MISCELLANEOUS NOTES REGARDING NLA DATA FILES

Prepared by D.V. Peck

Updated 11/30/09

This is intended to provide users with some additional information regarding the various NLA data files. This information includes filenames of final data files, notes regarding file structure (e.g., number of records, key variables) that is important to know if you intend to merge different files, and any "quirks" specific to individual files that might affect how on selects records for analysis or uses/interprets the values of various variables.

GIS COVERAGE FILES (from Horizon Systems)

National_LakePoly_withMetrics_20090929.*

Lake polygons: lake area, lake perimeter, and shoreline development

.prj
.dbf
.sbn
.sbx
.shp
.shx

National_Basin_withMetrics_20090928.*

Basin (catchment) area

.prj
.dbf
.sbn
.sbx
.shp
.shp.xml
.shx

GIS BASIN AND BUFFER LANDUSE METRIC FILES

GIS file (basins): National_NLA_Basin_Allocations_20090928.dbf
GIS file (buffers): Lake_LU_Allocation20090929.dbf

These files were used to create the final landuse metric files., which contain both the areal estimates as well as the percent landuse variables

Lake Basin Metric File: NLA_BASIN_LU_METS_20091006
Variable names and labels: NLA_BASIN_LU_METS_INFO_20091006
Lake Buffer Metric File: NLA_BUFFER_LU_METS_20091006
Variable names and labels: NLA_BUFFER_LU_METS_INFO_20091006

Includes areal estimates of NLCD landuse classes for lake basins (catchments) and 200-m buffer area around each lake, and metrics representing the percent area of each category.

There is one record per site (1157 records), VISIT_NO is set to 1. Merge with other data with records for multiple visits using SITE_ID, rather than SITE_ID and VISIT_NO.

Two sites (0567 and 2378) were were delineated but were never sampled (too shallow and not needed). These have been dropped from the metric file but are still included in the shapefiles and attribute tables.

NLA DESIGN AND LAKE INFORMATION FILES (WED-Corvallis)

DESIGN FILES:

GIS coverages: NLA_LAKES_ALL_20090917.* (all lakes in design file). This file was prepared at WED-Corvallis from NHD Plus. There will be slight differences in values (lake areas, perimeters) from the GIS files for lake polygons provided by Horizon Systems. Values of lake area, lake perimeter, and shoreline development from the WED file are used in the Lake Information file.

.prj (use 20090904 file)
.dbf
.shp
.shp.xml
.shx

Attribute table served as the basis for the final lake information file.

Also lake polygon files (NLA_2009_ANALYSIS_POLY_20090917.*) for the set of sampled lakes.

.prj (may not be there)
.dbf
.sbx
.shp
.shx
.shp.xml
.shx

There is also a .kmz file (LAKES.KMZ) that can be used with Google Earth to locate sampled sites.

LAKE INFORMATION FILES

NLA_LAKEINFO_ALL_20091113.* (all lakes in the design file)
NLA_LAKEINFO_SAMPLED_20091113.* (only lakes that were sampled)
NLA_LAKEINFO_INFO_20091113 (variable names and labels)

There are

SAMPLED file includes all sites where SAMPLED=YES (probability sites, hand-selected sites, and some non-target sites that were sampled anyway). There are 1252 records (incuding revisits), and 1157 records if repeat visits are eliminated.

These files have one record per site visit, so much of the design information is replicated.

A set of repeat visit sites (including the original visit) can be obtained using the variable REPEAT=Y.

In the original design file, one lake was selected twice (as sites 0285 and 0365). Site o285 was sampled and 0365 was listed as Non Needed. It was later discovered that these two lakes were in fact a single lake split by a bridge. The final lake information file has the correct information for site 0285, and site 0385 should not be used for any analyses.

Site NELP-4896 may be retained in the original design files. It was sampled as lake 3846. All data are associated with the probability site (3846), and site NELP-4896 should be deleted before analysis.

