E

TWC/26/20 Add.
ORIGINAL: English
DATE: September 24, 2008

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Sixth Session
Jeju, Republic of Korea, September 2 to 5, 2008

ADDENDUM

CORRELATION BETWEEN DIFFERENT TYPES DISTANCES/SIMILARITY ON A SET OF WINTER OILSEED RAPE CHARACTERISTICS OF DIFFERENT TYPES (NOMINAL TO RATIO)

Document prepared by experts from Germany

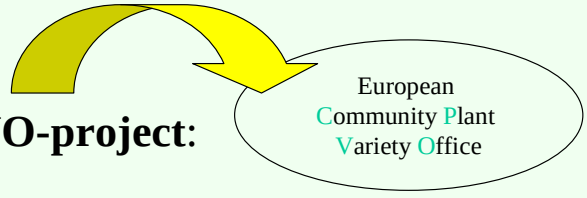
CORRELATION BETWEEN DIFFERENT TYPES OF DISTANCE/SIMILARITY MEASURES ON A SET OF WINTER OILSEED RAPE CHARACTERISTICS OF DIFFERENT TYPES (NOMINAL TO RATIO SCALE)

Uwe Meyer
Bundessortenamt Hannover – Germany

TWC/26/20

Introduction

Aims of the CPVO-project:



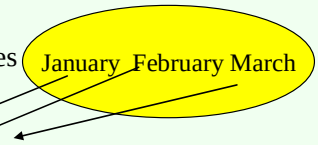
- Study of management of WOSR reference collections (see also TWC/26/18)
- Identification of appropriate statistical procedures to analyze morphological data

Datasets

Dataset 1: Notes and measurements
UK, FR, DK and DE in 2003, 2004 and 2005

Dataset 2: Consolidated Notes and measurements
UK, FR, DK and DE in 2003, 2004 and 2005

Consolidation = Harmonization between countries



Example: Harmonization of dates (Add 31+28+31=60 days)
(char: time of flowering has different starting points in the countries:
1th January, 1th April, …)

Definitions

- Similarity measures
  - Cosinus, Dice, Jaccard, M, RR, Kulczinski, Gower
- Dissimilarity measures
  - Minkowski metric , Cityblock, Euclidian distance, maximum distance
- Correlation measures
  - Pearson

Notation

v	number of variables (here characteristics) or the dimensionality
x_j	data for observation x on the i^{th} variable (characteristic), where i=1 to v (here observation = variety per year)
y_j	data for observation y on the i^{th} variable (characteristic), where i=1 to v
w_j	weight for the i^{th} variable. w_i=0 when either x_i or y_i is missing
W	the sum of total weights
$\bar{x}$	mean for observation x
$\bar{y}$	mean for observation y

Weighted means

$$\bar{x} = \frac{\sum_{i=1}^{v} \left(w_i * x_i \right)}{\sum_{i=1}^{v} w_i}$$

$W_i = 1/v \quad i=1,\ldots,v$

Standardization

- Z-Score standardization: $z_i = \frac{x_i - \bar{X}}{\sigma}$

-Range standardization: $z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$

Minkowski metric

$$d(x,y) = \sqrt[p]{\sum_{i=1}^{v} \left| x_i - y_i \right|^p}$$

For p = 2 → Euclidian distance
p = 1 → Cityblock distance

Cityblock distance (p=1)

$$d(x,y) = \sum_{i=1}^{v} \left| x_i - y_i \right|$$

Scale levels – TGP/8

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

Influence

- Similarity measures
- Dissimilarity measures
- Correlation measures
- Standardization

Gower's index (1)

-S(x,y) = $\frac{\sum_{i=1}^{v} w_i * \delta_{x,y}^i * d_{x,y}^i}{\sum_{i=1}^{v} w_i \delta_{x,y}^i}$

$\delta_{x,y}^i = 1;$ for nominal, ordinal, interval and ratio chars

-Special case:

- for asymmetric nominal variable

$\delta_{x,y}^i = 1;$ - if either x_i or y_i is present

$\delta_{x,y}^i = 0;$ - if both x_i and y_i are absent

Gower's index (2)

- S(x,y) = $\frac{\sum_{i=1}^{v} w_i * \delta_{x,y}^i * d_{x,y}^i}{\sum_{i=1}^{v} w_i \delta_{x,y}^i}$

