Photo by Dan Milchunas

A Team Approach to Data Synthesis: The Playbook for Creating a Centralized, Dynamic, and Sustainable ANPP Database

Nicole E. Kaplan[1], Kristin Vanderbilt[2], Lee Zeman[3], Judith B. Cushing[3], Christine Laney[4], Juli Mallett[3], Ken Ramsey[4], Jincheng Gao[5], Judith Kruger[6], Carri Leroy[3], Daniel Milchunas[7], Esteban Muldavin[2]

[1] Dept. of Soil and Crop Sciences, Colorado State University, Ft. Collins, CO 80523-1170, USA, [2] ept. of Biology, University of New Mexico, Albuquerque, NM, 87131-1091, [3]The Evergreen State College, Olympia, WA, 98505, [4] Dept. of Biology New Mexico State University, Las Cruces, NM 88003-8001, [5]Department of Biology, Kansas State University, Manhattan, KS 66506-4901, [6] South African National Parks, Scientific Services, Skukuza, South Africa, 1350, [7] Dept. of Forest, Rangeland, & Watershed Stewardship, Colorado State University, Ft. Collins, CO 80523-1472

Background: Ecologists are interested in synthesizing regional and/or cross-site aboveground Annual Net Primary Productivity (ANPP) values to answer questions related to ecosystem structure and function in a changing world. Knapp and Smith (2001) assessed temporal dynamics of ANPP with mean total ANPP values from eleven Long Term Ecological Research (LTER) sites. Today, ecologists are interested in more refined analysis of ANPP values for different life forms and species of plants, and making predictions of community and population responses to variability in precipitation and global change phenomena. Such efforts rely on integrating large datasets and are hindered by the lack of standard methodologies for data collection and detailed metadata documentation across sites.

Project Description: The Grasslands Data Integration (GDI) project has brought together ecologists, information managers and computer scientists to address the interdisciplinary challenges of integrating ANPP data from multiple sources. In this poster we present 1) the necessity to coordinate expertise and information to integrate ANPP data and metadata from five national and international grassland LTER sites, 2) the data model we designed to archive and serve the data, and 3) analysis planned for the future. This collaboration is an example of how professionals with inter-related work experience build a community of experts and a successful data product for the LTER (Baker and Millerand 2007).

Methods at The Sites: Data formats, experimental design, and methodologies for measuring ANPP differed considerably among sites. Furthermore, different sites use specific codes to record species level data. These data require significant transformation, restructuring, or standardization to extract useable measures for cross-site analyses, but guidelines and tools for these processes are not readily available.

Site	Sampling Method	Number of Sub-Sites	Number of Years of Data	Number of Vegetation Types	Number of Sampling Units	Number of Experimental Units within each Sampling Unit
Kruger National Park (Kruger)	Regression relationships	35	17	35	35	9-41
Konza Prairie (KNZ)	Biomass harvest	1	5	1	2	40
Jornada Basin (JRN)	Regression relationships	15	17	5	15	49
Sevilleta Wildlife Refuge (SEV)	Regression relationships	3	8	3	15	16
Shortgrass Steppe (SGS)	Biomass harvest	6	23	1	3	5

Table 1. Sampling methods and experimental designs for ANPP at each site within the GDI database. Measurements are made directly from total harvesting of standing crop biomass at SGS and KNZ (Milchunas et al. 1994), and estimations are based on species-specific regression relationships between biomass and plant volume (Muldavin et al. 2008, Huenneke et al. 2002) at SEV, JRN, and Kruger. A number of years of data from each site are contained in the GDI. The number of vegetation types sampled for ANPP at each site was determined by local ecologists, as well as the number of sampling units, which make-up a replicate for statistical analysis. The number of experimental units are the number of plots or quads within each sampling unit or replicate.

The GDI Data Model:

- Centralized, physical database designed by computer scientists, information managers and ecologists
- Integration from all sites into one database schema
- Tool standardizes site-specific plant species codes with USDA PLANTS codes (http://plants.usda.gov/, USDA, NRCS 2008)
- Granularity of data is representative of observed ANPP:
  - What – mass of a single plant species (grams per square meter)
  - When - specified date
  - Where – specified location
- Important metadata joined to observed ANPP (i.e. location)
- Data exploration and quality control enabled by preliminary analysis (Figure 1)

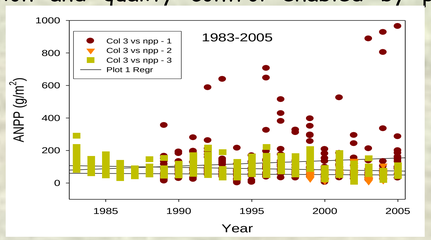

Figure 1. This figure illustrates the value of preliminary exploration and analyses by information managers and ecologists after integrating the data. The early comparison of JRN ANPP data (npp-1) to other sites (npp-2 is SEV, npp-3 is SGS) erroneously indicated that the JRN is significantly more productive than similar grassland sites despite being the warmest and driest of these sites. This verified for JRN the updates required for their regressions and highlighted the influence of a single species (*Yucca elata*) on the analysis.

