Supplementary material to the paper Wells AF et al. Vegetation classification for northwestern Arctic Alaska using an EcoVeg approach: tussock tundra and low and tall willow groups and alliances. Vegetation Classification and Survey. DOI: 10.3897/VCS.65469

Supplementary material 1. Data compilation methods for the Ecological Land Survey Legacy Database (ELD).

Ecological Land Survey Legacy Database (ELD)

Data Compilation Methods

We began by identifying all existing ELS datasets from ABR projects across Alaska, and obtained permission to use proprietary data in advance of data compilation and analysis. Next, we compiled the final Quality Assurance and Quality Control (QA/QC) reviewed data from their original sources, typically in Microsoft Access format, and reorganized to conform with our standardized ELS database schema. After the original data sources were identified, we prepared a table of data compilation steps needed between individual columns in the original tables and the required columns in the standard schema. These steps included data type conversions, SQL-based logic statements, and external join tables necessary to convert the variety of older data formats to a standard form. A series of scripts written in the R programming language (R Core Team 2020) executed these steps, which involved reading data from the original files, transforming them, and inserting the data into a PostgreSQL database containing all the data from all projects. In PostgreSQL, each source project was assigned an individual schema, or workspace, within the database that represented our first, semi-automated attempt at migrating the data. Because of the strictly enforced relationships between data and reference tables in our standardized ELS database, these intermediate schemas allowed us to resolve the issues with these relationships before integrating the data into the final data tables. Other conversions between old and new codes or values were built into database tables that automated the migration of data from the project schema tables to the final tables within the database. As part of this process, conversions from old codes to new codes were made as necessary to maintain referential integrity between the final data tables and reference tables. In some cases, original source files (e.g., paper data sheets) were reviewed to resolve questions on possible transcription errors. Missing or NULL values were assessed in the project-level schema and populated where possible. For example, a project performed 20 years ago may have required a complete species list, but did not have fields to record the cover of vegetation structure classes (e.g., cover of dwarf shrubs) as in contemporary plots. In this case, the structure fields were populated based on the originally recorded species-level cover values using a cross-reference table between species and structure classes. Where NULL values could not be populated, the data were updated to reflect that data were missing (no data, -999). After converting old codes and addressing missing data, data from each project were migrated to public schema tables designed to store data on environmental variables, vegetation, and soils. Public schema tables also contain information on voucher specimens for species that were difficult to identify in the field as well as soil laboratory analyses. The public schema tables contain the final QA/QC-reviewed data for use in subsequent analyses. Data from 6,986 relevé plots sampled across 31 individual field studies from across Alaska, between 1992 and 2019, were successfully compiled into the ELD.

Literature Cited

R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org