

Assessing Geo-Location and Gender Information in Han Chinese Personal Names

**Bruce Brown and Deryle Lonsdale
Brigham Young University**

**Sixth Annual Family History Technology Workshop
Provo, Utah
March 9, 2006**

List #	Country	Total
1	Albania	20
2	Angola	1
3	Argentina	32
4	Armenia	7
5	Australia	19
6	Austria	4
7	Bangladesh	3
8	Barbados	1
9	Belgium	1
10	Bolivia	15
11	Brazil	93
12	British Virgin Isles	1
13	Bulgaria	13
14	Canada	247
15	Chile	38
16	China P.R.	171
17	Colombia	40
18	Costa Rica	1
19	Croatia	5
20	Czech Republic	4
21	Denmark	2
22	Dominican Republic	3
23	Ecuador	40
24	Egypt	3
25	El Salvador	11
26	Estonia	4
27	Fiji	5
28	Finland	8
29	France	17
30	French Polynesia	3
31	Georgia	5
32	Germany	34
33	Ghana	9
34	Guatemala	21
35	Haiti	7
36	Honduras	5
37	Hong Kong	32
38	Hungary	5
39	Iceland	4
40	India	34
41	Indonesia	6
42	Iran	2
43	Ireland	1
44	Israel	5
45	Italy	24
46	Ivory Coast	2
47	Jamaica	6
48	Japan	96
49	Jordan	24
50	Kazakhstan	2

List #	Country	Total
51	Kenya	8
52	Korea	194
53	Kuwait	1
54	Kyrgyzstan	1
55	Latvia	1
56	Lithuania	5
57	Macao	1
58	Madagascar	2
59	Malaysia	2
60	Mali	6
61	Mauritius	2
62	Mexico	235
63	Moldova	3
64	Mongolia	37
65	Morocco	4
66	Namibia	3
67	Nepal	35
68	Netherlands	5
69	New Zealand	14
70	Nicaragua	1
71	Niger	2
72	Nigeria	10
73	Norway	17
74	Pakistan	11
75	Panama	2
76	Paraguay	3
77	Peru	65
78	Philippines	6
79	Poland	7
80	Portugal	5
81	Romania	15
82	Russia	37
83	Sierra Leone	1
84	Singapore	24
85	Slovak Republic	4
86	Slovenia	1
87	South Africa	11
88	Spain	24
89	Sri Lanka	1
90	Sudan	1
91	Sweden	16
92	Switzerland	8
93	Syria	2
94	Taiwan	50
95	Tajikistan	1
96	Tanzania	1
97	Thailand	7
98	Tonga	2
99	Turkey	2
100	Uganda	7

List #	Country	Total
101	Ukraine	29
102	United Kingdom	62
103	Uruguay	11
104	Uzbekistan	2
105	Venezuela	22
106	Vietnam	20
107	West Bank	3
108	West Samoa	1
109	Yugoslavia	5
110	Zimbabwe	4

TOTAL = 2198

Winter Semester 2005

Count of Brigham Young University Students from Each of 110 Nations (Winter Semester, 2005)

List #	Country	Total
1	Albania	34
2	Argentina	713
3	Armenia	9
4	Asia North	1
5	Australia	186
6	Austria	11
7	Baltic	62
8	Baltic States	10
9	Belgium	137
10	Bolivia	120
11	Brazil	1372
12	Bulgaria	62
13	Cambodia	30
14	Canada	357
15	Cape Verde Praia	4
16	Chile	619
17	China Hong Kong	107
18	Colombia	66
19	Costa Rica	56
20	Croatia	37
21	Czech Republic	51
22	Denmark	46
23	Dominican Republic	180
24	Ecuador	212
25	El Salvador	71
26	England	237
27	Fiji	28
28	Finland	41
29	France	199
30	Germany	377
31	Ghana	8
32	Greece	25
33	Guatemala	223
34	Haiti	15
35	Honduras	133
36	Hungary	74
37	India	15
38	Ireland	30
39	Italy	325
40	Ivory Coast	13
41	Jamaica	25
42	Japan	468
43	Kenya	24