Note that hand-selected lakes are identified as LAKE_SAMP=REF_Lake. **This does not imply these are "reference" lakes, only that they are candidates**. Use RT_NLA='REF' to identify lakes that passed all of the "filters" used to select reference sites for assessing condition.

<u>Merging Files for Analysis and Estimation</u>

You will need to merge all or pats of this file with other data files prior to analysis. Key variables to identify unique records in each file for merging are SITE_ID, and VISIT_NO (in that order).

<u>Selecting Index Visits and Samples</u>

To select the index visit for a particular indicator, use the variable INDXVISIT_*, where * is a particular indicator (e.g., CHEM for water chemistry). This will select a single visit record from the lake information file, and a corresponding sample record from the indicator file. Typically, index visits are defined as the first visit with valid data, but in some cases alternatives have been selected. In the indicator file, INDXVISIT_* needs to be coupled with INDXSAMP_* to select the actual index sample record.

To select only the probability-sample lakes, use the variable SITE_TYPE=PROB_Lake for extent estimates, and SITE_TYPE=PROB_Lake and LAKE_SAMP=TargetSampled to select the samples to be used for condition estimates. WGT_NLA is the adjusted weight variable to be used for population estimates (both extent and condition).

LAKE PROFILE DATA

NLA_LAKEPROFILE_VALID_20091008.*
NLA_LAKEPROFILE_INFO_20091008 (variable names and labels)

Maximum length of many cables is 50 m (some were shorter), so some profiles may not have been taken at deepest point of lake, or else profile is incomplete. Hopefully these have been identified with SAMPLED_PROFILE=YES, PARTIAL PROFILE.

Values determined to be definitely invalid due to probe failures have been set to missing and assigned a "K' flag. Suspicious values have been given a "U" flag.

Reviewed calibration notes and comments and flagged values with "X" where a calibration failure was noted (but did not set to missing). Did not flag cases where the dilute phosphate pH buffer did not pass—there were apparently preparation problems with some batches, though if the conductivity passed then it might be a pH probe issue—probably deserves some more scrutiny.

SECCHI DATA:

NLA_SECCHI_VALID_20091008
NLA_SECCHI_INFO_20091008

In final data set, values that were noted as clear to bottom were set to missing (SECMEAN=.), and SAMPLED_SECCHI=YES, CLEAR TO BOTTOM.=Y.

Sites where Secchi was not determined were assigned SAMPLED_SECCHI=NOT DONE, FLAG_SECCHI=K.

EPILIMNETIC DISSOLVED OXYGEN:

NLA_EPI_DO2_VALID_20091007;
NLA_EPA_DO2_INFO_20091007;

Has mean values of all DO measurements from lake profile taken between 0 and 2 m, or top 50% of profile if depth was < 2 m.

There are no flags or comments with this file. Missing values are associated with either missing or invalid profile measurements for dissolved oxygen.

Data was used to assign condition classes for the NLA report.

VISUAL ASSESSMENT DATA:

NLA_VISASSESS_VALID_20091015
NLA_VISASSESS_INFO_20091015

Combined ASSESSMENT_VALID and ASSESSMENT_COMMENT files.

Larger lakes (> 5000 ha area) did not have shoreline observations. Set SAMPLED_ASSESS=YES, PARTIAL, FLAG-ASSESS='U' (for non-standard measurement), and explained in COMMENT_ASSESS.

Lakes less than 5000 ha that did not have shoreline observations but that had assessment results were treated the same, but the COMMENT_ASSESS explains why. Smaller lakes were presumed to have valid visual assessments despite the lake of shoreline observations (i.e., presumed that crew could see entire shoreline from index and/or launch site).

Needed to sanitize the "comment" variables to remove reference to personal information/ Also removed references to presence of youth or girls' camps. Otherwise left entries as is (no spelling corrections, e.g.).

File includes "indices" of stressor activities (residential, industrial, recreational, agricultural, and lake management). The intensity classes (Low, Medium, and High) from all the stressor activities were combined into "*_STRING" variables, and then converted into numerical *_SCORE variables by setting L=1, M=3, and H=5 and summing them within each major class of stressor.

