Methodology for Assessing the Utility of Existing Data for Vegetation Mapping

National Park Service Vegetation Inventory

October 2009

Prepared For:

United States Department of Interior
National Park Service

1996 version prepared by:

The Nature Conservancy
1815 N. Lynn St.
Arlington, VA 22209

2009 version edited by:

National Park Service

1201 Oakridge Drive
Fort Collins, CO 80525

7333 West Jefferson Avenue
Lakewood, CO 80235

Table of Contents

Section **Page**

List of Tables & Figures

I. Introduction

Vegetation data have been previously gathered in many National Parks in which vegetation mapping projects will occur. The use of existing data in National Park Service Vegetation Inventory can offer great cost savings. These data have potential uses in a vegetation mapping project, but may be limited by a variety of factors, including accessibility of the data, the nature of the data itself, and how the data meets the objectives for vegetation mapping in the National Parks. In order to distinguish which data should be used in a mapping project and for what purpose, a standardized methodology is needed that can be applied to evaluating data in all parks. This paper proposes such a methodology based upon the identification and assessment of seven key components of data quality for each existing data set. The methodology results in a categorization of each data set according to its utility for vegetation mapping. These standards were developed for and implemented in an assessment of existing data for Great Smoky Mountains National Park.

The first step in the mapping of any National Park should be to locate and review pertinent literature and other materials associated with past vegetation studies in the park unit, such as maps, original field forms, unpublished reports, and data summaries. Information about each data set's content, format, location, availability and how it could be used in the vegetation mapping project should be documented.

Vegetation data can be gathered for different purposes, each of which can employ varying methodologies. The purposes and methodologies for data collection for other vegetation studies are often not compatible with the standards that have been developed for the purpose of standardized mapping across all National Parks. Some of the data sets may be useful for the characterization of a vegetation type, but not for the assignment of a classification attributes to map polygons. Other data sets may require significant effort to locate and convert to a compatible form such that its use would not be cost-effective. Data also varies in its intrinsic quality and the amount of information it carries; some data may be useful, but only minimally so, or with low confidence. A decision about the utility of a given data set may therefore involve a cost/benefit analysis, and this analysis may be affected by consideration of other available data sets.

In order to be useful for a vegetation mapping project a data set must, at a minimum, represent existing vegetation and contain enough structural and compositional information to place the sample within some level of the National Vegetation Classification hierarchy (Federal Geographic Data Committee 2008). Vegetation samples meeting these minimum criteria can be used to refine the vegetation classification, document the variation in vegetation types, and to develop park-specific vegetation characterizations. To aid in polygon attribution or map accuracy assessment, the vegetation samples must, in addition to the minimum criteria, be geographically referenced with sufficient accuracy for vegetation mapping.

The methodology outlined below requires that the data assessor become familiar with key data

quality factors which affect the utility of data in various applications. This assessment requires unrestricted access to each data set. Each data set will need to be assessed for each data quality factor, as described below. The assessment will be used to determine the potential value of each data set for the purposes of the vegetation mapping project.

II. A Methodology For Evaluating Existing Data Sets

The evaluation of existing vegetation data sets requires an assessment of key factors that define data quality for the specific purposes of the NPS Vegetation Inventory. This results in a categorization of each data set according to utility for vegetation classification, description, and mapping. Seven key data quality factors have been identified for completing this assessment:

(1) Accessibility of the data sets,
(2) Compatibility in sampling approach and methods,
(3) Utility of data format and condition,
(4) Recognition of special value of a particular data set,
(5) Inclusion of core vegetation and environmental characteristics,
(6) Currency of the data set, and
(7) Adequacy of geo-referencing.

Figure 1 presents a step-by-step process to evaluate existing data sets according to these seven components for their utility in vegetation mapping. Starting from the top of the figure, data sets are assessed and placed into appropriate categories of utility, which are characterized in Table 1. Each step of the process, as it is linked to the seven factors, is further described below.

Figure 1. Flow chart for evaluating existing datasets.

