis of variance with plots as a random effect and observer as fixed effect testing the null hypothesis that there was no effect of observer for general and fine cover categories (PROC MIXED, SAS 2001). To evaluate how variability between observers changed with plant community composition, we compared the standard deviation and coefficient of variation between observers in each plot to the field cover values for each general category. Additionally, we conducted two statistical tests to compare the level of agreement and the cover class frequency of data collected with paired calibration plots (One-to-One) and with non-paired calibration plots (Many-to-One). First, we compared the level of agreement on each point between the four observers in the Many-to-One to that of four randomly selected observers from the One-to-One approach using a nonparametric randomized block analysis of variance (ANOVA, Friedman's test blocking on plot, PROC FREQ, SAS 2001). To test the null hypothesis that the cover class frequency of the coarse categories did not differ between the two approaches, we used a paired *t* test analysis to compare the two approaches (One-to-One and Many-to-One) and treated the observers as subsamples (PROC TTEST, SAS 2001). Again, cover class frequency data was arcsine square-root-transformed to meet normality assumptions.

To evaluate the relationship between data collected using IIT with the One-to-One approach and field data collected using LPI, we used two strategies. For both strategies, analyses were done at the plot-level (not point-by-point) by averaging the 300 point classifications to estimate percent cover in the 50×50-m plot for each cover type of interest for each method. IIT observers were treated as subsamples, and the LPI data was assumed to have no measurement error. Although we acknowledge that there is measurement error associated with LPI cover estimates, evaluation of the field method was not an objective of this study. First, we used simple regressions with LPI as the independent variable and IIT as dependent variable (PROC REG, SAS 2001) for each coarse and selected fine cover types. Then, to evaluate if the regression model differed among study areas for the cover classes, we used a mixed model with IIT as our response variable, LPI as a continuous effect, study area as fixed effect (testing for differing intercepts), and a study area by LPI interaction effect (testing for different slopes; PROC MIXED, SAS 2001).

Results

Repeatability

Analyses of IIT data indicate the calibration protocol we developed can produce repeatable cover measurements among observers. There was no detectable bias between the seven observers for any of the three general cover types (Table 2). For the nine fine cover types, there was only a detectable observer effect ($p<0.05$) for the Forb cover type. Furthermore, the range in average cover over all 54 plots as measured by each observer was less than ~3% for all categories.

Evaluation of plot-level standard deviations in cover among observers for each general category as a function of field-measured cover indicate that agreement among observers was highest (i.e., low standard deviation) in plots with either high or low cover class frequency values and lowest in plots with intermediate levels of cover for a given type (Fig. 3a). Examination of points with low among-observer agreement (i.e., those where four or fewer of the seven observers agreed, ~8% of the points in the study) indicate that the fine cover types causing the most confusion were Litter, Forb, Sub-shrubs, and Succulents (Table 2). When expressed as a coefficient of variance (CV), variability among observers was highest for low cover values and generally decreased

Table 2 Observer bias analysis of variance, descriptive statistics, and relative frequency of occurrence of fine cover types in IIT points with low agreement

Category	Numerator DF	Denominator DF	*F* value	*p* Value	Average[a]	Range[b]	Freq. at low agree. points[c]
No-canopy	6	314.00	0.71	0.644	24.4	1.9	–
Litter	6	314.05	0.81	0.562	1.2	0.4	2.0
Rock	6	314.00	0.47	0.830	6.3	0.4	0.2
Soil	6	314.00	1.58	0.152	17.0	1.7	0.4
Herbaceous	6	314.00	1.05	0.396	54.2	3.0	–
Forb	6	314.02	3.50	0.002	1.9	0.9	1.5
Grass	6	314.00	1.62	0.141	52.3	3.3	0.2
Woody	6	314.00	1.57	0.156	21.4	1.7	–
Shrub	6	314.00	1.22	0.297	16.9	1.1	0.5
Sub-shrub	6	314.92	2.01	0.063	0.1	0.1	3.0
Succulent	6	314.19	0.75	0.608	0.0	0.0	2.4
Tree	6	314.01	0.50	0.806	4.4	0.9	0.5

[a] Average IIT percent cover across all plots and observers

[b] Range between the seven observers in average IIT percent cover across plots

[c] Values greater than 1 indicate cover class was selected at the low agreement points (four or fewer observers agreed, 1,396 points) by at least one observer more frequently than expected based on the overall average IIT cover for that type

with increasing cover for all three general cover types (Fig 3b). CV is directly related to the power to detect change (Conquest 1983), and in this context, a CV of 1.0 means that the standard deviation of measurements among observers was equal to the mean over all observers. Among-observer variability in CV was highest in plots with cover values of less than 20% for at least one cover type (Woody, Herbaceous, or No-Canopy). Of plots with field-measured cover less than 20%, No-Canopy IIT measurements tended to have higher CV than the Woody and Herbaceous cover types.