Take home messages for dynamic, successful and sustainable data integration:

1) Collaboration between Information Managers, Ecologists, and Computer Scientists is necessary to create a valid, updateable, and sustainable data model to support dynamic integration and analysis
2) Comparison of methodologies and identification of statistically valid sampling units should be performed early
3) Standardization of units of measurements and derivations maintain data quality
4) Standardization of species codes, vegetative characteristics and other metadata facilitates detailed analysis
5) Performance of exploratory analysis aids in quality assurance
6) Design data model to support important & interesting analyses

References:

Baker, K.S. and F. Millerand. 2007. Articulation Work Supporting Information Infrastructure Design: Coordination, Categorization, and Assessment in Practice. *Proceedings of the 40th Hawaii International Conference on System Sciences.* 1530-1 605/07.

Knapp, A.K. and M.D. Smith. 2001. Variation among biomes in temporal dynamics of aboveground primary production. *Science* 291:481-484.

Milchunas, D.G., J.R. Forwood & W.K. Lauenroth. 1994. Productivity of long-term grazing treatments in response to seasonal precipitation. *Journal of Range Management*. 47:133-139.

Muldavin, E.H., · D.I. Moore, S.L. Collins, K.R. Wetherill & D.C. Lightfoot. 2008. Aboveground net primary production dynamics in a northern Chihuahuan Desert ecosystem. *Oecologia*. 155:123-132.

Acknowledgements: NSF Canopy Database Project (NSF Grants: DBI-0417311, DBI-0319309), JRN-LTER (NSF Grant: DEB-0080412), KNZ-LTER (NSF Grant: DEB-0218210), SEV-LTER (NSF Grant: DEB-0080529), & SGS-LTER (NSF Grant: DEB-0217631).

Next Steps:

- Development of GDI Browser
- Generate automatic QAQC reports
- Establish data warehouse
- Conduct multivariate analysis (samples of analyses in figures 2, 3, and 4)

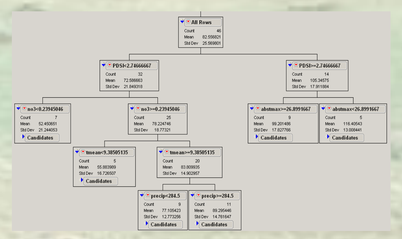

Figure 2. CART model describing variation in ANPP at SEV, SGS and JRN LTER sites. CART model explains 64% of the variation in ANPP across a 23-year integrative study as affected by Palmer Drought Severity Index (PSDI), maximum and mean temperatures, and precipitation.

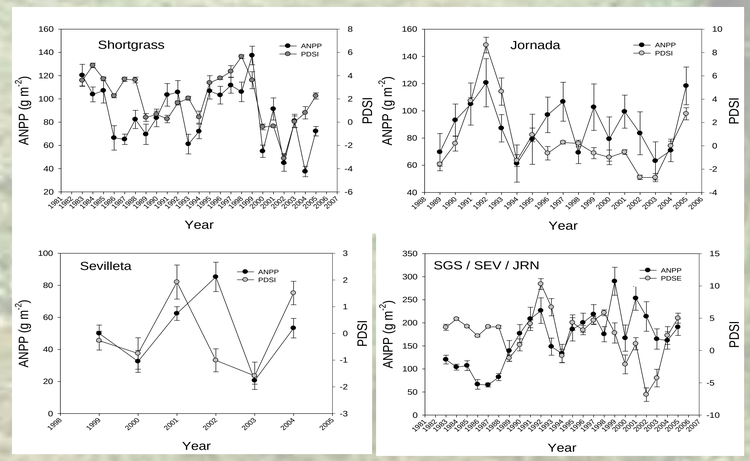

Figure 3. Average ANPP at individual LTER sites (A-C) and averaged across all LTER sites (D) through time. ANPP is plotted with PDSI through time to show temporal patterns, strong correlations between ANPP and PDSI (Pearson's *r*) and possible lags in ANPP.

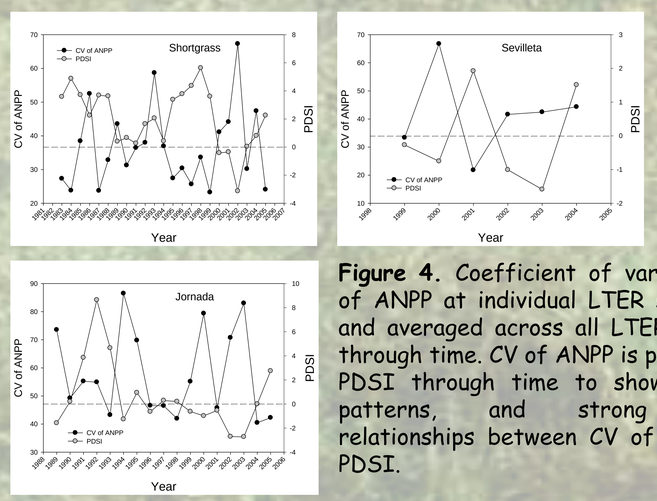

Figure 4. Coefficient of variation (CV) of ANPP at individual LTER sites (A-C) and averaged across all LTER sites (D) through time. CV of ANPP is plotted with PDSI through time to show temporal patterns, and strong inverse relationships between CV of ANPP and PDSI.