

ver.3.14.

**Métodos de geografía lingüística
con estadística multivariante
Análisis de asociación**

Hiroto Ueda, Universidad de Tokio

14 de marzo, Universidad de Barcelona

Análisis de asociación

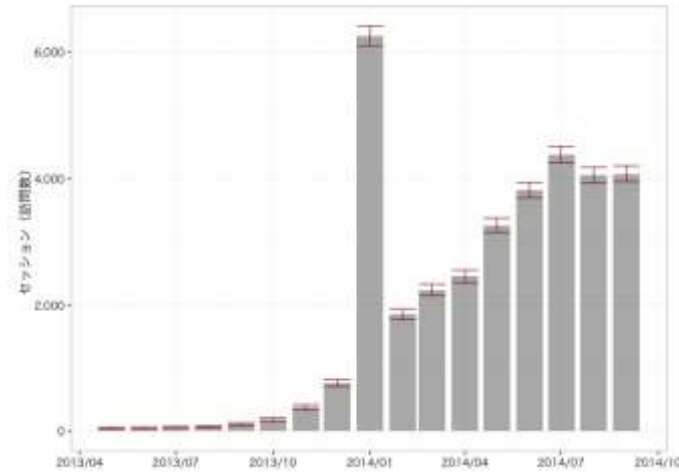


"Comprados juntos habitualmente:"

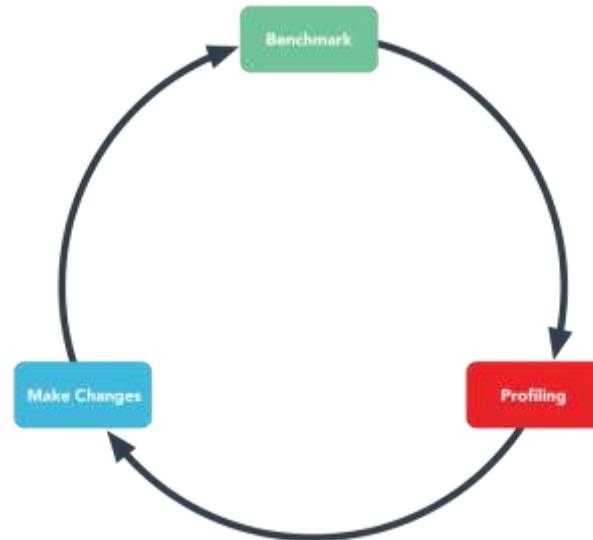
Libro A => Libro B

Libro A, B => Libro C

1. Análisis descriptivo:



2. Causa y efecto:



Tres cifras propuestas:

1. «Support» (Soporte)

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

$$\text{Soporte}(A, B) = 2 / 5 = .400$$

$$\llcorner\text{Soporte}\llcorner(A, B) = \llcorner\text{Soporte}\llcorner(B, A)$$

Rango: [0, 1]

2. «Confidence» (Confianza)

probabilidad condicionada de $A \Rightarrow B$.

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

$$\text{Confianza}(A, B) = 2 / 2 = 1.000$$

$$\text{Confianza}(B, A) = 2 / 3 = 0.667$$

Rango: [0, 1]

3. «Lift» (Elevación).

«Confianza»(A, B) con respecto a la probabilidad de B:

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

Elevación(A, B)

$$= \text{Confianza}(A, B) / \text{Probabilidad}(B)$$

$$= (2 / 2) / (3 / 5) = 1 / 0.6 = 1.667$$

«Elevación»(A, B) \neq «Elevación»(B, A)

Rango: [0, ~]

Nuestra propuesta:

$$\text{Síntesis} = (\text{Soporte} * \text{Confianza} * \text{Elevación})^{1/3}$$

$$\text{Máximo: } c/N * 1 * 1/(c/N) = 1 \text{ (!)}$$

$$\text{Rango} = [0, 1]$$

(donde c: coocurrencias, N: número de datos)



(1) Análisis de asociación simple

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

#	L	R	LHS =>	RHS	L.f.	R.f.	L.p.	R.p.	Cooc.	Sup.	Cnf.	Lift	Synt.
1	1	2	A =>	B	2	3	.400	.600	2	.400	1.000	1.667	.667
2	2	1	B =>	A	3	2	.600	.400	2	.400	.667	1.667	.444
3	4	3	D =>	C	1	3	.200	.600	1	.200	1.000	1.667	.333