List #	Country	Total
44	Korea	319
45	Madagascar	28
46	Mexico	587
47	Micronesia Guam	19
48	Mongolia Ulaanbaata	33
49	Netherlands	22
50	New Zealand	45
51	Nicaragua	45
52	Nigeria	3
53	Norway	52
54	Panama	41
55	Paraguay	106
56	Peru	222
57	Philippines	383
58	Poland	80
59	Portugal	169
60	Puerto Rico	68
61	Romania	80
62	Russia	436
63	Samoa	12
64	Scotland	28
65	Singapore	26
66	South Africa	60
67	Spain	408
68	Sweden	74
69	Switzerland	125
70	Tahiti	12
71	Taiwan	318
72	Thailand	86
73	Tonga	5
74	Ukraine	170
75	Uruguay	129
76	Venezuela	246
77	West Indies	45
78	Zimbabwe	10

TOTAL = 10252

5387
15639

Fall Semester 2004

Count of Brigham Young University Students Who Have Served Missions in Various Foreign Nations (Fall Semester, 2004)

ten additional nations

Study 1. Pilot Study of Subjective Judgments

- **Purpose: To identify subjective collateral information in Han Chinese personal names.**
- **Six native Chinese informants provided judgments of (1) gender, (2) location, (3) ethnicity, (4) language/dialect, and (5) religion.**
- **There were **four parts** to the electronic questionnaire process.**

Part A. Categorization and confidence rating of 269 names.

Part B. Textual explanation of the reasons for the categorizations.

Part C. Ratings of 269 names on scales reflecting basis of judgment.

Study 1. Pilot Study

Part A. Categorization & Confidence

Chín

Example of form used to obtain categorization of Chinese names and ratings of confidence.



A. Name Categorization

B. Text

Chinese Name				Location	Location Confidence	Ethnic Group	Ethnic Group Confidence	Language of Home	Language Confidence
Mainland		Traditional							
1	安	朝俊		3. Hebe	2	Han 汉	1	Wu (Sh)	3
2	白	寿彝		B. North	1	Tibetan 藏	2	Zhongguo	2
	薄	一波							
	蔡	子民							
	岑	国荣							
	陈	春明							
	陈	昊苏							
	陈	雷							
	陈	其通							
10	陈	欣							
11	陈	又遭							
12	程	法光							

Record: 2 of 269

Gender Identification Accuracy

Six Native Chinese Informants:

Female vs Male

NCI 1	71%
NCI 2	69%
NCI 3	68%
NCI 4	65%
NCI 5	76%
NCI 6	71%

Study 1. Pilot Study

Part A. Categorization & Confidence

Signal Detection Theory Applied to Gender Identifications

Judged:		F					M					
		1	2	3	4	5	5	4	3	2	1	
Actual	F	14	1	1	0	0	7	2	2	2	33	62
	M&b	2	0	0	1	0	8	2	2	3	189	207
		16	1	1	1	0	15	4	4	5	222	269

		Test of known Female and "Other"					
		F		M			
F		16	46	62	0.26	0.74	1
M		3	204	207	0.01	0.99	1
		19	250	269	0.27	1.73	2

82% = percent correct

		conservative						ideal receiver						liberal														
frequencies:		F	M		F	M		F	M		F	M		F	M		F	M		F	M							
Actual	F	14	48	62	15	47	62	16	46	62	16	46	62	23	39	62	25	37	62	27	35	62	29	33	62			
	M&b	2	205	207	2	205	207	2	205	207	3	204	207	11	196	207	13	194	207	15	192	207	18	189	207			
		16	253	269	17	252	269	18	251	269	19	250	269	34	235	269	38	231	269	42	227	269	47	222	269			
probabilities:		F	M		F	M		F	M		F	M		F	M		F	M		F	M		F	M				
Actual	F	0.23	0.77	1	0.24	0.76	1	0.26	0.74	1	0.26	0.74	1	0.37	0.63	1	0.4	0.6	1	0.44	0.56	1	0.47	0.53	1			
	M&b	0.01	0.99	1	0.01	0.99	1	0.01	0.99	1	0.01	0.99	1	0.05	0.95	1	0.06	0.94	1	0.07	0.93	1	0.09	0.91	1			
		0.24	1.76	2	0.25	1.75	2	0.27	1.73	2	0.27	1.73	2	0.42	1.58	2	0.47	1.53	2	0.51	1.49	2	0.55	1.45	2			
		0.75	d' = 2.34	1.59	0.7	d' = 2.34	1.64	0.65	d' = 2.34	1.69	0.65	d' = 2.18	1.53	0.65	d' = 2.18	1.53	0.33	d' = 1.62	1.29	0.25	d' = 1.53	1.29	0.16	d' = 1.46	1.3	0.08	d' = 1.36	1.28

