Data Use Notes for the 3rd Generation ISCN Database (12/2015 version)

By Luke Nave, ISCN Coordinator

These data use notes are intended to help data users orient to, interpret, and understand the limitations of the ISCN Database. They are not intended as a replacement for, but rather a supplement to, other documentation on the ISCN website, including the Data Policy, the Data User Quick Start Guide, the Variables and Calculations pages, and other information linked from the Dataset Information page (http://iscn.fluxdata.org/Data/Pages/DatasetInformation.aspx). We recommend every data user familiarize her/himself with all of these materials before beginning an analysis, as the size and complexity of the ISCN Database and its products as made available here are not trivial. We strive to create data products that are useful as a starting point for the data user; ultimately, individual data uses are different enough from project to project that we expect data users to invest their own time and effort into understanding and developing what we have assembled here into what is needed for a specific project. Please contact ISCN support if you have made a reasonable attempt to understand the documentation and still need help with data interpretation. Note that documentation for the 2nd generation Database is still available at by scrolling down, as is technical support through ISCN support.

1. Submission dataset vs. SOC dataset

It is important to understand the distinction between submission and SOC datasets before data use. Each submission dataset- a named collection of data values at one or more geolocations- bears a unique submission dataset name (dataset_name_sub). That submission dataset may or may not have contained calculated SOC stocks when it was contributed. If calculated SOC stocks were contributed with the dataset, the name of the SOC dataset (dataset_name_soc) will be the same as the submission dataset name. However, most of the data contributed to ISCN do not include computed SOC stocks. When possible, the ISCN Database computes SOC stocks for individual soil layers (and sums these up to the profile level) using the contributed information (layer thicknesses, bulk densities, and carbon concentrations). Because these computations are not performed by the data contributor, they are considered their own dataset- a derivative of the submitted data, performed by ISCN Database curators- and are identified as such with a distinct dataset_name_soc. At the profile level, there are two variants: 'ISCN SOC stock computation' and 'ISCN SOC stock to 1m computation;' the former is for the entire profile (whether more, less, or by coincidence equal to 1m), the latter applies only to profiles ≥1m deep and is the SOC stock truncated to a standardized 1m depth. If some of data needed to compute the SOC stock are lacking for at least one layer in a profile, that profile will have a dataset_name_soc of 'ISCN No SOC stock computation.' Because any submission dataset can have multiple SOC datasets associated with it, the same profile or site name can appear multiple times in a profile-level database report. Because very few users will want to analyze datasets with multiple SOC values for the same location, some example use cases are provided below with guidelines for filtering out unwanted "duplicates" (i.e., rows of repeated profile_names with different SOC values).

1. Plug-and-play uses (e.g., land modelers quick-validating profile SOC stock predictions, beginning users comparing broad differences in profile total SOC stocks)

Data users wanting the easiest route to profile total SOC stocks for a widespread number of locations should not filter on dataset_name_sub. Rather, filter on dataset_name_soc, and analyze all profiles that have dataset_name_soc of either 'ISCN SOC stock computation' or 'ISCN SOC stock to 1m computation' (depending on whether full profiles or profiles to 1m are desired). The primary benefit of this approach is that it requires very little data manipulation before use; the primary drawback is that it significantly decreases the number of profile observations available to the data user.

2. Intermediate data uses (e.g., land modelers or broad-scope data users willing to invest modest effort in dataset preparation)
Data users willing to invest modest effort into preparing a more extensive dataset can first perform the data filtering steps described under the first case. However, many more profile SOC stocks can be included in the analysis by performing several additional steps. First, select all profiles that have dataset_name_sub = dataset_name_soc (i.e., profiles with contributor-computed SOC stocks) and include these in the analysis. Next, verify that these profile_names are not already represented by profiles with an ISCN SOC stock computation value (i.e., the same profile_name with a dataset_name_soc of 'ISCN SOC stock computation…'). If the profile does appear more than once, compare the SOC stock computations done by the contributor vs. ISCN and determine which to use based on the SOC computation methods as described in the soc_method column (see Section 3, Data Quality Flags).

3. Advanced data uses (e.g., mechanistic models, vertical SOC distributions, pedology studies)
Data users, whether beginning or advanced, are always urged to conduct their analyses using datasets that they have built from the bottom up, with careful attention to QAQC, project-specific protocols for handling missing or overlapping data, and detailed documentation of the procedures used to convert the database reports ISCN provides as a starting point into the datasets used for final analyses. In this regard, beginning with layer-level database reports, and performing SOC computations using the raw, contributed data is the best practice. Indeed, ISCN SOC stock computations use the available data to compute SOC stocks without regard to the uncertainties introduced into the SOC computation by those data. For example, either %organic carbon (oc) or %total carbon (c_tot) may be the source of the carbon concentration used in the computation, depending on which is available. Likewise, multiple forms of bulk density are contributed by data submitters and sometimes the only bulk density value available for a layer is a derived from an estimate or a biased method. Thus, data users are encouraged to begin their analyses by downloading the relevant layer-level database report, inspecting and removing any duplicate layers, and performing their own SOC computations (including, if desired, gap-filling for layers with missing data).