4	3	4	C =>	D	3	1	.600	.200	1	.200	.333	1.667	.111
5	1	3	A =>	C	2	3	.400	.600	1	.200	.500	.833	.083
6	3	1	C =>	A	3	2	.600	.400	1	.200	.333	.833	.056
7	2	3	B =>	C	3	3	.600	.600	1	.200	.333	.556	.037
8	3	2	C =>	B	3	3	.600	.600	1	.200	.333	.556	.037
9	1	4	A =>	D	2	1	.400	.200	0	.000	.000	.000	.000
10	2	4	B =>	D	3	1	.600	.200	0	.000	.000	.000	.000
11	4	1	D =>	A	1	2	.200	.400	0	.000	.000	.000	.000
12	4	2	D =>	B	1	3	.200	.600	0	.000	.000	.000	.000

(2) Análisis de asociación **doble**

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

#	L.	R.	LHS	=>	RHS	L.f.	R.f.	L.p.	R.p.	Cooc.	Sup.	Cnf.	Lift	Synt.
1	2:3	1	B:C	=>	A	1	2	.200	.400	1	.200	1.000	2.500	.794

2	1:3	2	A:C	=>	B	1	3	.200	.600	1	.200	1.000	1.667	.693
3	1:2	3	A:B	=>	C	2	3	.400	.600	1	.200	.500	.833	.437
4	1:4	3	A:D	=>	C	0	3	.000	.600	0	.000	.000	.000	.000
5	1:2	4	A:B	=>	D	2	1	.400	.200	0	.000	.000	.000	.000

(3) Análisis de asociación triple

...

(4) Análisis de asociación cuádruple

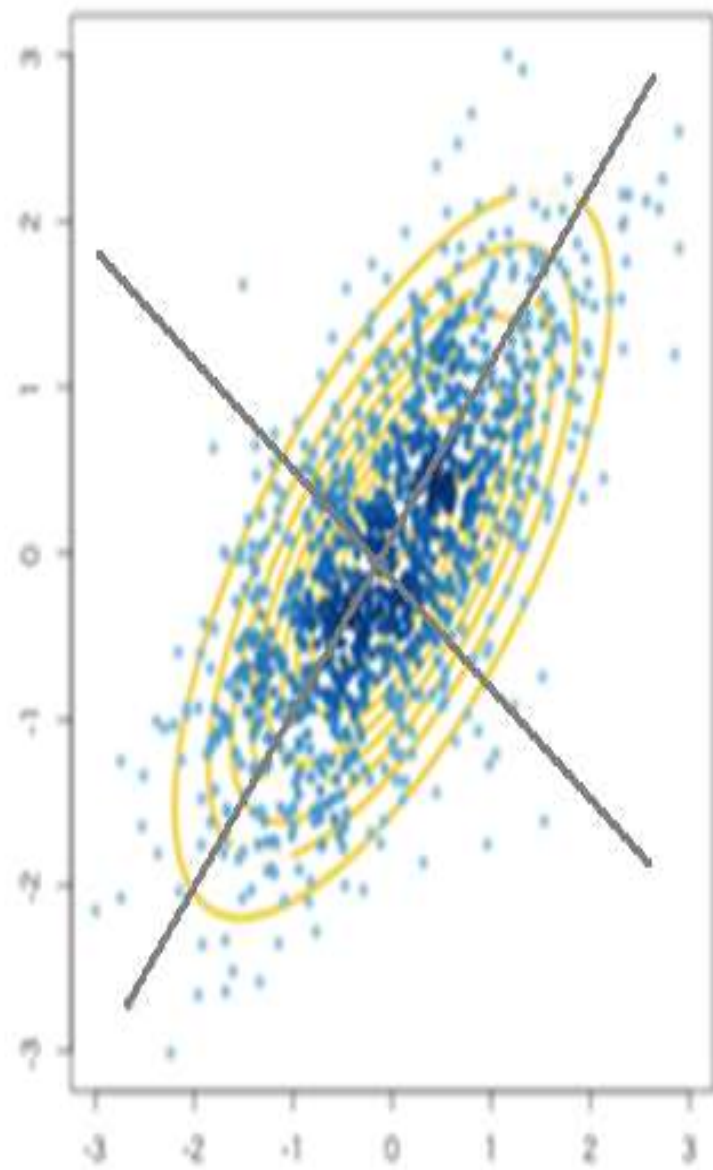
...

... (?)

Addenda: Análisis de componentes principales

- Álgebra lineal
- Matrices
- Producto
- Matriz inversa
- Valores eigen (valores propios),
etc.