A *Signal Detection Theory (TSD)* paradigm was used to evaluate the accuracy and the confidence level of native Chinese informants in identifying gender from the 269 names.

The d-prime statistics are stable across confidence boundaries and also similar across the six native Chinese informants.

Location Identifications

Surprisingly, native Chinese informants were able to identify location from the names 20% or better, well beyond the chance level.

Six Native Chinese Informants:	Number of Judgments	Percent Correct	Normal Deviate	Probability	Chance of Guess
NCI 1	235	19.6%	$z = 4.89$	.0000005	five in ten million
NCI 2	229	21.8%	$z = 5.97$	.000000001	one in a billion
NCI 3	11	54.5%	$z = 4.92$	.0000004	four in ten million
NCI 4	15	26.7%	$z = 2.15$	.0157122	1.6 in a hundred
NCI 5	0				
NCI 6	0				

Study 1. Pilot Study

Part A. Categorization & Confidence

Example Accuracy Matrix for Identification of Location

Actual:	Judged										
	A.Northeast	B.North	C.Northwest	D.East	E.Central	F.South	G.Southwest	II.Taiwan	III.Hong Kong	IV.Singapore	
A.Northeast	0	9	3	5	1	1	0	1	0	0	0.0%
B.North	6	26	4	13	6	6	3	8	0	0	36.1%
C.Northwest	0	11	2	6	3	2	3	2	0	0	6.9%
D.East	6	5	1	10	4	2	4	8	0	0	25.0%
E.Central	2	9	1	3	2	2	0	1	0	0	10.0%
F.South	0	2	1	1	6	7	2	2	0	0	33.3%
G.Southwest	0	7	0	1	1	2	2	1	0	0	14.3%
II.Taiwan	0	4	0	0	0	1	2	1	1	0	11.1%
III.Hong Kong	0	1	0	1	0	1	0	0	0	0	0.0%
IV.Singapore	0	0	0	0	0	0	0	1	0	0	0.0%
	14	74	12	40	23	24	16	25	1	0	229
	0.0%	35.1%	16.7%	25.0%	8.7%	29.2%	12.5%	4.0%	0.0%	0.0%	21.8%
		north	northwest	east		south					

$$z = (.218 - .10) / \sqrt{(.1) * (.9) / 229} = 5.97$$

$$\text{probability} = .000000001$$

one in a billion probability by chance

Study 1. Pilot Study

Part B. Textual Explanation

Exit

	Chinese Name				Meaning of Given Name/ Comments on Given Name or Surname	Location		Ethnic		Language			
	Mainland	Traditional				L	R	Reason	E	R	Reason	L	R
268	张	川田	張			E	2		汉	1		Zh	1
269	张	五岳	張			B	2		汉	1		Zh	1
1	安	朝俊				A	2		汉	3		Zh	1
2	白	寿彝				G	2		回	2		OT	3
3	薄	一波				B	1		汉	1		Zh	1
4	蔡	子民				II	2		汉	1		MN	1
5	岑	国荣				III	2		汉	1		Yu	1
6	陈	春明				D	2		汉	1		Zh	1

Record: 1 of 269

A. Name Categorization

B. Textual Reasons

Example of form used to obtain textual commentary with respect to name properties that help with categorization.

Part B of the electronic questionnaire obtained textual commentary on the basis by which names were categorized.

The results of this qualitative aspect of the study were used to create rating scales for quantitative classification of the names in Part C.

Part C. Analysis of Rating Scales

Example of form used to obtain ratings of names, to quantify the basis on which they are categorized.

