Ajustes de datos: transformación de datos.

Capítulo 9 de McCune y Grace
2002

Razones estadísticas para transformar datos

- Mejorar las suposiciones de algunas técnicas estadísticas: normalidad, linealidad, homocedasticidad, etc.
- Hacer que datos medidos en escalas diferentes sean más comparables entre sí.

Razones ecológicas para transformar datos

- Mejorar el desempeño de las medidas de distancia composicional
- Reducir el efecto de los totales; enfocar en medidas relativas.
- Asemejar las importancias relativas de especies comunes y raras.
- Enfatizar en las especies más informativas

Tipos de transformaciones

- Transformaciones monotónicas
- Transformación probabilística (Beals)
- Relativizaciones
- Eliminación de especies raras
- Combinación de entidades
- Diferencias entre fechas
- Diferencias primarias (en series temporales)

Asuntos de notación

- En las ecuaciones que siguen se usa la siguiente notación:

x_{ij} = the original value in row i and column j of the data matrix

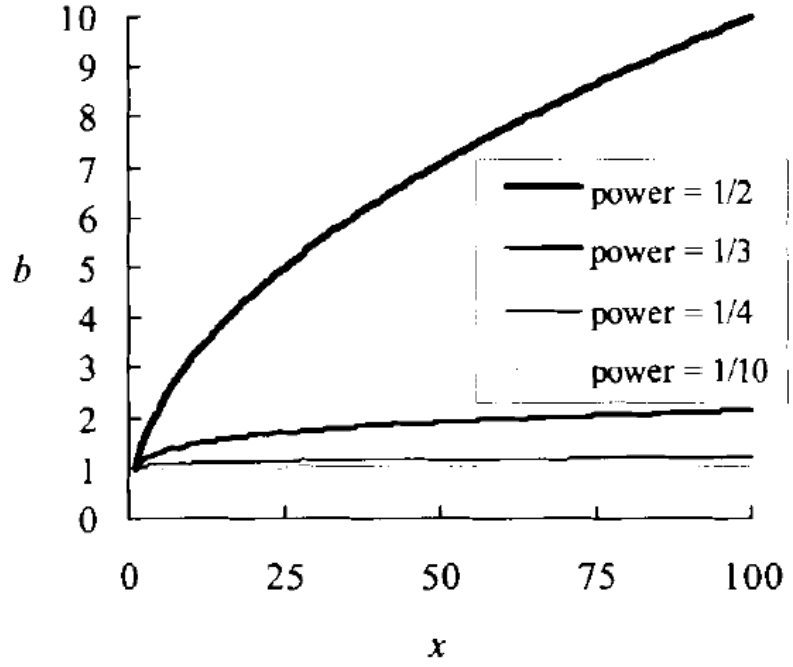
b_{ij} = the adjusted value that replaces x_{ij}.

Transformaciones monotónicas

- Se aplican a cada elemento de la matriz independientemente de los demás elementos.
- Monotónicas porque cambian la magnitud de los valores sin cambiar su posición relativa.

Transformaciones de potencia

- Ecuación general: $b_{ij} = x_{ij}^{p}$
- Mientras menor el valor de p más se comprime la magnitud de los valores altos
- La más utilizada es p=0.5, o raíz cuadrada de x