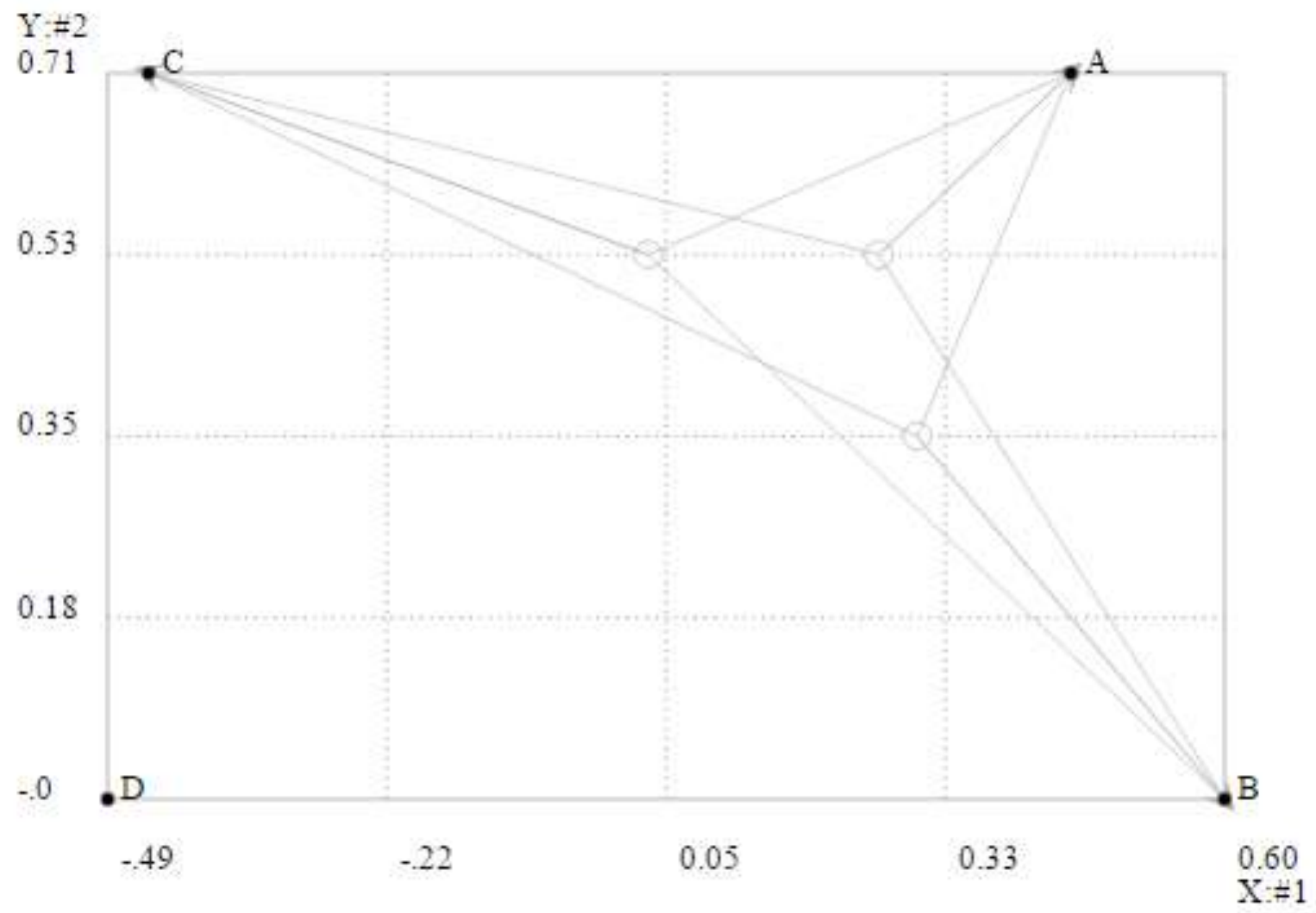
$$[1] \quad Z_n = X_{np} W_p$$

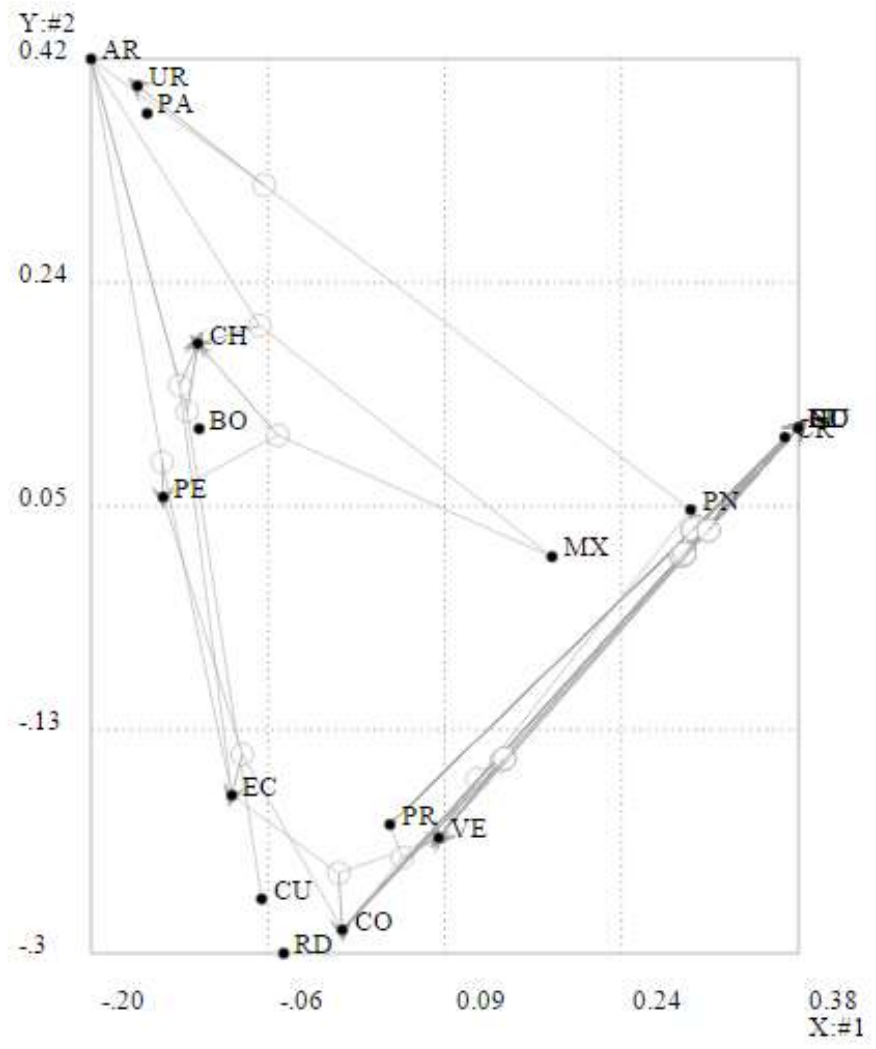
Varianza (V) del vector compuesto:

$$\begin{aligned} [2] \quad V &= (Z_n^T Z_n) / N \\ &= (X_{np} W_p)^T (X_{np} W_p) / N \quad \leftarrow [1] \\ &= W_p^T X_{np}^T X_{np} W_p / N \quad \leftarrow (A B)^T = B^T A \\ &= W_p^T (X_{np}^T X_{np} / N) W_p \quad \leftarrow N \text{ es escalar, movable} \\ &= W_p^T R_{pp} W_p \quad \leftarrow R_{pp} = X_{np}^T X_{np} / N \end{aligned}$$