WATER CHEMISTRY AND CHLOROPHYLL

NLA_WATQUAL_20091123 (Water chemistry, chlorophyll a, and Secchi data)
NLA_WATQUAL_INFO_20091123 (Variable names and labels)

This file combines the water chemistry, chlorophyll a, and Secchi data (see SECCHI DATA section) for all samples (including field duplicates; total number of records=1,326). Secchi data are merged with SAMPLE_CATEGORY=P (Primary samples) records. Field duplicate samples are identified by SAMPLE_CATEGORY='D'. Field duplicates are associated with the set of re-visited lakes.. One can subset out the set of samples from re-visit lakes after merging (by SITE_ID and VISIT_NO) with the Lake Information file and selecting REPEAT=Y'.

For major cations and anions, there are variables that present results in both mg/L (*_PPM) and μeq/L (*).

Reporting limits were established to be approximately 2 × the long-term MDL, and the lowest calibration standard was supposed to be equal to the RL. Values below the RL are identified using *_RL-ALERT='Y". Data below the MDL were operationally assigned a value equal to 0.5 × the target MDL. The RL for chlorophyll was adjusted from the original target value to 0.1 μg/L, after reviewing the QC sample data again and finding the wrong set of low-level QC samples were used.

Samples that exceeded the target holding times (*_HT_ALERT=Y) were evaluated to determine if sample integrity had been compromised (e.g., ion balance check, comparison to other variables that had been analyzed within the target holding time). IF a value was determined to be invalid, it was flagged and set to missing.

SEDIMENT DIATOM DATA

Sample Information:

NLA_SEDDIA_SAMINFO_20091102
NLA_SEDDIA_SAMINFO_INFO_20091102

Organized as one record per sample ID (top and bottom slices have different SAMPLE_ID numbers, ending in 7 and 8). The variable SAMPLE_TYPE or SAMPLE CLASS can be use to subset out top and bottom core samples. Total number of records is

Note there are a few "field duplicate" core samples (SAMPLE_CATEGORY='D'), although duplicate cores were not required to be taken. These should all have INDXSAMP_CORE=NO, and should be dropped before any analysis.

Samples with SAMPLED_CORE='YES (Lost sample), may represent "ghost samples" (recorded on the field form but never collected).

Unique record identifiers: SITE-ID, VISIT_NO, SAMPLE, SAMPLE_ID, in that order.

Count file:

NLA_SEDDIA_COUNT_ALL_20091026 (SAS data set)
NLA_SEDDIA_COUNT_TOP_20091026 (Excel file—top samples only)
NLA_SEDDIA_COUNT_BOT_20091026 (Excel file—bottom samples only)
NLA_SEDDIA_COUNT_INFO_20091026 (Variable names and labels)

Unique record identifiers: SITE_ID VISIT_NO SAMPLE_ID NLA_TAXANAME in that order).

SAS file of all samples has 73,351 records!! Top and bottom samples (SAMPLE_TYPE=SEDT and SEDB) were split into two Excel files (*-top_*, *_bot_*)

OTU_LUMP1 OTU_LUMP2, and OTU_LUMP3 are variables created by NASP and MSU, and no information about these was provided with the data file.

Diatom Lake Disturbance Condition Index (aka Diatom IBI)

NLA_E99_LDC_DATA_20090623 (LDC values)
NLA_E99_LDC_DATA_INFO_20090623 (variable names and labels)

This file only has the final LDC index scores (no metrics or scores). Includes scores for both top and bottom slices (SAMPLE_TYPE=SEDT and SEDB, respectively). Only the SEDT samples were used to assess condition.

Also, in a few cases, the visit 2 core was used as the index sample. Use PRIMESEDT='1' to select the index samples prior to merging with the Lake Information file..

DIATOM-INFERRED CHEMISTRY:

Input data files (from lab):
NLA_TOP_BOT_INFER_20091103
NLA_TOP_BOT_INFER_INFO 20091103

Initial input file was provided by the Academy of Natural Sciences Philadelphia (NLA_TOP_BOTTOM_INFERENCES_DVP; DVP made minor changes to make import to SAS work). Final input data file has some corrections from original file.