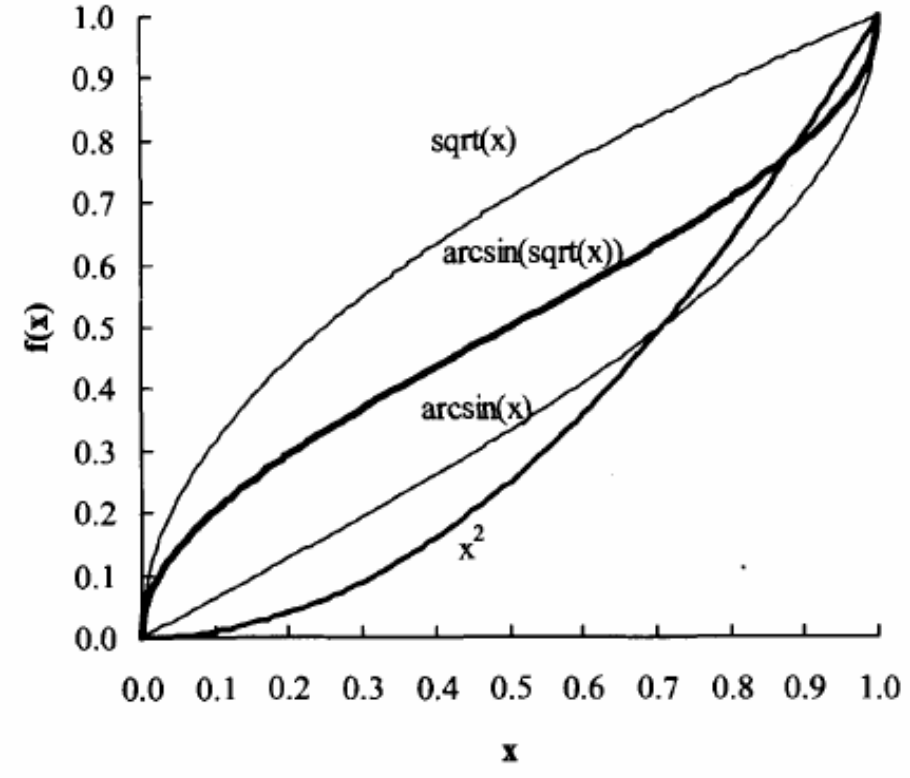
Transformación logarítmica

- Ecuación general: $b_{ij} = \log(x_{ij})$
- Comprime valores bien altos y riega los valores bajos
- Útil cuando hay una variación grande en los valores
- Ya que log(0) no está definido se acostumbra utilizar: $b_{ij} = \log(x_{ij} + 1)$
- Pero puede tener consecuencias indeseables en ciertos casos.
  - Pag. 69 del texto describe una ecuación alternativa para evitar estas consecuencias.

Transformación raíz cuadrada del arco-seno

$$b_{ij} = 2/\pi * \arcsin\left(\sqrt{x_{ij}}\right)$$

- Recomendada para datos de proporción
- Riega los extremos y comprime el centro de la escala



Suavización de Beals

- Sustituye cada celda de la matriz por la probabilidad de que la especie ocurra en esa unidad de muestra.

$$b_{ij} = \frac{1}{S_i} \sum_k \left(\frac{M_{jk}}{N_k} \right)$$

$$b_{ij} = \frac{1}{S_i} \sum_k \left(\frac{M_{jk}}{N_k} \right)$$

	sp1	sp2	sp3	sp4	sp5	S_i
SU1	1	0	1	1	1	4
SU2	0	0	0	1	0	1
SU3	1	1	0	0	0	2
N_j	2	1	1	2	1	

Species k

Species j	1	2	3	4	5
1	2				
2	1	1			
3	1	0	1		
4	1	0	1	2	
5	1	0	1	1	1

	sp1	sp2	sp3	sp4	sp5
SU1	0.88	0.13	0.75	0.88	0.75
SU2	0.50	0.00	0.50	1.00	0.50
SU3	1.00	0.75	0.25	0.25	0.25

Relativizaciones

- Muy util para datos de comunidades
- La decisión sobre relativizar o no, y cual relativización utilizar debe basarse en la pregunta que se hace sobre los datos.
- También conviene determinar cual es la variación en los totales; si es poca la relativización tendrá poco efecto.
  - La variación se puede estimar con el coeficiente de variación (CV)

- Si CV < 50%, relativizacion generalmente tiene poco impacto en los resultados.
- Si CV >100%, tiene gran impacto.

Relativización general

Por columnas $$b_{ij} = \frac{x_{ij}}{\left(\sum_{j=1}^{q} x_{ij}^{p}\right)^{1/p}}$$ $$b_{ij} = \frac{x_{ij}}{\left(\sum_{i=1}^{n} x_{ij}^{p}\right)^{1/p}}$$ Por filas

- Si p=1, relativización es por totales
  - Apropiado cuando la técnica se basa en distancias de bloques de ciudad.
- Si p=2, es el equivalente Euclidiano

Relativización por máximo

$$b_{ij} = x_{ij} / \mathbf{xmax}_j$$

- Tiende a igualar las especies comunes y raras.
- Es conveniente cuando los datos que fueron tomados en diferentes unidades de medida (e.g., cobertura y área basal) se quieren analizar juntos.

Relativización binaria según la media

$$b_{ij} = 1 \text{ if } x_{ij} > \bar{x} \ , \quad b_{ij} = 0 \text{ if } x_{ij} \leq \bar{x}$$

- Abundancias son convertidas a presencia o ausencia; 1 o 0
- Enfatiza las porciones óptimas de las distribuciones de especies

Ponderación por ubicuidad

$$b_{ij} = U_j \, x_{ij} \quad \text{where} \quad U_j = N_j / N$$

- Las especies que ocurran en mayor numero de muestras llevaran valores mas altos.