Comparison of repeatability of IIT data collected using paired calibration plots (One-To-One) to data collected using non-paired calibration plots (Many-To-One) showed no detectable difference in the level of point-by-point observer agreement between the two methods (p=0.62). The level of agreement was similarly high with all four observers agreeing on 77% and 80% of the points in the Many-To-One and One-To-One, respectively. Average plot cover class frequencies generated using the two methods did not differ for Herbaceous and No-Canopy but was different for Woody (p<0.05; Fig. 4), though the average difference in Woody cover between the two methods was very small (<1%).

Relationship of image interpretation to field data

For the general cover types, plot-level comparisons of LPI-estimated (field-based) cover to the average IIT-estimated (image-based) cover indicates that there was a strong relationship among all the general cover types with R^2>=0.94 and root mean square error (RMSE) <10.0% (Fig. 5). The mixed-effects model results showed a significant linear relationship between cover estimated with IIT and LPI (LPI effect; Table 3). In the Herbaceous and No-Canopy classes, the estimated slope (as indicated by the LPI by Study Area interaction) and intercept (as indicated by the Study Area effect) differed for at least one study area (Table 3). Examination of individual study area slope and intercept estimates indicate that the significant effect of Study Area (intercept) and Study Area by IIT (slope) were primarily due to differing regression models for ID and PA (data not shown; Fig. 5).

For the fine cover types analyzed, comparison of LPI-estimated to the average IIT-estimated cover indicate a strong relationship for shrubs and trees (R^2>=0.90, RMSE<=3%), but a weaker relationship for the Grass and Bare Ground classes (R^2=0.73 and RMSE=19.6% for Grass, R^2=0.72 and RMSE=13.8% for Bare Ground; Fig. 6).

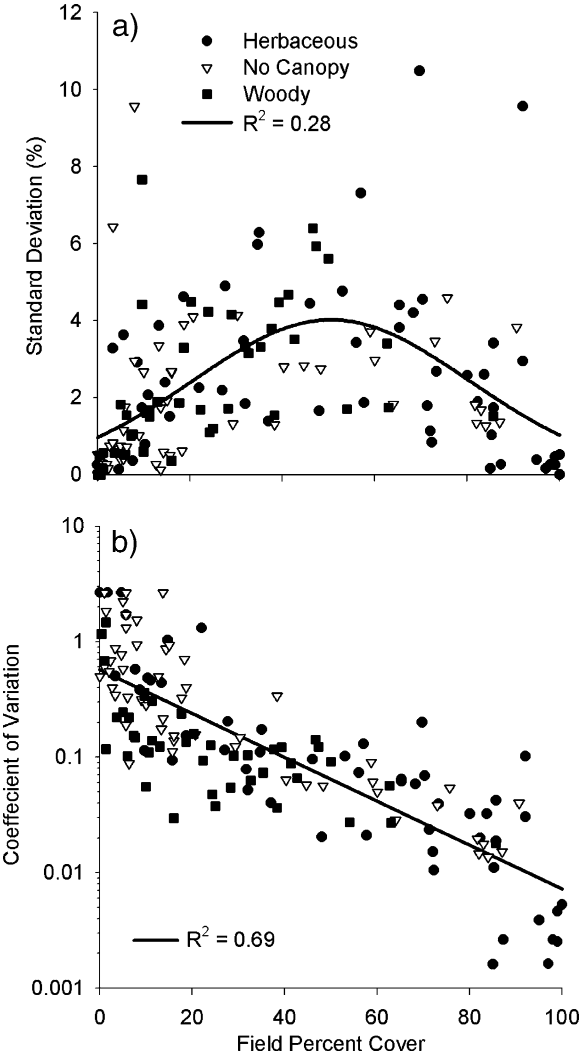

Fig. 3 **a** Standard deviation and **b** coefficient of variation in IIT percent cover among observers within a plot as a function of field cover data for the three general cover types