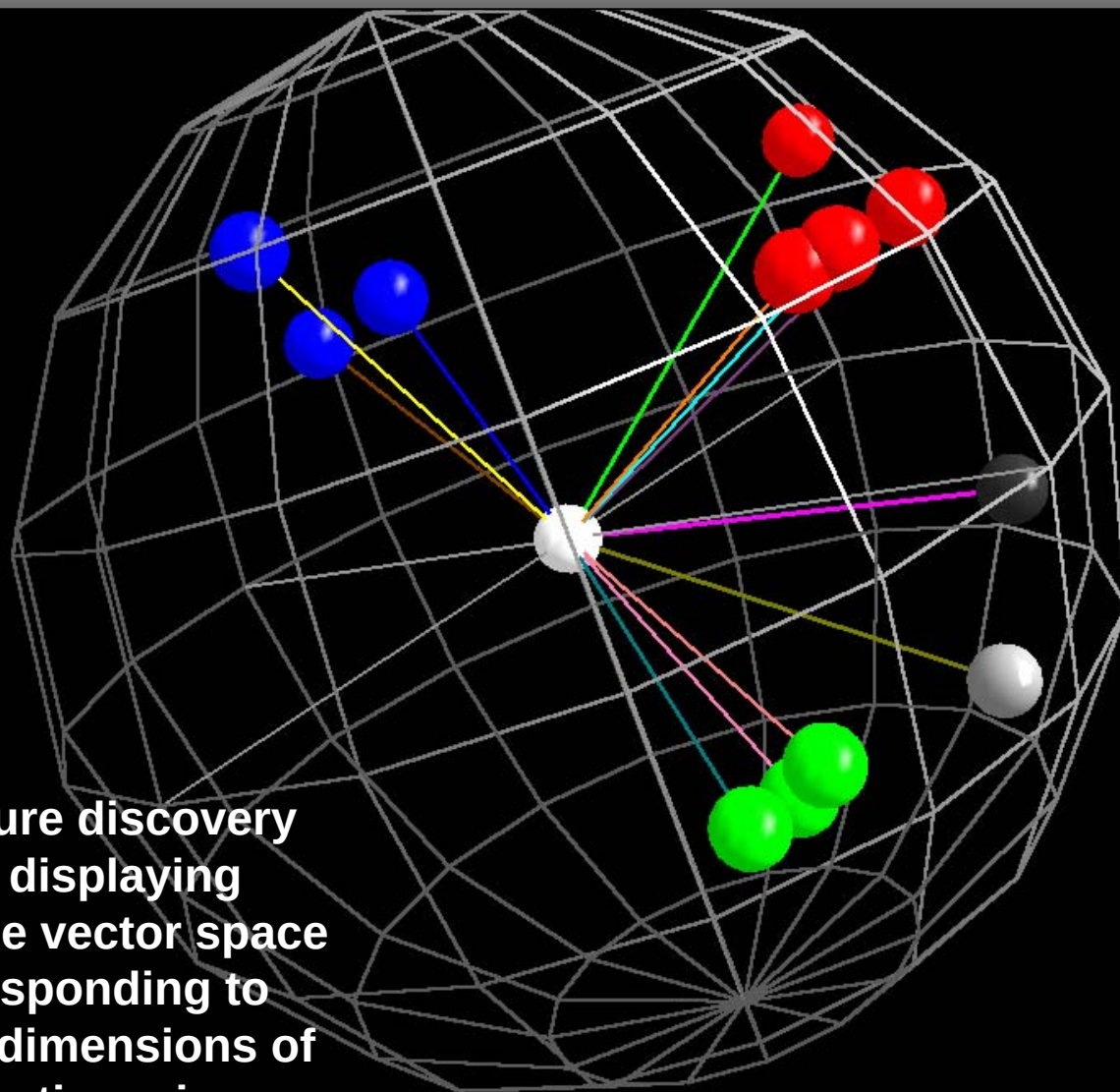
D. Combined Decision

vectors.xml
FactorA.xml
FactorB.xml
FactorAsp.xml
FactorBsp.xml

Study 1. Pilot Study

Part C. Analysis of Rating Scales

Structure discovery
tool displaying
possible vector space
corresponding to
eleven dimensions of
onomastic variance.