The percent urban, crops, agriculture, and wetland estimates agree quite well with those from the basin landuse metric file (the vast majority within ±1 percent, and all within 10 percent).

Final data files (with field and lab data):

NLA_DIA_INF_CHEM_20091124
NLA_DIA_INF-CHEM_INFO_20091124

File has 1254 records, and includes two "field duplicates" (SAMPLE_CATEGORY='D'). These will be dropped when the index samples are selected (INDXSAMP_INF=YES).

Inferences were made for one sample per site (index visits only, INDXSAMP_INF=YES) where both top and bottom samples had count data. There are 593 sample records that have SAMPLE-TYPE=SEDB (bottom slice) that were not considered reservoirs by the field crew. However, only 499 visits have inferred chemistry (SAMPLED_INF=YES). SAMPLED_INF=NO CORE SAMPLE where no core sample was collected. SAMPLED_INF=YES (LOST SAMPLE) when it appeared a bottom sample was collected but never shipped to the lab. SAMPLED-INF=NO LAB DATA when it appeared a sample was collected and sent to the lab, but there are no inferred chemistry results (e.g., there is no inferred chemistry results for re-visits). Sites where no bottom sample was collected because it was believed to be a reservoir have SAMPLED_INF=NOT DONE.

Note that results will not necessarily match up with the final determination of lake origin (LAKE_ORIGIN). Assessment was done using data as is, rather than basing it on LAKE_ORIGIN.

Only those variables directly pertaining to inferred chemistry were retained in the final data file. The input data files from the lab have all of these variables.

ZOOPLANKTON

Field sample information

NLA_ZOOP_SAMINFO_20091020
NLA_ZOOP_SAMINFO_INFO_20091020

The SAMINFO file has 1 record per sample ID, and has the final sample information from the field and lake information files.

Only one sample pair from a given site is identified as the "index" sample (INDXSAMP_ZOOP=YES). In most cases, this is VISIT_NO=1 and SAMPLE-CATEGORY=P. In cases where the original primary sample was invalid, the field duplicate samples was used and assigned SAMPLE_CATEGORY=P. When all visit 1 samples were invalid, the visit 2 sample was used.

Count files:

NLA_ZOOP_COUNT_20091022
NLA_ZOOP_COUNT_INFO_20091022

Unique record identifiers= SITE_ID VISIT_NO SAMPLE_ID TAXANAME (in that order).

The COUNT file contains 1 record per taxon (TAXANAME) per sample ID, combined with field sampling information from the SAMINFO file (has almost 18,000 records!!). Sites where one or both samples were lost or incomplete had ZOOP_FLAG_FLD='X' so they can be deleted before further analysis. Note that some taxa names may be present I both samples from a site (coarse and fine-mesh).

PHYTOPLANKTON

Sample information file:

NLA_PHYT_SAMINFO_20091023
NLA_PHYT_SAMINFO_20091023

This file contains on record per sample, and includes both field and laboratory sample information.

Soft Algae Count file:

NLA_PHYT_SOFTCOUNT_20091023
NLA_PHYT_SOFTCOUNT_INFO_20091023

Count file has 31,971 records!!

Data received was inconsistent across different laboratories for some sample-related information such as volumes..

Some samples had missing sample volumes on the field form (these are flagged with COMMENT_FLD_PHYT='IM'. Sample volume information from label not provided in lab count file, only the initial volume of the subsample. In some cases, volume might be inferred to be 1000 mL (e.g., when INIT_VOL=333 mL).

There is an "extra" sample from NLA-06608-0105, visit 1 (SAMPEL_ID=502403). IT was not recorded on the field form, and there are already primary and field duplicate samples for VISIT_NO=2. Probably can use as an extra field duplicate. It has SAMPLE-CATEGORY='X' to allow one to delete it easily.