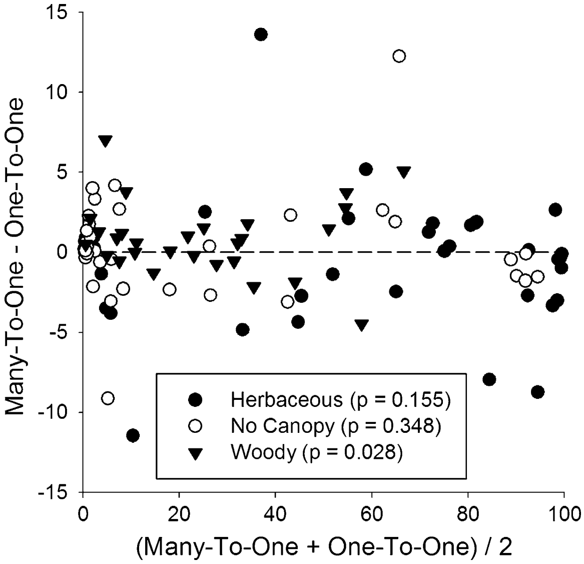

Fig. 4 Comparison of the difference between One-To-One and Many-To-One method plot cover estimates (*Y*-axis; values used are the average from four observers for each method) as a function of average plot cover (*X*-axis). *P* values are from paired *t* tests

Discussion

Our results suggest the image-interpretation approach tested is sufficiently repeatable for use in measuring cover of general cover types for broad-scale surveys. This is evidenced by low among-observer variability (Fig. 3) and almost no detectable observer bias (Table 2). We believe this approach is generally feasible for two reasons. (1) The high-resolution imagery (~3 cm GSD) used was collected with a standard digital mapping camera available from many aerial imagery contractors, and (2) image interpretation does not require expertise in image analysis and can be done with basic GIS software available to many land management and monitoring agencies. The developed calibration technique allowed collection of data by observers with no prior experience. Thus, the approach has the potential for applicability at national scales by land management (e.g., BLM) and monitoring agencies (e.g., NRCS) whose staff commonly have access to and familiarity with basic GIS software (e.g., ArcMap) but most of whom lack expertise in and access to image analysis software. The close relationship of image interpretation to field estimates of plot cover (Fig. 5) indicate both that the IIT estimates are valid and that IIT-LPI regression equations can be used to adjust IIT data to be comparable to LPI data. This adjustment can serve two important purposes. First, it will allow IIT data and LPI data to be analyzed together (e.g., in evaluating trends). Second, by using LPI data as a reference, we can account for changes in IIT due to changes in the method or image resolution.

Ability of observers to resolve cover types

Both the repeatability and the relationship between IIT and LPI data were affected by the experts' and observers' ability to resolve different cover types. Low resolvability