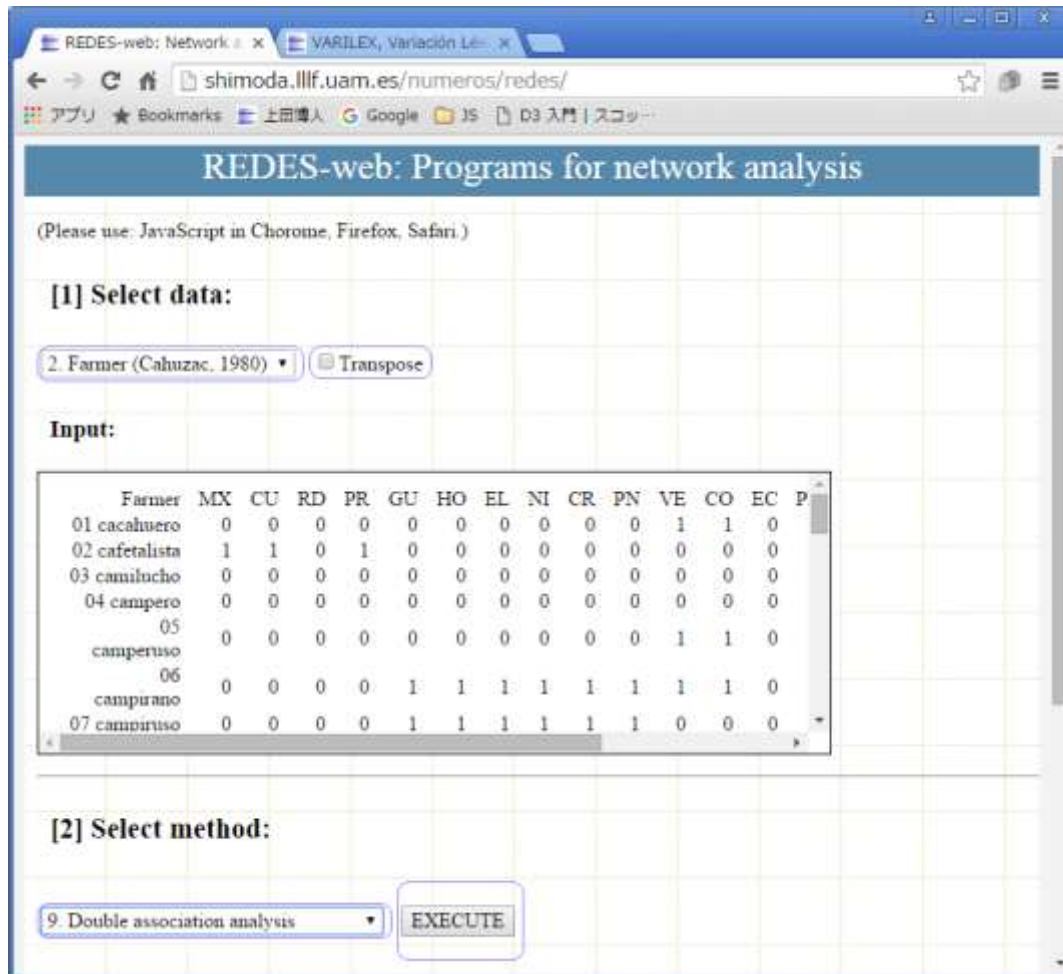
(...)

E.mat.	#1	#2	#3	#4
A	.449	.707	.317	-.445
B	.599	-.000	.230	.767
C	-.449	.707	-.317	.445
D	-.489	.000	.864	.122





(a) Input:



REDES-web: Programs for network analysis

(Please use: JavaScript in Chorome, Firefox, Safari.)

[1] Select data:

2. Farmer (Cahuzac, 1980)

Input:

Farmer	MX	CU	RD	PR	GU	HO	EL	NI	CR	PN	VE	CO	EC	P
01 cacahuero	0	0	0	0	0	0	0	0	0	0	1	1	0	
02 cafetalista	1	1	0	1	0	0	0	0	0	0	0	0	0	
03 camilocho	0	0	0	0	0	0	0	0	0	0	0	0	0	
04 campero	0	0	0	0	0	0	0	0	0	0	0	0	0	
05 camperuso	0	0	0	0	0	0	0	0	0	0	1	1	0	
06 campirano	0	0	0	0	1	1	1	1	1	1	1	1	0	
07 campiriso	0	0	0	0	1	1	1	1	1	1	0	0	0	

[2] Select method:

9. Double association analysis

(b) Seleccionar el método

The screenshot shows a web browser window with two tabs: "REDES-web: Network a" and "VARILEX, Variación Léxi". The address bar shows the URL "shimoda.llf.uam.es/numeros/redes/". The page content includes a dropdown menu for "2. Farmer (Cahuzac, 1980)" and a "Transpose" checkbox. Below this is an "Input:" section containing a table of data. Further down is a "[2] Select method:" section with a dropdown menu and an "EXECUTE" button.

2. Farmer (Cahuzac, 1980) Transpose

Input:

Farmer	MX	CU	RD	PR	GU	HO	EL	NI	CR	PN	VE	CO	EC	P
01 cacahuero	0	0	0	0	0	0	0	0	0	0	1	1	0	
02 cafetalista	1	1	0	1	0	0	0	0	0	0	0	0	0	
03 camilucho	0	0	0	0	0	0	0	0	0	0	0	0	0	
04 campero	0	0	0	0	0	0	0	0	0	0	0	0	0	
05 camperuso	0	0	0	0	0	0	0	0	0	0	1	1	0	
06 campirano	0	0	0	0	1	1	1	1	1	1	1	1	0	
07 campiruso	0	0	0	0	1	1	1	1	1	1	0	0	0	

[2] Select method:

0. Input matrix

- 0. Input matrix
- 1. Sum
- 2. Mean
- 3. Standard deviation
- 4. Standard score
- 5. Cooccurrence matrix (Product sum)
- 6. Correlation matrix
- 7. Principal component analysis
- 8. Single association analysis
- 9. Double association analysis

(c.1.) Output 1. Análisis de asociación

The screenshot shows a web browser window with three tabs: 'REDES-web: Network', 'LETRAS-web', and 'REDES-web: Network'. The address bar shows the file path: 'file:///C:/Users/ueda/Desktop/redes/index.html'. The page title is 'REDES-web: Programs for network analysis'. The interface is divided into two main sections: '[1] Select data: (Please use: JavaScript in Chorome, Firefox, Safari.)' and '[2] Select method:'. In the first section, a dropdown menu is set to '2. Farmer (Cahuzac, 1980)' and a 'Transpose' checkbox is checked. Below this is a table with 17 columns (Farmer, MX, CU, RD, PR, GU, HO, EL, NI, CR, PN, VE, CO, EC, PE, BO, CH) and 5 rows (01 cacahuero, 02 cafetalista, 03 camilucho, 04 campero, 05). In the second section, a dropdown menu is set to '9. Double association analysis' and an 'EXECUTE' button is visible. Below this is a table with 15 columns (#, L, R, LHS => RHS, L.f, R.f, L.p, R.p, Cooc., Sup., Cnf., Lift, Synt.) and 4 rows of results.

[1] Select data: (Please use: JavaScript in Chorome, Firefox, Safari.)

2. Farmer (Cahuzac, 1980) Transpose

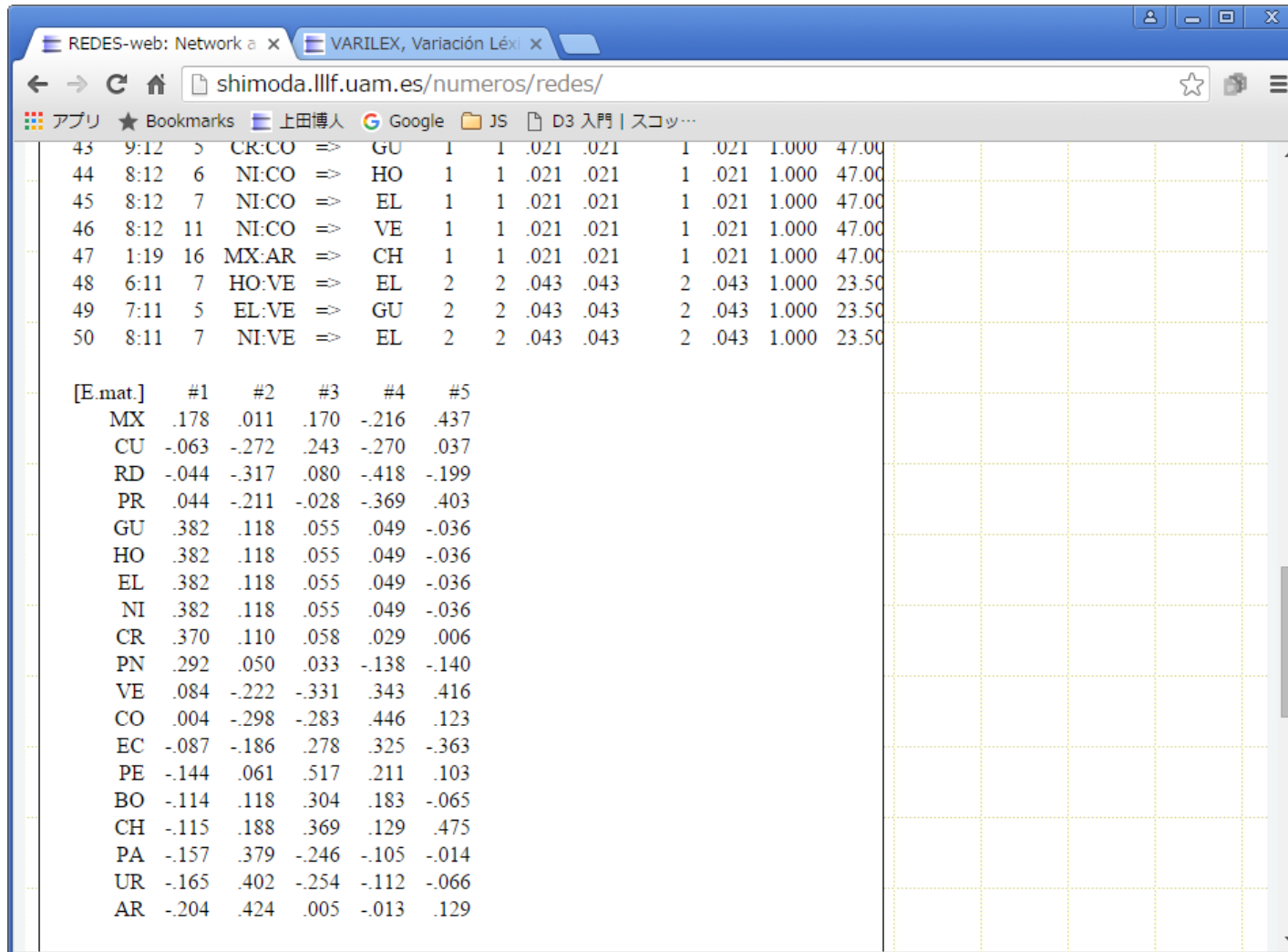
Farmer	MX	CU	RD	PR	GU	HO	EL	NI	CR	PN	VE	CO	EC	PE	BO	CH
01 cacahuero	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
02 cafetalista	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
03 camilucho	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
04 campero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
05	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0