2. Data quality flags
ISCN Database reports contain quality flags that are intended to serve as easily-sorted variables for restricting observations according to some commonly useful criteria.

soc_carbon_flag

Layer-level database reports- Syntax examples: 'no_soc[no c or bd]' and 'soc[c_tot*bd_samp]' This flag is present in layer-level database reports and identifies whether a layer has a computed SOC stock. If so, it identifies which variants of %C and bulk density were used in the computation; if no SOC was computed, it identifies which variables were missing. In the case of SOC stocks computed by the Alaska Deep Soil Carbon Project (Johnson et al. 2011), the soc_carbon_flag codes are described under previous documentation (the second-generation database; see below).

Profile-level database reports- This flag is present in profile-level database reports to indicate the depth and completeness of the profiles and whether any gap-filling was used in computing their SOC stocks. Profiles flagged as 'Complete' vs. 'Short' constitute >70% of the observations; these connote an intact profile with a depth of at least (Complete) or less than (Short) 1m. Profiles flagged as 'Gap' have an un-sampled gap ≥5cm thick somewhere within the profile. A flag containing the string '(5cm)' means that there is an un-sampled gap ≤5cm thick somewhere within the profile.

soc_method

Layer-level database reports- At this time, this flag is not utilized in layer-level database reports.

Profile-level database reports- This flag describes the approach used to compute profile SOC stocks. For profiles that were contributed along with computed SOC stocks, this describes something about the approach, formula, or gap-filling used in the computation. Flags prefixed with 'no fill…' are for profiles for which the SOC stock is computed by ISCN.

bd_method- (in layer-level database reports) as provided by the data contributor, this flag identifies the approach, computation, or portion of the sample was used for the determination of bulk density.

bdNRCS_prep_code and cNRCS_prep_code- (in layer-level database reports) many data in the ISCN Database come from the USDA-NRCS, National Soil Survey Lab. Those familiar with NRCS data and lab protocols will recognize the sample preparation codes that identify the methods used to prepare samples for C and bulk density determinations, as described in Method Code 1B of the SSL Lab Methods Manuals. The vast majority of data contained in the ISCN Database are from prep code S (air-dried and sieved <2mm before analysis, with final data corrected to an oven-dry basis).

xxxxx_note- (in layer-level and profile-level database reports) Database reports contain a variety of columns named 'xxxx_note' (where 'xxxx' is typically a separate variable [e.g., layer_name and layer_note or site_name and site_note]); these columns contain additional information provided by the data contributor and are sometimes useful for data interpretation. The contents of these 'xxxx_notes' are quite variable, thus use of PivotTables or other filtering techniques is a useful way to inspect them more closely to ensure important data usage notes are not missed.

3. Limitations to spatial and temporal data

Given the number and diversity of datasets contributed to ISCN, there are a variety of disclaimers that are necessary to spell out here. The first key limitation pertains to the precision (and in some cases, probably the accuracy) of geocoordinates. This is especially important when point values are extracted from other spatial data products and then overlaid on top of the locations of sites in the ISCN Database. First, note that the reference datum for geocoordinates is specified for most contributed data; however, the LAT/LONG values in the database are not harmonized to any one standard reference datum. Expected deviations between reference data (e.g., WGS84 and NAD27) are likely only on the tens of meters for most sites, but this may be important for very high-resolution work (e.g., using DEMs). For that matter, while up to 5 decimal degrees (~1m precision) are accommodated by the ISCN Database, experienced researchers know that this is very often unattainable in field (or especially forest) settings. Thus, geocoordinates should be considered, at best, site-level attributes, and their precision and accuracy should be considered provisional unless demonstrated otherwise by independent means (e.g., validation against aerial imagery, communication with data contributor). Finally, it is worth considering that, for sites lacking geocoordinates, it is still often possible to generally constrain locations using information in the COUNTRY or STATE columns. For many analyses, knowing the exact location of the observations is not as important as having internally consistent characterization data and an ability to broadly constrain the location.

There are important considerations regarding the timing of when data in the ISCN Database were collected. First, considering the date of profile observation and sampling (obs_date), be aware that the data received from contributors spans many decades of sampling, in some cases running back into the early decades of the 20th century.

Also, note that some of the attributes contained in the database (e.g., landuse derived from remote sensing observations) are extracted from spatial data products with a distinct date stamp. Thus, these attributes may not be correct for sites/profiles/layers that were sampled many years before the derived, remote sensing data.