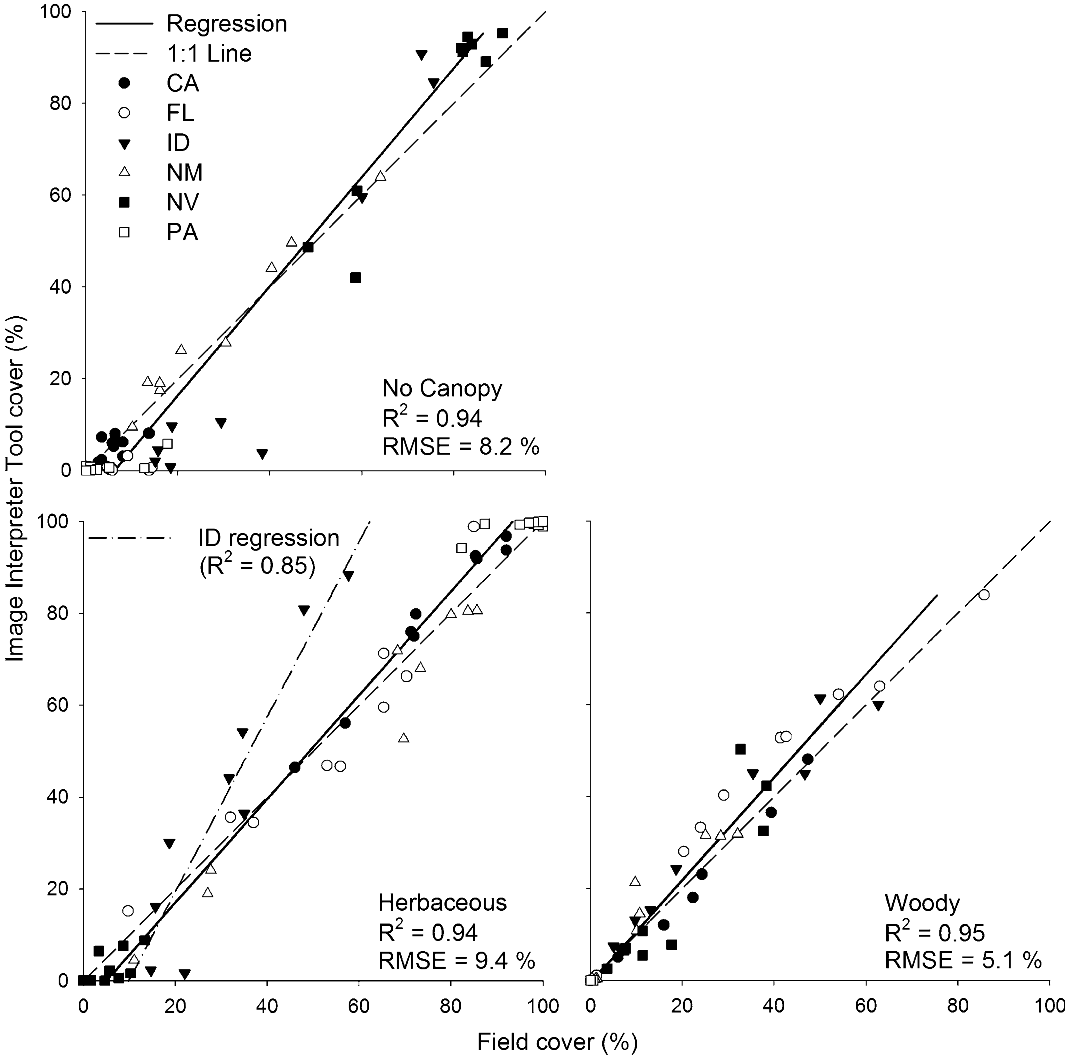

Fig. 5 Results of regressions of field cover data (line-point-intercept) and data collected using image interpretation (average of all seven observers) for the general cover types (*CA*, California; *FL*, Florida; *ID*, Idaho; *NM*, New Mexico; *NV*, Nevada; *PA*, Pennsylvania)

Table 3 Mixed-model results in general cover types

Type	Effect	Numerator DF	Denominator DF	*F* value	*p* Value
No-canopy	LPI	1	42	24.6	<0.001
	Study area	5	42	8.8	<0.001
	LPI by study area	5	42	5.7	<0.001
Herbaceous	LPI	1	42	69.6	<0.001
	Study area	5	42	3.5	0.010
	LPI by study area	5	42	7.9	<0.001
Woody[a]	LPI	1	35	417.7	<0.001
	Study area	4	35	2.4	0.071
	LPI by study area	4	35	0.9	0.499