Diatom count file:

NLA_PHYT-DIATCOUNT_20091125
NLA_PHYT-DIATCOUNT_INFO_20091125

This file has the diatom subsample data for those samples where diatoms made up > 2% of the sample.

File has 16,126 records!!

PLANKTON O/E DATA

Count file:

NLA_OE5_OTUCOUNT_20091019
NLA_OE5_OTUCOUNT_INFO_20091019

Has combined phytoplankton and zooplankton taxa with OTU names assigned by C.P. Hawkins, Note OTU_OE5 may not be the same as OUT_CPH in the zooplankton and phytoplankton files.

Only has visit 1 samples, Samples all assigned SAMPLE_CATEGORY='P', so linkage back to original zooplankton and phytoplankton samples and count data is probably lost.

Relative density estimates (ABUND_OE5) are in indiv/mL. For phytoplankton OTUs, ABUND_OE5 is obtained directly from the ABUND variable in the phytoplankton count data file. For zooplankton OTUs, ABUND_OE5 is derived from the ABUND variable in the zooplankton count file as follows:

Original Volume used for counting

$$CV = \frac{\frac{VOL\_COUNT}{INIT\_VOL}}{2}$$ (2 is used because sample was split at the lab.)

The total volume of water sampled (diameter of the zooplankton nets was 65 mm), 1000 converts m^3 to L

$$VOL\_SAMPLED = DEPTH\_OF\_TOW \times \left[\left(0.0065 \right)^2 \times \pi \right] \times 1000$$

The density (indiv/mL) is then calculated as:

$$ADJ\_ABUND = \left(\frac{ABUND}{CV \times VOL\_SAMPLED} \right) \times 0.001$$ (0.001 converts from L to mL)

ABUND_OE5 for zooplankton is the sum of the ADJ_ABUND values for all taxa grouped into a single OTU.

Use ABUND_OE5 with great caution,. It is not clear that CV values can be calculated correctly, since some samples had VOL_COUNT > INIT_COUNT, and in some cases these are recorded in mL, while in others they represent a proportion of the sample. Problem is that no one (field or lab) ever recorded the original volume of the sample in the bottle.

O/E data file:

NLA_OE5_VALID_20091021
NLA_OE5_VALID_INFO_20091021

Contains O/E scores for visit 1 samples only. Potentially invalid scores were obtained when one or both plankton samples were either lost or incomplete. These have SAMPLED_OE5=YES, PARTIAL, and are set to missing before any condition estimates are made.

In the WSA-ECO3=PLNLOW region (Plains and Lowlands), the O/E model required shoreline habitat variables. For lakes > 5000 ha, habitat data was not available. SAMPLED_OE was set to NOT DONE for these samples.

PHYSICAL HABITAT

Metric files:

NLA_PHABMET_20091116 (all metrics calculated from field form data)
NLA_PHABMET_INFO_20091116 (variable names and labels)
NLA_PHABMET_20091116 A (subset of entire PHABMET file)
NLA_PHABMET_20091116_B (subset of entire PHABMET file)

PHABMET has too many variables to import into Excel, so only tab-delimited (*.txt) and comma-delimited (*.csv) are provided. Variables are arranged as follows (basically in order they appear on the field form):

Littoral depth, surface films, littoral bottom substrate (size class, color, odor), macrophyte cover, littoral fish cover, riparian vegetation (structure and cover), shoreline substrate (size class), riparian human influence/disturbance, littoral fish macrohabitat characteristics, bank features and lake level fluctuation, and some initial indices for riparian disturbance, riparian vegetation quality, littoral cover complexity, and littoral-riparian cover complexity. These indexes serve as source material for the final O/E-based habitat indicators.

PHABMET was split into two Excel files (*_A and *_B). Common variables in each file are SITE-ID, YEAR, VISIT_NO, UID, SAMPLED_PHAB, DATEPHAB, FLAG_PHAB, and COMMENT_PHAB.

File A Littoral depth, surface films, littoral bottom substrate (size class, color, odor), macrophyte cover, littoral fish cover, riparian vegetation (structure and cover), (VARNUMs 7-186 in the INFO file).

