ASM 2006, September 23, 2006
Judy Cushing, organizer, Grasslands ANPP, Data Integration Case Study

Background of our Case Study:
This is a prototype data integration tools for 3 grassland sites (SGS, SEV, JRN). We discussed preliminary analysis and future plans in consultation with Alan Knapp, Dan Milchunas, and Este Muldavin.

Project Overview:
Data integration strategy includes data from different sites, into an intermediate database and then out. We described sampling design at each of the 3 sites: (biomass per species) (see slides)

Project Goals:
Our consultant researchers made suggestions regarding analyses which are discussed below and including other sites in the prototype. Judith Kruger, Kruger Nat'l Park is also interested in including ANPP data from South Africa. We plan to include scientific analysis of ANPP with species and form from participating sites (SGS, SEV, and JRN). KNZ has ANPP by form, but not species so may be included in the near future.

We discussed 3 outcomes of our work.
(1) Feasibility study - integration of a few datasets from different sites with a preliminary statistical analysis of these data. From a computer scientist perspective it is interesting to identify the challenges that ecologists have in integrating relatively large data sets. Computer scientists traditionally abstract the problem to come up with a simplified solution. A simplified solution may be challenging to apply to the larger problem(s) of data integration. In addition, we may be able to contribute this project to the TRENDS projects.

Project Status:

Notes for SGS-
14-17% under-estimate of cool seasons based on C14 data.
Owl creek – different community type, others are different soil and landscape position type. Clipping time- labor intensive, so not initially designed to look at species level dynamics. Look at cover data for species level questions. LTNPP is appropriate for grasses, forbs, and total. Early data does not have cactus in it and it is very difficult to each ANPP for cactus, cactus is also so patchy that it is difficult to capture with this sampling design.

Notes for Kruger (biomass per site)-
39 landscape or Veld conditions. 100 Discs pasture readings, species composition, and tuft diameter readings are performed on 522 1 m2 plots, then regressions for standing crop (kg/ha). Exclosure in the park vary where harvesting for ANPP is performed at 1m2 (?).

Notes from Integration of SGS, SEV and JRN datasets (Lee Zeman)-

1 big database, for now only fall sites, no piñon juniper from SEV, and KNZ not in the database b/c not split up by species

Notes on Discussion of Database Creation –
Data and metadata come from websites, and conversations were necessary to have with IMs, particularly regarding species codes and how they jive with the USDA plants database. Integration of USDA plants database with codes from each site is a useful tool to have and to use in the future for cross-site research. Identifying major vegetation types and growth forms to the lowest common denominator was another challenge.

Finally, changing species codes over time presented another challenge. Questions arise as to what to do when different species are pooled and recorded together over time at one site and not another. The granularity of the species identification may need to at the broadest level of information that is available. However, is this IM's or biologist's decision? If the finer grained information is tracked in the database, it will preserve the opportunity for the biologist "end-user" to make decisions. One recommendation is to establish plants database codes as a standard. However, some codes may be created at a site to avoid identifying a plant to species level, an example includes (ASOX at SGS), which can be either an Asragalus or Oxytropis species.

We discussed the issue of proper usage of the data and how metadata should be used to document flags that identify criteria that guard against mistakes that may happen when synthetic researchers are integrating or analyzing data improperly. For example, diversity analysis may not be appropriate for these datasets. Need closer connection between metadata, warnings and data.

Discussion of Analysis of grassland plant communities at JRN, SEV and SGS LTER sites (Carri)-
Used 1999 only
species richness : # of different species in 1m2
NPP : total net annual Primary Productivity
Community – weighted by NPP, or species presence or absence, weighted can tell us how productive are species in a plot, not just presence or absence – this may be comparable to cover
Indicator species

Results-
NPP: SGS – multiply by 4 for NPP, significant differences between SEV and JRN, but no significant differences between SGS and SEV or SGS and JRN
Species richness: – not measured at SGS – plan to add 4 of 5 sub-plots at each plot, but still can not perform species area curves. There were significant differences in species richness between JRN and SEV.
At JRN Mesquite is most productive, playa least productive. SEV? SGS is presented as Blue Grama b/c there is no dominant species overlap between SGS and other sites.

Ordination – measures significant differences among groups within each site, sub-site, plot and sub-plot based on Presence or Absence of species in plot. We see differences among vegetation types (sub-sites) at a site. JRN spans vegetation types, SGS is the least

broad. JRN has significant differences among plots; b/c range of low to high productivity is represented in location of plots and sub-plots, which are not located randomly. Owl Creek at SGS separates out nicely as a shrubland – other plots are not significantly different from one another. When plots are weighted by NPP we see **fascinating results**. SEV has significant differences between black grama and creosote shrubland plots. We see statistical backings for patterns that are there!

Indicator species analysis - shows species that are relatively abundant and frequent given in a certain ecosystem or vegetation type, plot or whatever level you are looking at. Is a measure of fidelity – relative abundance and frequency together!