[a] No Woody species were recorded in PA

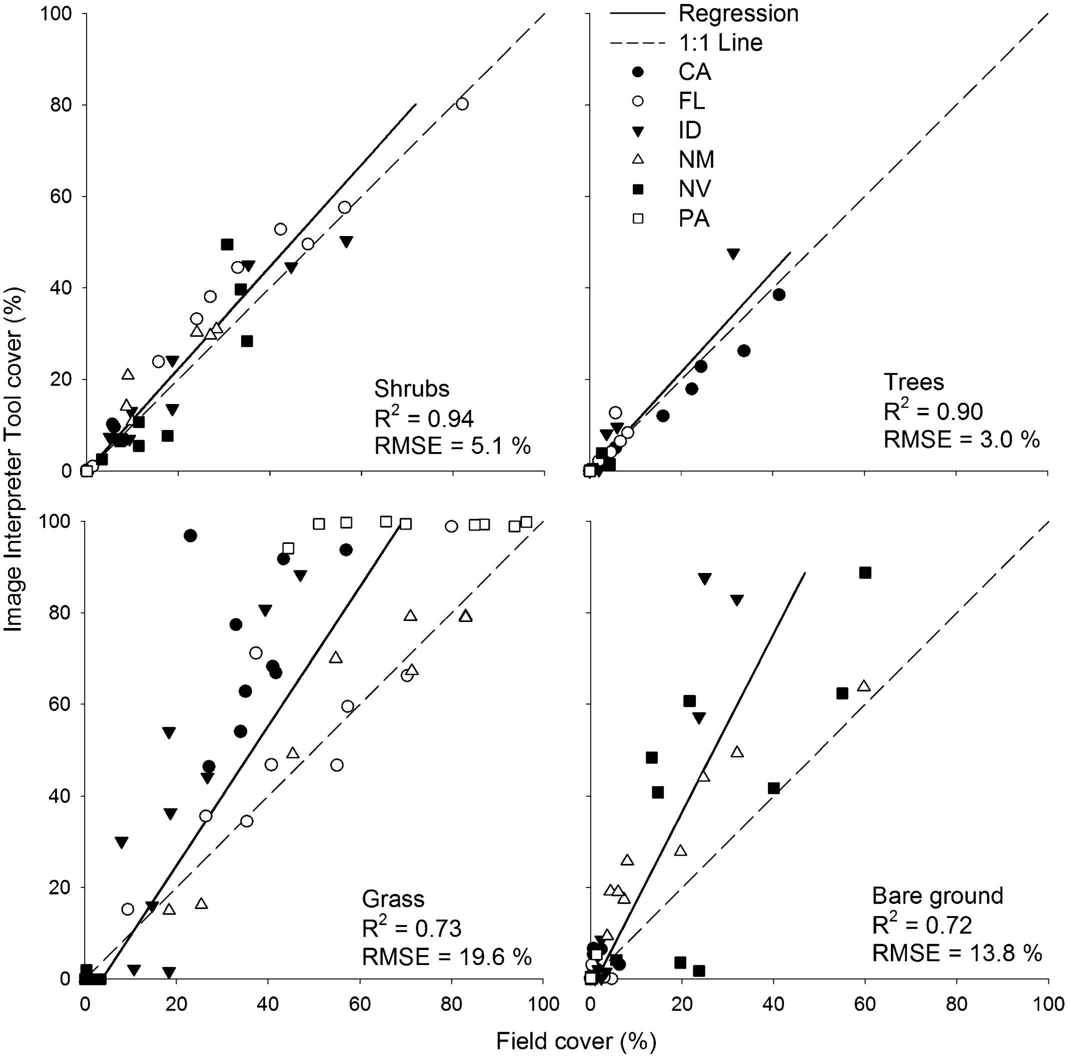

Fig. 6 Results of regressions of field cover data (line-point-intercept) and data collected using image interpretation (average of seven observers) for select fine cover categories (*CA*, California; *FL*, Florida; *ID*, Idaho; *NM*, New Mexico; *NV*, Nevada; *PA*, Pennsylvania)

is due to two or more types having similar colors, textures, or shapes. An example of this is the distinction between litter and plant canopy. In the field, the distinction is based on whether the plant part intercepted is rooted (Herrick et al. 2005), a distinction not always possible viewing an image and which likely contributed to the low resolvability in litter (Table 2). This difficulty added to the confusion in ID plots where there was a high cover of annual grasses (primarily *Bromus tectorum*). Experts and observers likely misclassified litter from annual grasses as grass, resulting in an overestimation of Herbaceous and Grass and underestimation of No-Canopy (Figs. 5 and 6). Accurate expert classification of the calibration plots is critical because the expert's classification and general interpretation of the image are subsequently used to develop the material used to orient and train the observer in the study area. Thus, if an expert misclassifies a cover type or is not consistent in their definition of cover types, then this error or uncertainty can be passed on to the observers and impair the ability to derive good cover estimates from image interpretation. Also, classes that are rare or typically occur in small patches can also be difficult to resolve (Forbs, Sub-Shrubs, and Succulents; Fig. 6; Table 2). Forbs and Grasses were indistinguishable in CA and PA imagery, so all points with non-woody canopy were classified as Grass, resulting in an overestimation of grass cover (Fig. 6).

Low resolvability may also explain cases where there was higher-than-average plot variability among observers, especially at low cover values. There were three plots with higher-than-average among-observer variability that account for the six points that are relative outliers in Fig. 3a. One plot in NM had very high standard deviations in Herbaceous (σ=10.8%) and Woody (σ=7.8%) covers. Herbaceous cover at this plot was a mix of both green and senescent vegetation. Senescent herbaceous plants near the margin of a woody patch appeared similar to the stem material of a woody species (in both color-infrared and true color), thus making it difficult to correctly classify. Two plots in CA also had high among-observer standard deviations for the No-Canopy class (σ=6.3 and 9.5%) at low cover levels (2% and 4%, respectively). For both of these plots, there was also a high variability in Herbaceous (σ= 7.1 and 9.5%, respectively) indicating that observers were having difficulty distinguishing between the two categories. In CA, this was likely due to the spatial distribution of Herbaceous and No-Canopy. In these annual grasslands, non-canopy patches were very small (most <20 cm in diameter), and litter was present in with both Herbaceous and Non-Canopy patches, making consistent classification difficult.

There also is inherent scale mismatch between LPI, which is measured using a ~1 mm diameter point and has a very high resolvability, and the IIT-based method of measuring cover using images with a pixel resolution of ~3 cm. Although IIT experts and observers were tasked with classifying points, in reality, classifications were not done on individual pixels but rather likely an area at least three pixels (6–9 cm) in diameter based on the fixed zoom level (map scale 1:40). This scale difference makes it very difficult to classify objects smaller than ~10 cm in a similar manner to what is done with LPI such as rocks (which can be as small as 5 mm in the LPI protocol), sparse or small vegetation patches, and disperse litter. Presence of small or diffuse objects likely contributed to the three plots in NV and ID that had much higher Bare Ground cover as estimated with IIT (40–85%) than measured with LPI (10–30%; Fig. 6). In both NV and ID, the problem plots were in the same segment and had a high LPI cover of rocks, diffusely arranged litter, or both.