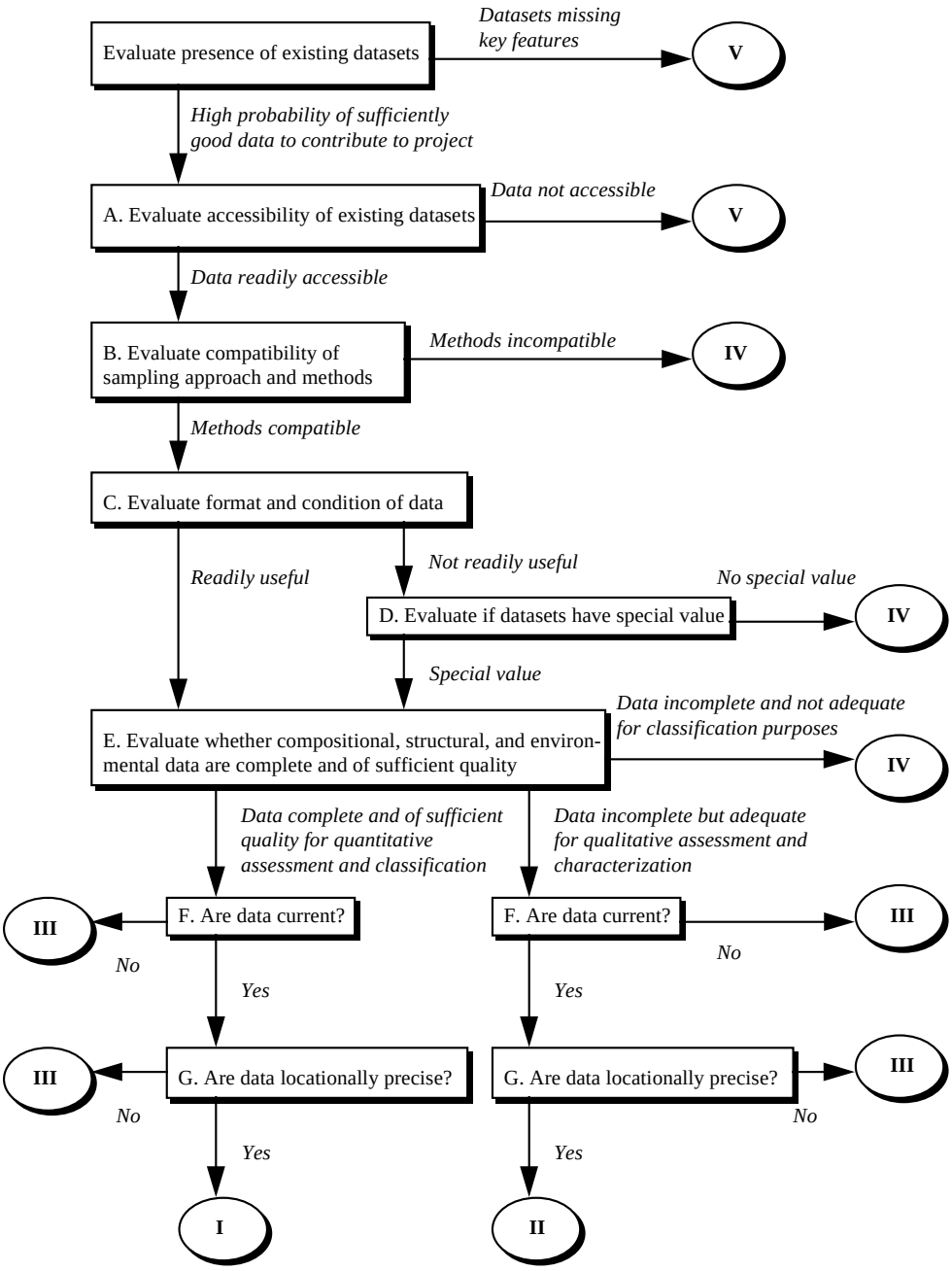

Table 1. Categories of Utility for Existing Data Sets.

Category	Description
I	Samples are adequate for classification and mapping (i.e., the data are geo-referenced, represent existing vegetation, and contain sufficient structural, compositional and site information to place the sample within the standard classification framework).
II	Data are adequate to assist in photo interpretation, photographic signature key development, or map accuracy assessment (i.e., the vegetation and site information are of lower quality, but the samples represent existing vegetation and are geo-referenced with reasonable confidence).
III	Data can be used for vegetation classification and characterization of a vegetation type within the park, but not for mapping or analysis because the sample is not adequately geo-referenced, contains inadequate detail in the vegetation information, and/or may not represent existing vegetation at the sample location.
IV	Data set was assessed and not found to be useful at any level
V	Data set not was not available for assessment.

Where existing data sets have a reasonable probability contributing to the vegetation information that is required in a vegetation mapping project at a park, the following procedures should be taken to assess the key data quality factors which will determine the utility of each data set.

A. Accessibility of Data Sets

Some data sets are readily available, either at park offices or at the offices of the academic institutions or agencies where the research was completed. Usually the data is held by the researcher who carried out the research, though sometimes it may be at the park office or held by other researchers. Data sets resulting from graduate research may be more difficult to access; this data may be held by the park, the major professor, or the student. In many cases, these data sets are more difficult to locate.