[2] Select method:

9. Double association analysis

#	L	R	LHS => RHS	L.f	R.f	L.p	R.p	Cooc.	Sup.	Cnf.	Lift	Synt.
1	8:9	6	NI:CR => HO	7	7	.149	.149	7	.149	1.000	6.714	1.000
2	6:8	5	HO:NI => GU	7	7	.149	.149	7	.149	1.000	6.714	1.000
3	5:9	6	GU:CR => HO	7	7	.149	.149	7	.149	1.000	6.714	1.000
4	8:9	5	NI:CR => GU	7	7	.149	.149	7	.149	1.000	6.714	1.000

(c.2.) Output 2. Vectores eigen

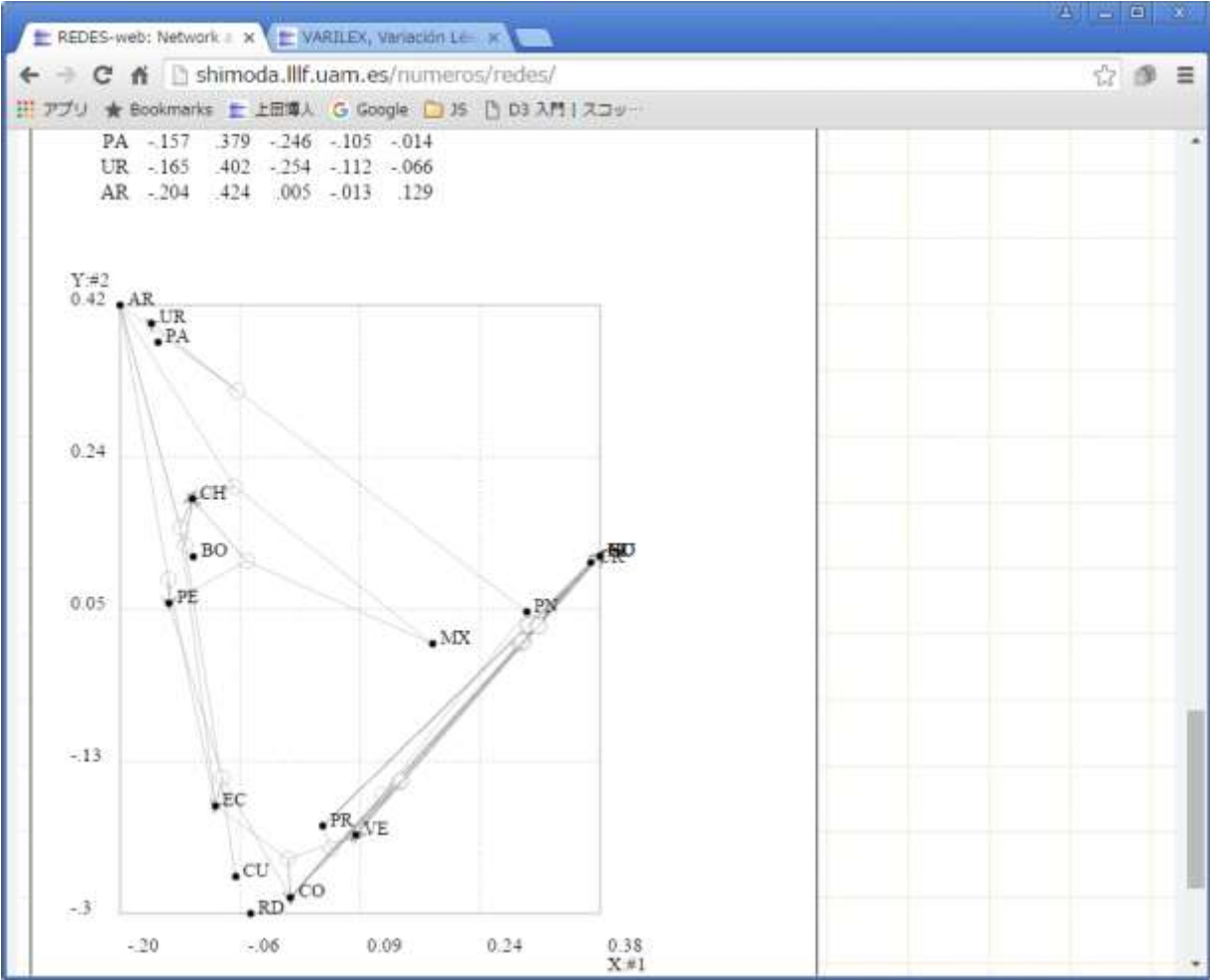


The screenshot shows a web browser window with two tabs: 'REDES-web: Network a' and 'VARILEX, Variación Léxi'. The address bar shows the URL 'shimoda.llf.uam.es/numeros/redes/'. The browser's bookmark bar includes 'アプリ', '★ Bookmarks', '上田博人', 'Google', 'JS', and 'D3 入門 | スコッ...'. The main content area displays a table of network data and a matrix of eigen vectors.

43	9:12	5	CR:CO =>	GU	1	1	.021	.021	1	.021	1.000	47.00
44	8:12	6	NI:CO =>	HO	1	1	.021	.021	1	.021	1.000	47.00
45	8:12	7	NI:CO =>	EL	1	1	.021	.021	1	.021	1.000	47.00
46	8:12	11	NI:CO =>	VE	1	1	.021	.021	1	.021	1.000	47.00
47	1:19	16	MX:AR =>	CH	1	1	.021	.021	1	.021	1.000	47.00
48	6:11	7	HO:VE =>	EL	2	2	.043	.043	2	.043	1.000	23.50
49	7:11	5	EL:VE =>	GU	2	2	.043	.043	2	.043	1.000	23.50
50	8:11	7	NI:VE =>	EL	2	2	.043	.043	2	.043	1.000	23.50

[E.mat.]	#1	#2	#3	#4	#5