- for nominal chars $d_{x,y}^i = 0, if\ x_i \neq y_i$ $d_{x,y}^i = 1, if\ x_i = y_i$

- for ordinal, interval and ratio chars $d_{x,y}^i = 1 - |x_i - y_i|$

  - for ordinal chars ranks has to be used

Pearson correlation coefficient

$$r(s,t) = \frac{\sum_{j=1}^{n}\left(s_j - \bar{s}\right)*\left(t_j - \bar{t}\right)}{\sqrt{\sum_{j=1}^{n}\left(s_j - \bar{s}\right)^2 * \sum_{j=1}^{n}\left(t_j - \bar{t}\right)^2}}$$

for assessing linear relation between two variables s and t

Variables s and t are here different distance measures.

Selection of appropriate methods

	Nominal		Ordinal	Interval	Ratio	Combination nominal/ ordinal/ interval/ratio
	Two Categories (notes)	>two Categories (notes)				
Cityblock			X	X	X	
Euclidian			X	X	X	
Chebychev			X	X	X	
Cosinus	X					
Dice	X					
Jaccard	X					
M coefficient	X					
RR coefficient	X					
Kulczinski coefficient	X					
Gower's index	X	X	X	X	X	X

Influence of the year

Data: Dataset 1 (german part)

Similarity measure: Gower‘s index

Results:

Sample 1	Sample 2	Similarity Measure	Correlation coefficient
DE2003	DE2004	Gower’s index	0.81926 (P<0.0001)
DE2003	DE2005	Gower’s index	0.82339 (P<0.0001)
DE2004	DE2005	Gower’s index	0.84790 (P<0.0001)

Influence of the years in German Dataset was very low.

Modification of dataset 1

Aim: Comparison of different distance/similarity measures

b1 (Seed: erucic acid; 1=absent, 9=present) =! ordinal
b6 (Leaf: lobes; 1=absent, 9=present) =! ordinal
b13 (Production of pollen; 1=absent, 9=present) =! Ordinal

Nominal with 2 categories (notes) = ordinal with 2 categories (notes)

b10 (Flower: Colour of petals; 1=white, 2=cream, 3=yellow, 4=orange-yellow) →dropped

It is forbidden to handle char b10 which is nominal scaled with more than two categories (notes) as ordinal, interval or ratio scaled characteristic.

Correlation coefficients

Sample	Measure 1	Measure 2	Correlation Coefficient
DE2003	Cityblock	Euclid	0.94925 (P<0.001)
		Chebychev	0.81139 (P<0.001)
		Gower	-0.95598 (P<0.001)
	Euclid	Chebychev	0.94786 (P<0.001)
		Gower	-0.85326 (P<0.001)
	Chebychev	Gower	-0.67777 (P<0.001)

Correlation coefficients

Sample	Measure 1	Measure 2	Correlation Coefficient
Consolidated dataset 2	Cityblock	Euclid	0.95687 (P<0.001)
		Chebychev	0.87801 (P<0.001)
		Gower	-0.92994 (P<0.001)
	Euclid	Chebychev	0.97336 (P<0.001)
		Gower	-0.81894 (P<0.001)
	Chebychev	Gower	-0.70844 (P<0.001)

Conclusions (1)

- Main efforts are to be made on harmonization of protocols, and harmonization of notations between experts that register the measures
- Statistical computations, as shown above, need to be selected according to the type of scale of the characteristics
- When some characteristics have a great influence on the synthetic (calculated) value (e.g. Gower's index) obtained over all characteristics, or when there are different types of scales in a dataset, one has to consider using either the whole dataset, or to drop some characteristics, or to compute subsets per type of characteristic
- The "Gower's index" is the most appropriate procedure for the structure of dataset 1 and 2 because it is the only one which allows a combination of the present data types
- It is not allowed to use nominal scaled characteristics like characteristic b10 (Flower: color of petals; 1=white, 2=cream, 3=yellow, 4=orange-yellow) with more than two categories (notes) for evaluation of the "Cityblock distance"

Conclusions (2)

- For comparison of different distance measurements dichotomous characteristics (b1, b6, b13) can be handled as ordinal characteristics. Nominal characteristics with more than two categories (b10) have to be dropped for that comparison.
- The best correlated measure to "Gower's index" on the basis of dataset 1 and 2 is the "Cityblock distance"

[End of document]