Data sets may have a sole location, or there may be duplicates. If there are multiple copies of a data set, one copy may be easier to access than another. However, it should not be assumed that the multiple versions are identical; each copy will change as it is prepared for analysis. In addition, new data and observations are often added to an existing data set.

Researchers have often moved since the data were gathered; it may or may not be easy to determine their current location and address. This is particularly so in the case of graduate students who did not continue to work in the geographic area or remain professionally in the discipline.

Even when physically accessible, the issue of 'intellectual property' must be addressed. Many scientists are reluctant to allow others to use their data. They may have plans to further analyze

and publish the data, concerns that their work will not be recognized, and generally perceive the vegetation mapping program as competitive with their intentions. This is particularly likely to be the case with researchers who have accumulated large data sets due to a long-term interest and involvement in a particular park. This should be considered in the initial development of the project; as it will usually be appropriate to formally or informally involve such researchers from the onset of the project.

In most cases, the National Park Service has legal rights to use of the data gathered on Park Service land, particularly for those project that have involved NPS funding and permitting. These rights may be worth invoking in some cases, though legal rights will often have little practical value. An uncooperative researcher will be unlikely to produce data in a useable form and a timely manner.

Researchers may be willing to provide the data, but it may be difficult, expensive, or a low priority for them to do so. The data often exists in paper files or old electronic files that need to be located and reproduced. Effective access to old data sets often will necessitate visiting the physical location of the data and assisting with its identification, sorting, and reproduction. Accessibility issues may entail a simple yes or no assessment. However, there is often a surprising complexity to this issue, and in cases where access to the data set may involve substantial effort, an assessment of the cost/benefit will be necessary. If the researcher is at considerable distance from the park and from the offices of the people working on the vegetation mapping project, this may be a significant cost of time and travel. The amount of effort needs to be balanced relative to the potential benefits, which will have to be assessed in relation to issues below (e.g., size of the data set, quality of the data, whether the data is duplicated by other data sets or would be easily replaced or improved on).

Data sets assessed as not accessible are considered to be Category V data.

B. Compatibility of Sampling Approach and Methods

In order for existing vegetation data to be useful, these methods for data collection must meet the standards of the NPS Vegetation Inventory.

Plots or other vegetation samples gathered with little attention to plot selection and placement may cause problems in interpretation, especially if the total number of plots is not very large. Plots gathered specifically for vegetation analysis are more likely to be homogeneous. Plots placed for other purposes are more likely to be heterogeneous. For instance, a plot (or other vegetation sample) placed to characterize nesting habitat of a bird species may be centered on the nest, and may include heterogeneous components of the actual vegetation types. In order to be useful for vegetation classification and characterization, samples must be homogenous in terms of floristics (species composition) and physiognomy (structure). The plot sizes must be adequate to serve as a sample of the vegetation type (i.e., capture adequate representation of species composition and vegetation layering).

Data sets compiled from incompatible methodologies are considered Category IV.

C. Compatibility of Data Format and Condition

Data in existing data sets may be in an electronic format incompatible with current formats, or may exist only as original field forms of variable and often poor legibility and interpretability. In some cases, it may be more efficient to collect new data of known quality and standards rather than access existing data, resolve questions of sampling methodology, perform quality control steps, standardize the species taxonomy, and enter it into a compatible digital format.
The amount of effort that is warranted is also dependent upon the actual size of a data set and the estimated cost of field sampling. Larger data sets offer greater economy of scale and cost-effectiveness as the quality control and standardization process is generally carried out once for the entire data set.

D. Special Value of Data Sets

All data sets (even those with data in not readily usable form) need to be assessed for any special value prior to discounting their contribution to the program. Some data sets (small or large) may be especially valuable and therefore worth the extra effort (e.g., converting data format) of incorporation into the mapping project. Examples are data sets of specialized communities unlikely to be adequately addressed as part of the park-wide sampling protocol, data sets covering remote or difficult-to-access areas, and data sets of especially high quality, comprehensiveness, or associated with other information (such as environmental data) likely to provide extraordinary insight into the communities of the park unit. Data sets with no special value and whose data are not readily usable are considered Category IV.

E. Completeness and Quality of Vegetation and Environmental Data

In order for existing vegetation data to be useful, these data must meet the information standards of the NPS Vegetation Inventory.