Informacion por ubicuidad

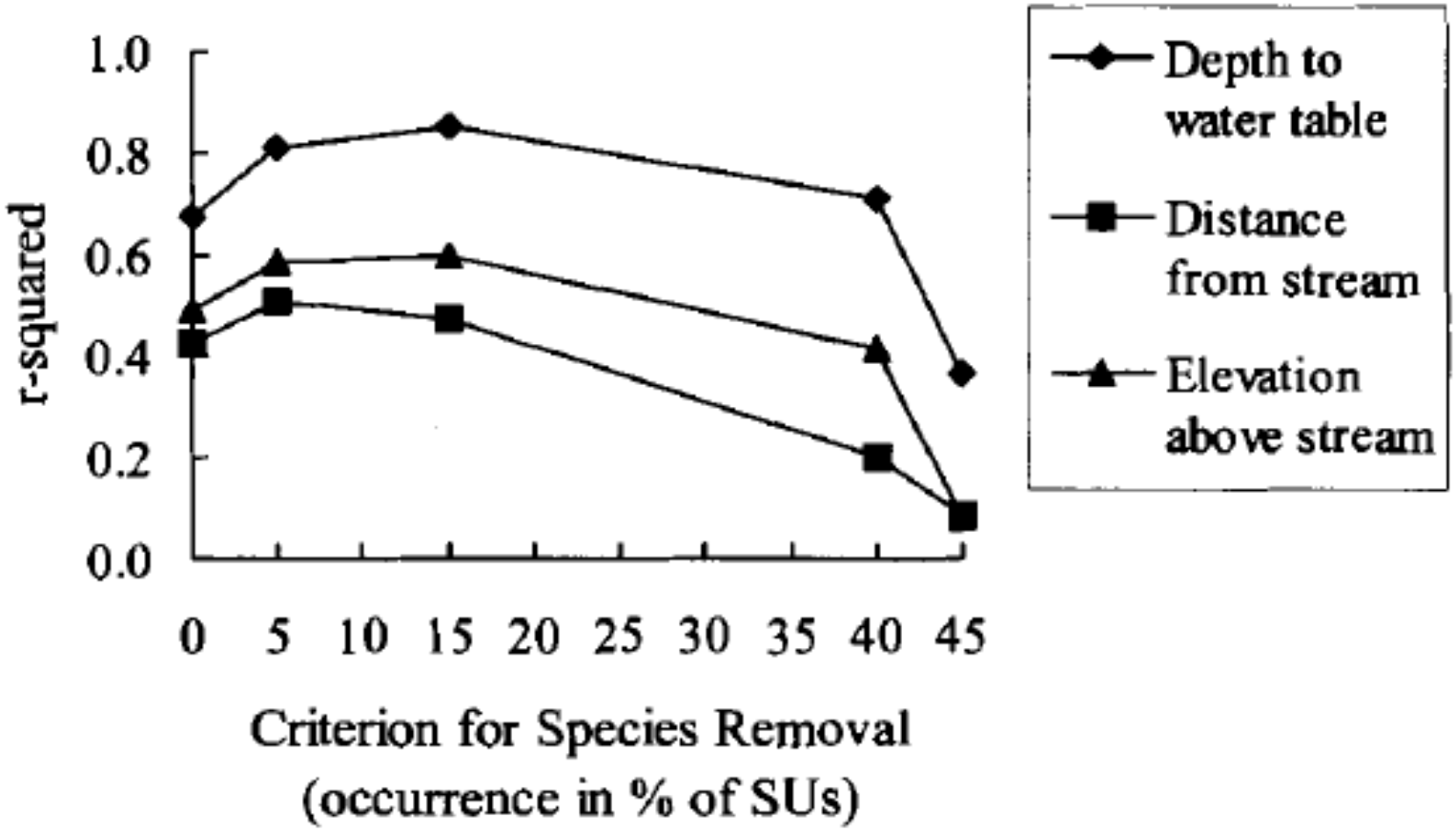
$$b_{ij} = I_j \, x_{ij}$$

- La mayor cantidad de informacion esta contenida en las especies que ocurran en la mitad de las unidades de muestra
- Especies bien comunes o bien raras llevaran el menor peso.

Eliminación de especies raras

- Eliminar especies raras reduce el ruido de los datos y a menudo mejora la detección de relaciones entre la composición y el ambiente.
- La regla general es eliminar especies que ocurran en <5% de las unidades de muestra.

Eliminación de especies raras



Diferencias entre fechas

- Cuando se mide la abundancia en el mismo lugar pero en mas de una fecha
- Las diferencias indican cambios (e.g., sucesión, degradación)
- Tienden a la normalidad y linealidad
- Son datos apropiados para técnicas que suponen esas características (e.g., PCA)

Secuencia de pasos: datos de especies

- Calcular estadísticas descriptivas
- Eliminar especies raras
- Transformaciones monotónicas
- Relativizaciones por filas o columnas
- Cotejar si hay rezagados

Secuencia de pasos: datos ambientales

- Calcular estadísticas descriptivas
- Transformaciones monotónicas
- Relativizaciones por filas o columnas
- Cotejar si hay rezagados