Species Area Curve - **but call it something different**!! – new ways based on re-sampling and permutation of the data. These results can determine what vegetation type we are located in. Re-sample data until you are not getting new species as you continue to sample.

Sub-plot (Quadrat) based analysis is below the scale of interest – summarize quadrat data to a given sub-site. Statistical n becomes 3 or 5, which is typical of ecological data and once all years of data are analyzed there will be lots of data!!!

Goals for Future Analysis (brainstorming):

More years analyzed, to look at trends through time – database should be annual
Add 2 S. African sites and 4 other lter sites: NWT, KBS, KNZ, CDR

Perform correlations w/environmental factors: ppt, temp, soil, texture, soil type, elevation, AET, soil moisture, PAR, soil, temp, days above -5 c temp = growing season – timing of ppt (highlighted factors are studied at KNZ)
Noted from biologists is there are lots of variables and adaptations from vegetation types to site niches. There are questions of what environmental factors drive the variability?

Questions regarding more data integration-

Quality control issues, scale, and versioning of data transformations are all questions.

An example from TRENDS (presented by Christine Laney) and a possible collaboration between databank and TRENDS: Store original dataset with link to dataset download and start with what Trends already have. PPT is already complete in TRENDS. Question: how far away is weather measured from NPP collection sites, because PPT can be variable across our research sites. Maybe we can represent that visually with GIS, and put it in the weather data.

Biologists would like to see-
Biomass by species
Presence/absence
Cover-based ordinations

Can't do-
Relative frequency, diversity, species abundance, species richness

Other challenges that biologists have –
Differences in PPT and NPP between grassland sites
SEV data in older data sets were collected with completely different methods. We see anomaly of high NPP compared to SGS, we see it not falling below SGS, but JRN. Different Regressions, methods, and weight analyses were used and this is a question for scientists. Biologists reported we would expect line for NPP relative to ppt. This is an important issue before we make x-site comparisons – how comparable are we?

Researchers know whether they are underestimating or overestimating NPP – maybe think of it as an index. Meta – analysis with 100s of sites will probably ok. We recommended making adjustments in derived NPP for under and over estimation and need to discuss ways to fine-tune the data calculations.

Talk about the bigger picture here-

Biologists and LTER community agree that NPP is a very important dataset. It is worthwhile to reconcile some of these issues. It is a challenge that how NPP is measured is so variable, so start our integrations and analyses with best data available.

Here we jumped into analysis without any questions from researchers. This is just an exercise, but we were able to show problems with data as well and it has been a good exercise that way!

Talk about including other sites-
How will this data model accommodate complexity of different NPP datasets across the network? We should scale up to common denominator, but flag in metadata proper uses for dataset given original scale on which it was measured. At what level do we report – is there advantages to reported at sub-plot level – report both was suggested – summary data to help us prevent errors! It was suggested that we should all scale to g/m2 for NPP
Stay with grasslands sites to keep it simpler and extend this data model for sharing data.
Data cleaning processes and species transformations would become more complicated and we need to work out protocols, and continue to talk about different methodologies.
This may be opportunity for post ASM money, for a NPP workshop(s) with researchers to clarify differences in experimental protocols. Other Challenges to address with Researchers include transformations and changes (updates) in species codes.
Plan winter 2006-7 NPP workshop. Judy spoke with Bob Waide at Network office, and would like to bring NPP people together and then follow-up with nuts and bolts with the technology.

<u>FUTURE NEXT STEPS:</u>
Standardize addressing and reported over/under estimation and integrate it into dataset
Species codes – transform to USDA plants database
Great to standardize the metadata for all the NPP data sets

PARTICIPANTS:

Name	Affiliation	e-mail
Judy Cushing	The Evergreen State College	judyc@evergreen.edu
Daniel G. Milchunas	SGS, CSU	Daniel.milchunas@colostate.edu
Nicole Kaplan	SGS-LTER	Nicole.kaplan@colostate.edu
Judith Kruger	S. Africa	Judith@sanparks.org
Jim Koelliker	KNZ	koellik@ksu.edu
Becky Riggle	SGS	beckyr@cnr.colostate.edu
Seth Munson	SGS	smunson@cnr.colostate.edu
Christine Laney	JRN	chrlaney@nmsu.edu
Harmony Dalgleish	KNZ	hjdal@ksu.edu
Emily Grman	KBS	grmanemi@msu.edu
Doug Moore	SEV	dmoore@sevilleta.unm.edu
Esteban Muldavin	SEV	muldavine@sevilleta.unm.edu
John Anderson	JRN	janderson@nmsu.edu
Kristin Vanderbilt	SEV	vanderbi@sevilleta.unm.edu
Lara Reichmann	JRN	Lara_reichmann@brown.edu
Ken Ramsey	JRN	kramsey@jornada.nmsu.edu
Lee Zeman	evergreen	zemanl@evergreen.edu
Carri LeRoy	evergreen	LeRoy@evergreen.edu
Terry Loecke	KBS	loeckete@msu.edu
Madeline Scheintaub	SGS	mscheint@lamar.colostate.edu