Application considerations

The approach outlined here was designed to facilitate repeatable collection of monitoring data with remotely sensed imagery in a way that was compatible with data collected in the field using standard methods. This protocol allows remote sensing to supplement field data such that the total number of plots sampled can be increased without increasing field data collection time and/or the number of plots visited by field crews can be reduced. To apply this approach, however, there are several factors that need to be considered to appropriately balance cost of data collection with data precision and accuracy.

A primary consideration for minimizing costs for our approach is determining how many field-measured plots are required for validation of the estimates. Because IIT-derived cover estimates tend to either over- or under-value field-estimated cover (Figs. 5 and 6), field data are necessary to develop regression models to adjust IIT cover estimates to be compatible with LPI data. Questions related to this issue include how many field plots are needed, how to spatially allocate those plots, and what types of plant communities can be grouped together in the regression models. Increasing the number of plots where LPI data are collected will increase the accuracy of the regression parameter estimates only to a point, and then additional field measurements will not significantly improve the model. Also, results from this study indicate that, for some indicators (e.g., general cover types), regression parameters are similar among many vegetation types (Fig. 5), while others that are more difficult to resolve might require more community-specific regressions (Fig. 6). Another consideration, though not one addressed in this study, is whether data collected across years can be combined to develop regression models.

Another important consideration in implementing image interpretation for collecting monitoring or assessment data is the selection of sample locations and their spatial distribution within the study area. Because of the current high costs of acquiring and assembling continuous high-resolution imagery for large landscapes, it is most likely to be used in a sampling manner (i.e., locations selected for measurement from some larger area or population). If inferences are to be drawn directly from IIT-derived estimates, then such sampling locations should be selected following some statistically valid (i.e., probability-based or systematic) procedure (see Thompson 1992). Clustering of sample locations may help reduce image acquisition costs for large landscapes, and survey designs such as two-stage sampling may be useful and still provide unbiased indicator estimates (Elzinga et al. 1998). If IIT data are intended to construct a statistical model of an indicator, then probability-based selection of sampling locations is not as critical as capturing the full range of indicator variability within the study area in an unbiased manner (Brus and de Gruijter 1997; Stevens 2006).

Another way to minimize costs is to reduce the number of IIT calibration plots that are needed. Development of IIT calibration data sets was very time-consuming (taking on average 2–3 h per segment for two people). This work needs to be done by experts capable of identifying plants in the study area and familiar with interpreting aerial imagery. Therefore, it would be desirable to minimize the number of IIT calibration plots needed by using only one or two IIT calibration plots for many similar IIT data collection plots. Our results suggest that calibration plots can be applied to data plots that were basically similar in types of vegetation but are fairly different in total cover without sacrificing repeatability or accuracy (Fig. 4). Further work is needed to determine how different the calibration plots can be in terms of plant community composition and if calibration data acquired from a different season, year, or imaging sensor can be used. Another very important aspect of the approach that was not addressed is the need for a method to check the consistency and accuracy of the experts' classification of the calibration plots. The focus of this work was on training the observers to collect IIT data. This additional quality control step of the experts' classifications needs to be developed before the approach discussed here can be implemented with confidence.

In contrast to developing the expert data sets, the collection of IIT data by the observers was typically very fast (median, 20 min; range, 10 to 50 min to train, test, and collect data on one plot) and, as shown by the results of this study, can be collected by persons with little previous experience. Increasing the number of individuals who collect IIT observations on a plot can increase the precision of the plot's cover estimate. The relationship between plot cover and repeatability, however, indicate that the level of repeatability (as estimated by the standard deviation among observers; Fig. 3) varies with community composition. This relationship could guide decisions for when multiple observations on an individual plot are necessary and, if so, how many observations are needed for a given level of precision. For example, in plots that are >80% cover in one class, such as those in PA (>80% Herbaceous) and several in NV (>80% No-Canopy), one observer might suffice. In plots that are more diverse, with cover more evenly distributed among cover classes of interest, several observations on each plot could be appropriate. More work is needed, however, to evaluate how broadly this diversity-repeatability relationship can be applied.