Species composition / abundance

Information on total floristic composition is a critical component of vegetation classification. Data sets with complete floristics and an abundance measure for each species are of the most utility for vegetation classification and characterization. At minimum, plot data should capture the dominant species in each stratum, with some indication of relative dominance. If existing data are to be quantitatively analyzed in combination with new plot data from the vegetation mapping program, abundance measures must be compatible or convertible.

Vegetation samples, especially if gathered for purposes other than vegetation classification, are sometimes based only on canopy, woody, dominant, or common species. Dominant or common species may be defined either loosely or strictly, by either cover or diameter cutoffs (variable from study to study), or by recording a set number of species in each stratum (e.g., the five most abundant species in each stratum). Such data can be useful, but their utility may be limited. It is also difficult to resolve data gathered with different criteria in a scientifically defensible and practical manner.

The problems that floristically incomplete data represent depends on the nature of the vegetation and existing level of vegetation knowledge. Canopy data may be sufficient to identify the alliance or even the association in some situations. In other situations the same level of data may not provide sufficient enough information to identify the alliance, especially if the plot size is small.

Structural information

The National Vegetation Classification (Federal Geographic Data Committee 2008) distinguishes between forests, woodlands, shrublands, dwarf shrublands, herbaceous and non vascular vegetation. A quantitative assessment of the coverage and height of each vegetation stratum is necessary to support the classification and characterization of vegetation types. Existing data sets must therefore include sufficient information on vegetation structure to allow these distinctions. Woody structure can sometimes be inferred from basal area, and canopy/subcanopy/shrub layers can often be reconstructed from diameter classes. However, inference of this sort has lower confidence than more direct measures.

Environmental data

Environmental information is extremely helpful for the interpretation and mapping of vegetation. Vegetation samples that lack critical components of environmental information may be difficult or impossible to classify. For example, hydrology can be very difficult to determine without direct observation. Other environmental information that can be very useful in vegetation classification and characterization include geology, soil, slope aspect and degree, elevation, landform, land use, and disturbance regime. In addition, information about vegetation adjacent to the sample is often helpful in placing the sample in a landscape context.

The quality of species identifications (accuracy, standardization, and level of classification).

Taxonomic standards must be maintained to develop a consistent vegetation classification. The taxonomic quality of existing data sets varies widely, due to the investigator's experience and knowledge, level of effort, changes in taxonomy over time, and the use of out-dated botanical manuals. In addition, different investigators often employ distinctive conventions in recording unknown species or unresolvable species groups. These conventions are rarely clearly recorded. Furthermore, the data in a single study may have been taken by different investigators with differing abilities.

Some data sets do not make the species-level distinctions that are important for the desired standard for vegetation classification. For instance, in maritime forests of the south Atlantic Coastal Plain (North Carolina, South Carolina, Georgia, and northern Florida) drier sites have abundant *Quercus hemispherica* and *Persea borbonia*, while wetter sites have abundant *Quercus laurifolia* and *Persea palustris*. The commonly used manuals for that area only distinguish the oaks and red-bays to the level of genus which limits the ability to differentiate between the two distinct vegetation types.

Data sets rarely have species identification quality sufficiently suspect as to warrant their elimination from consideration. More commonly, an assessment of the quality of species identification will reveal some problems requiring cautious interpretation.

Combining multiple data sets gathered by different investigators can be particularly complicated, and the resolution of taxonomic differences can be very time-consuming. Taxonomic resolution can rarely be brought to completion, and the unresolved differences can be significant. The effort put into taxonomic resolution should be calibrated to the likely affect on classification, characterization, and mapping. The taxonomic quality needs to be an additional consideration in the cost/benefit analysis for the utility of the existing data set.

Data sets with incomplete vegetation and/or environmental data such that no characterization or assessment of the type can be made are considered Category IV. Those with data that are incomplete but can be still be used qualitatively are considered either Category II or Category III, depending on the currency of the data and whether it has been adequately geo-referenced. Those data sets that are suitable for quantitative characterization and assessment are considered either Category I or Category III, again depending on the currency of the data and whether it has been adequately geo-referenced.

F. Currency of Data

Data sets need to be assessed for their currency. There are two different components of currency; the first relates to the classification of a vegetation type and the second relates to individual polygons on a map. The NPS Vegetation Inventory specifies the need to map existing vegetation,

as compared to 'climax', or 'presettlement', or 'potential natural' vegetation. The assessment of the utility of an existing data set requires a consideration of both the amount of change and the nature of that change that has occurred in a park unit since the vegetation samples were taken.