MX	.178	.011	.170	-.216	.437
CU	-.063	-.272	.243	-.270	.037
RD	-.044	-.317	.080	-.418	-.199
PR	.044	-.211	-.028	-.369	.403
GU	.382	.118	.055	.049	-.036
HO	.382	.118	.055	.049	-.036
EL	.382	.118	.055	.049	-.036
NI	.382	.118	.055	.049	-.036
CR	.370	.110	.058	.029	.006
PN	.292	.050	.033	-.138	-.140
VE	.084	-.222	-.331	.343	.416
CO	.004	-.298	-.283	.446	.123
EC	-.087	-.186	.278	.325	-.363
PE	-.144	.061	.517	.211	.103
BO	-.114	.118	.304	.183	-.065
CH	-.115	.188	.369	.129	.475
PA	-.157	.379	-.246	-.105	-.014
UR	-.165	.402	-.254	-.112	-.066
AR	-.204	.424	.005	-.013	.129

(c.3.) Output 3. Redes



Bibliografía

Cahuzac, Philippe. (1980) "La División del español de América en zonas dialectales: Solución etnolingüística o semántico-dialectal." *Lingüística Española Actual*, 10, pp. 385-461.

Michael Hahsler, Bettina Grun, Kurt Hornik, Christian Buchta.
"Introduction to arules – A computational environment for mining association rules and frequent item sets"
<https://lyle.smu.edu/IDA/arules/> (24/2/2016)

Ueda, Hiroto. *Análisis de datos cuantitativos para estudios lingüísticos*.
<http://shimoda.llf.uam.es/numeros/redes/numeros-es.pdf>

5.3.2. Análisis de componentes principales

5.4. Análisis de asociación

¡FIN!

¡¡MUCHAS GRACIAS!

Google: Hiroto Ueda