A final consideration, though not one directly addressed by this study, is the scale of imagery required. We sought to obtain the highest-resolution imagery that was commonly attainable from commercial aerial imagery providers using standard digital mapping cameras. Other work has found significant differences between cover measurements obtained with very-high-resolution imagery (1 and 2 mm GSD) and imagery comparable to that used in this study (13 and 21 mm; Booth and Cox 2009). While our results demonstrated good resolvability for some cover types, poor relationships between field and IIT for grass and bare ground cover types (Fig. 6) support Booth and Cox's (2009) conclusion that very-high-resolution imagery is preferred for measuring vegetative cover. Taken together, our results and these previous studies suggest: (1) Reliable estimates of cover can be derived from interpretation of imagery with ~3 cm resolution for general cover types in most situations; (2) the precision of cover estimates can improve as pixel size becomes smaller, and (3) very-high-resolution imagery may be necessary for precise estimates of cover for fine cover types. Two practical implications of these conclusions are that the highest-resolution imagery attainable should be used to measure canopy cover via image interpretation, and that the resolution used should be included in accompanying metadata. Because image-interpreter cover measurements can be sensitive to image resolution (Booth and Cox 2009), it is important to have a standard against which to validate or adjust image-interpretation cover measurements used in monitoring programs. The results from this study suggest continued collection of field data at select plots and development of image-field regression equations could be used to account for effects of changing image resolution.

Conclusion

The intent of this study was to develop and test an approach for measuring plant community composition and ground cover with image interpretation that is applicable to national-scale surveys of grazing lands. To meet this objective, we developed an approach within GIS that uses expert point classifications to calibrate observers, including point-by-point training and quantitative quality control limits and then related image-derived cover estimates to field-based estimates using regression-based adjustments. We demonstrated that, through this approach, novice observers can derive repeatable estimates of some important ecosystem indicators that have a predictable relationship with field-based estimates in many ecosystems. Such a system extends the utility of expensive-to-collect field data by allowing them to serve as a validation data source for image interpretation. A training and calibration system like the one we described above is critical if high-resolution, remotely sensed data are to be used in national-level surveys that collect fine-scale data on plant community composition and ground cover. While our research demonstrates that image interpretation using multiple observers can be a viable technique for estimating ecosystem properties, more research is needed on factors affecting accuracy of cover measurements from aerial imagery, techniques for minimizing among-observer variability and bias, and efficient training, calibration, and data collection protocols.

High-resolution earth imagery is becoming increasingly available due to new digital airborne sensors, both piloted and unmanned, increased access to satellite imagery, and scanning of historical photos. Additionally, such imagery sources are becoming more easily accessible through Web services such as Google Earth. We have likely reached a point where the amount of accessible imagery exceeds the availability of experts to turn those images into indicators of ecosystem processes, either through image interpretation or image analysis. Novel procedures are needed that collect reliable data with a minimal time requirement by experts. Results presented here indicate the image-interpretation system developed could help fill that resource gap by transferring expert knowledge of plant communities to non-experts such that data on key indicators can be collected on a large number of images in a smooth and efficient manner.

Acknowledgments This study was supported by the USDA-ARS, USDA-NRCS Conservation Effects Assessment Project (CEAP), and the National Park Service Lake Mead National Recreation Area and Mojave Desert Network. We would like to thank Michelle Mattocks for her assistance with field data collection and expert classifications, the NMSU students and other Jornada staff who contributed to this project, and the assistance we received from our collaborators including Melvin George, Neil McDougald, Paul Garner, George Poole, Chris Roberts, Toshi Yoshida, Jeanne Taylor, Joe Chigbrow, John Wilford, Mark Murdock, Lance Kosberg, Jeff Gonet, Sarah Goslee, Frank Stoltzfus, and Lloyd Rietz. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Booth, T. D., & Cox, S. E. (2008). Image-based monitoring to measure ecological change in rangeland. *Frontiers in Ecology and the Environment, 6*(4), 185–190.

Booth, D. T., & Cox, S. E. (2009). Dual-camera, high-resolution aerial assessment of pipeline revegetation. *Environmental Monitoring and Assessment, 158*(1–4), 23–33.

Booth, T. D., & Tueller, P. T. (2003). Rangeland monitoring using remote sensing. *Arid Land Research and Management, 17*(4), 455–467.

Booth, D. T., Cox, S. E., & Johnson, D. E. (2005). Detection-threshold calibration and other factors influencing digital measurements of ground cover. *Rangeland Ecology & Management, 58*(6), 598–604.

Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma, 80*(1–2), 1–44.

Conquest, L. L. (1983). Assessing the statistical effectiveness of ecological experiments: Utility of the coefficient of variation. *International Journal of Environmental Studies, 20*, 209–221.

Elzinga, C. L., Salzer, D. W., & Willoughby, J. W. (1998). *Measuring and monitoring plant populations*. Denver, CO, US Department of the Interior, Bureau of Land Management. National Applied Resource Sciences Center. BLM/RS/ST-98/005 + 1730.

Feld, C. K., Sousa, J. P., da Silva, P. M., & Dawson, T. P. (2010). Indicators for biodiversity and ecosystem services: Towards an improved framework for ecosystems assessment. *Biodiversity and Conservation, 19*(10), 2895–2919.