Currency for polygon attribution requires recent data that was gathered after the last major disturbance event. Major disturbance events could include fires, floods, logging, hydrologic alterations, hurricanes, and extirpation of dominant species. Some disturbances may be very localized, which could leave portions of a data set useful and other portions not useful for vegetation mapping. The data should also be gathered recently enough such that successional changes have not resulted in a different matrix of vegetation types. Successional changes occur at different time scales under different environmental and disturbance regimes. Polygons of long-lived vegetation types that are not disturbance-maintained and occur in relatively pristine landscapes may be tentatively attributed from older data.

Any kind of catastrophic disturbance or simply the passage of enough time since the data were taken render the data suspect for vegetation mapping. In most landscapes, data should be considered suspect if older than 10 years (this may be longer in other areas). If the data set is considered suspect for currency, it may be assessed on a sample by sample basis, or simply used with caution.

In some parks, especially recently established parks, changes in land use have occurred relatively recently. Examples of land use change include modification of fire regime, change or elimination of grazing, and release of successional processes following agricultural uses. In such parks where the vegetation has not stabilized, older vegetation samples should be considered particularly suspect.

A good example where caution must be employed in the use of older vegetation samples is Great Smoky Mountains National Park. Miller (1938) systematically sampled 1300 plots on a grid system throughout the park in the early 1930's. Since then, one of the most important canopy trees, American chestnut (*Castanea dentata*), was eliminated from the canopy by chestnut blight. In addition, fires have burned in many areas changing the locations and extent of fire-dependent vegetation types, grassy balds have filled in with woody growth, old fields have undergone succession, and hurricanes have caused landslides. Thus, Miller's plots cannot be used for locating existing vegetation. However, the plots might still be useful for classification and characterization of specific vegetation types. They cannot be used for the sites (particular elevation and topographic settings) where American chestnut was a significant component of the vegetation, since its loss has permanently altered the vegetation patterns and composition. Our conclusion was that this extensive data set was not worth using due to additional data quality problems concerning poor taxonomic resolution, poor geo-referencing, non sampling of the herbaceous stratum, and variable sampling protocols. Miller's data set is a Category IV, not found to be useful at any level.

Data that is too dated for polygon attribution or accuracy assessment purposes may still be of

utility for the classification and characterization of classification types. In some landscapes, the polygons may shift in response to disturbance, but the vegetation types themselves may be very stable. In such cases, older data could be used to assist in classification and characterization, with caution proportional to its age.

Current data, if it meets other requirements, may be Category I or Category II data (depending on data quality issues). Non-current data that meet other information requirements would be considered Category III.

G. Geo-referenced Data

The uses of vegetation samples for mapping and accuracy assessment require precise locational information. The determination of adequate geo-referencing is not a simple process. Many older data sets did not have the benefit of Global Positioning System (GPS) technology to help geo-reference the plots. GPS data is generally considered to have the highest accuracy in geo-referencing, and older methods have a higher level of error, although some pre-GPS or non-GPS data may actually be more accurately located than some GPS data. This is particularly true for steep landscapes or under dense vegetation canopy where GPS reading are often ineffective and inaccurate. It may also be true in landscapes where landmarks in the landscape are frequent, reliable, and closely correlated with the vegetation. For example, a vegetation plot along a narrow riparian zone may be more accurately referenced to photography and topographic maps through use of ground control points, than through conventional use of GPS technology. The required precision of geo-referencing is also landscape-dependent. Locational precision will be less critical in landscapes dominated by large vegetation polygons. Conversely, in landscapes dominated by small vegetation polygons, greater locational precision will be necessary to reference a plot to the correct polygon.

Data sets with sufficiently precise geo-referenced data can be used as Category I or Category II. Inadequately geo-referenced data with sufficient information quality may still be useful as Category III data.

III. Conclusion

A standard process to assess the utility of existing data sets for vegetation classification, characterization and mapping for the NPS Vegetation Inventory is presented (see Figure 1). Through an assessment of seven key factors, data sets can be placed into different categories of utility (see Table 1). This assessment provides the information required to assess requirements for additional data to complete an accurate vegetation map for a park. Standardized assessment of existing data will result in the most cost-effective, efficient, and accurate approach to develop the vegetation classification and maps in the National Parks.

IV. References

Federal Geographic Data Committee, Vegetation SubCommittee. 2008. National Vegetation Classification Standard, version 2. FGDC-STD-005-2008 (Version 2). Reston, VA. Available: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation.

Miller, F. H. 1938. Brief narrative descriptions of the vegetative types in the Great Smoky Mountains National Park. Unpubl. manuscript. Great Smoky Mountains National Park Archives. 10 p.