File B contains shoreline substrate (size class), riparian human influence/disturbance, littoral fish macrohabitat characteristics, bank features and lake level fluctuation, and some initial indices for riparian disturbance, riparian vegetation quality, littoral cover complexity, and littoral-riparian cover complexity (VARNUMs 187-362 in the INFO file).

You can merge selected variables from each file into a new file using either SITE_ID and VISIT_NO or UID.

All files contain 1,252 records (so they include re-visits). SAMPLED_PHAB='YES' if any habitat data were collected. SAMPLED_PHAB='NOT DONE" for lakes where no habitat data were collected (supposed to be only lakes with area > 5000 ha, but there are a few with smaller areas that were also not done for various reasons). The visual assessment data file flags and comments can be used to explain why habitat data is not available for some lakes with area < 5000 ha.

"Super metrics" files:

NLA_PHABSUPERMET_20091120 (metrics need to produce the final habitat indicators)

NLA_PHABSUPERMET_INFO_20091120 (variable names and labels)

This file includes additional metrics derived from those in PHABMET, and the final four habitat indicators (riparian disturbance, riparian vegetation, littoral cover, and riparian-littoral cover complexity). For riparian vegetation, littoral cover complexity, and riparian-littoral cover complexity, different variants of a particular metric were used in different reference site clusters and/or aggregated ecoregions. There are variables (RVegQ_Var, LitCvrQ_Var, and LitRipCVQ_Var) that contain the name of the "base" metric used for that particular site.

POPULATION ESTIMATE FILES

FINAL DATA folder contains condition class files derived from the final data files. There may be some slight differences between these estimates and thos used in the final report (mostly dealing with assignment of NOT ASSESSED vs. NO DATA).

Condition Class files:

NLA_CHEM_COND_20091123 (Condition based on various water quality variables [e.g., nutrients, turbidity, ANC, salinity])
NLA_EPI_DO2_COND_20091123 (Condition based on epilimetic dissolved oxygen)
NLA_INFERCHEM_COND_20091124 (Condition based on diatom-inferred chemistry)
NLA_LDC_COND_20091123 (Biological condition based on sediment diatom lake disturbance index)
NLA_OE5_COND_20091125 (Biological condition based on plankton O/E indicator)
NLA_PHAB_COND_20091130 (Condition based on physical habitat quality indicators)
NLA_REC_COND_20091123 (Condition based on recreational indicators [microcystin, chlorophyll a, and cyanophyte density])
NLA_TROPHIC_COND_20091123 (trophic state assignments based on total P, chlorophyll, and Secchi)

These files serve as the input files to the extent and status estimation process. They contain all the variables from the lake information file needed to describe various subpopulations, all original variables used to derive condition classes, and the condition class variables themselves.

Population estimation script:

NLACatAnalyses_20091130_dvp.R:
Geodalbers.R

The NLACatAnalyses file does 2 things:

1) computes extent estimates for various components of the sample frame, using the NLA_LAKEINFO_ALL_* file (e.g., target vs. non-target, target-sampled vs. target-not sampled). Produces output file NLA_Extent_Estimates_20091130.csv.

2) Computes status estimates for each of the condition class files (estimated number of lakes in each condition class). The Geodalbers file is needed to convert geographic coordinates to an Albers projection to calculate the local variance estimate. Produces output file NLA_Condition_Class_Estimates_20091130.csv

Extent and status estimates are expressed as both percentages (Estimate.P) and actual lake numbers (Estimate.U). Confidence intervals are provided as both the actual values of the upper and lower limits, and as the width of the interval (MarginErr.P and MarginErr.U). The upper and lower limits are required when producing cdfs, while the interval width is needed for error bars on bar charts.

Output files can be imported into Excel for graphics. Note in cases where a particular condition class was not present within a given subpopulation, there will be no record, and you may need to insert a record with 0 values for everything to make bar charts come out correctly.

The script is set up to produce the same estimates as those used in the final report. You will need to edit the script if you want different subpopulations (e.g., State).