Fensham, R. J., & Fairfax, R. J. (2007). Effect of photoscale, interpreter bias and land type on woody crown-over estimates from aerial photography. *Australian Journal of Botany, 55*(4), 457–463.

Herrick, J. E., Van Zee, J. W., Havstad, K. M., Burkett, L. M., & Whitford, W. G. (2005). *Monitoring Manual for Grassland, Shrubland and Savanna Ecosystems. Volume I: Quick Start*. Tucson, AZ: The University of Arizona.

Herrick, J. E., Lessard, V. C., Spaeth, K. E., et al. (2010). National ecosystem assessments supported by scientific and local knowledge. *Frontiers in Ecology and the Environment, 8*(8), 403–408.

Holthausen, R. , R. Czaplewski , D. DeLorenzo , G. Hayward , W. Kessler , P. N. Manley et al. (2005). Strategies for monitoring terrestrial animals and habitats. Gen. Tech. Rep. RMRS-GTR-161. Fort Collins, CO US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 34 p.

House, C. C., Goebel, J. J., Schreuder, H. T., Geissler, P. H., Williams, W. R., & Olsen, A. R. (1998). Prototyping a vision for inter-agency terrestrial inventory and monitoring: A statistical perspective. *Environmental Monitoring and Assessment, 51*(1–2), 451–463.

Hunt, E. R., Jr., Everitt, J. H., Ritchie, J. C., et al. (2003). Applications and research using remote sensing for rangeland management. *Photogrammetric Engineering and Remote Sensing, 69*(6), 675–693.

Karl, J. W. (2010). Spatial predictions of cover attributes of rangeland ecosystems using regression kriging and remote sensing. *Rangeland Ecology and Management, 63*(3), 335–349.

Laliberte, A. S., Rango, A., Havstad, K. M., et al. (2004). Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern New Mexico. *Remote Sensing of Environment, 93*(1–2), 198–210.

Laliberte, A. S., Herrick, J. E., Rango, A., & Winters, C. (2010). Acquisition, orthorectification, and object-based classification of unmanned aerial vehicle (UAV) imagery for rangeland monitoring. *Photogrammetric Engineering and Remote Sensing, 76*(6), 661–672.

Luscier, J. D., Thompson, W. L., Wilson, J. M., Gorham, B. E., & Dragut, L. D. (2006). Using digital photographs and object-based image analysis to estimate percent ground cover in vegetation plots. *Frontiers in Ecology and the Environment, 4*(8), 408–413.

MacKinnon, W. C., Toevs, G., Spurrier, C., & Taylor, J. J. (2011). *BLM Core Terrestrial Indicators and Methods*. Washington, D.C: Bureau of Land Management.

Marsett, R. C., Qi, J. G., Heilman, P., et al. (2006). Remote sensing for grassland management in the arid Southwest. *Rangeland Ecology & Management, 59*(5), 530–540.

Morgan, J. L., Gergel, S. E., & Coops, N. C. (2010). Aerial photography: A rapidly evolving tool for ecological management. *Bioscience, 60*(1), 47–59.

National Research Council (NRC). (1994). *Rangeland Health: New Ways to Classify, Inventory and Monitor Rangelands*. Washington: National Academy Press.

Nusser, S. M., & Goebel, J. J. (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics, 4* (3), 181–204.

Parr, T. W., Sier, A. R. J., Battarbee, R. W., Mackay, A., & Burgess, J. (2003). Detecting environmental change: Science and society—Perspectives on long-term research and monitoring in the 21st century. [Proceedings Paper]. *Science of the Total Environment, 310*(1–3), 1–8.

Stevens, D. L., Jr. (2006). *Spatial properties of design-based versus model-based approaches to environmental sampling*. 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal.

Thompson, S. K. (1992). *Sampling*. New York: John Wiley and Sons, Inc.

United States Department of Agriculture- Natural Resources Conservation Service (USDA-NRCS) (2006). Major land resource area (MLRA). Land resource regions and major land resource areas of the United States, the Caribbean, and the Pacific Basin. U.S. Department of Agriculture Handbook 296. http://soils.usda.gov/survey/geography/mlra/. Accessed 12 November 2010.

United States Department of Agriculture- Natural Resources Conservation Service (USDA-NRCS) (2007). National Resources Inventory—Handbook of instructions for rangeland field study data collection. http://www.ncgc.nrcs.usda.gov/products/nri/range/2007range.html. Accessed 12 November, 2010.

United States Department of Agriculture- Natural Resources Conservation Service (USDA-NRCS) (2010). National Resources Inventory—Rangeland resource assessment. http://www.nrcs.usda.gov/technical/nri/rangeland/. Accessed 12 November 2010.