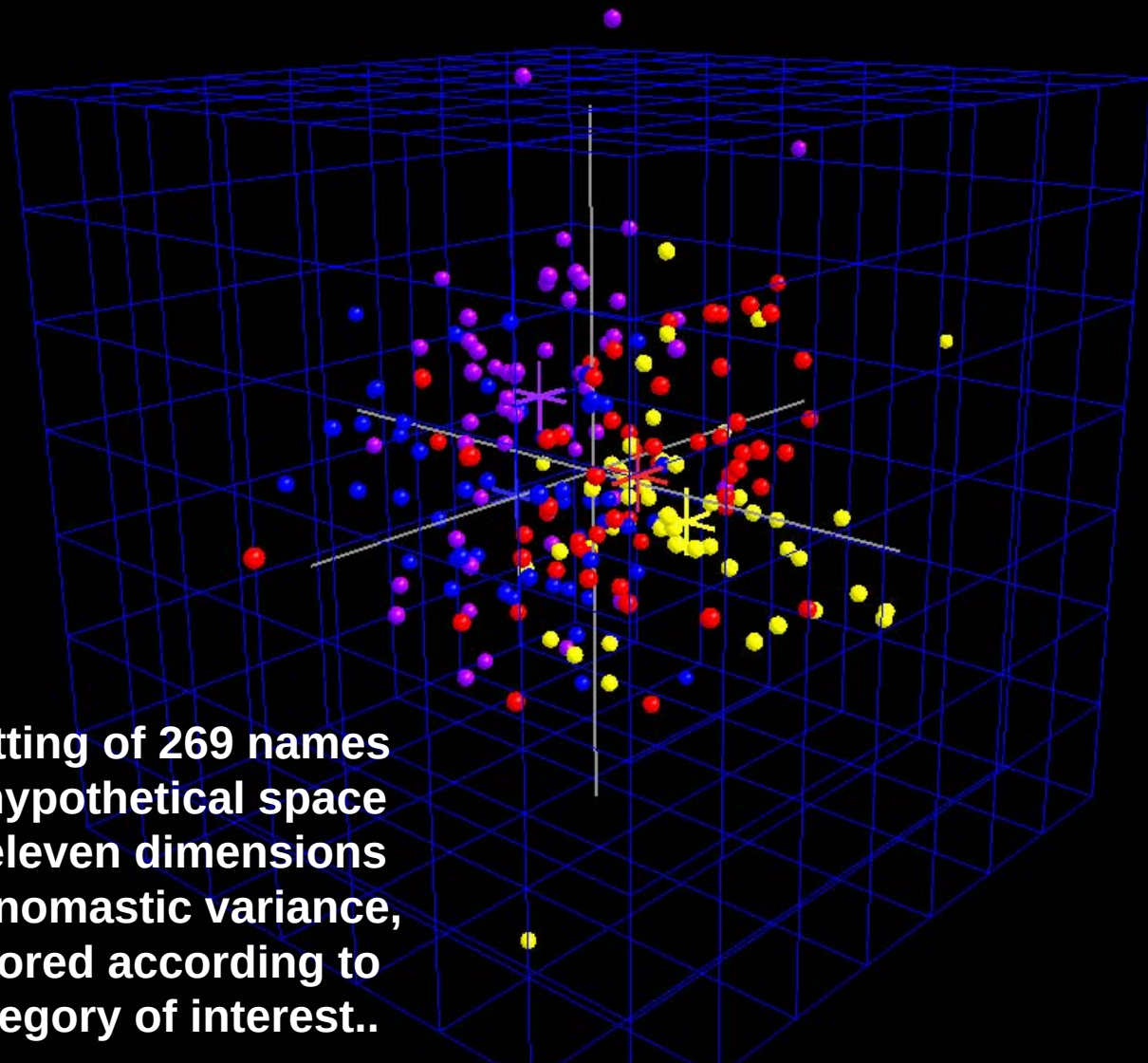


vectors.xml
FactorA.xml
FactorB.xml
FactorAsp.xml
FactorBsp.xml

Study 1. Pilot Study

Part C. Analysis of Rating Scales

**Plotting of 269 names
In hypothetical space
of eleven dimensions,
colored according to
category of interest..**



Label

A1B1

Study 3. Statistical Analysis and Comparison

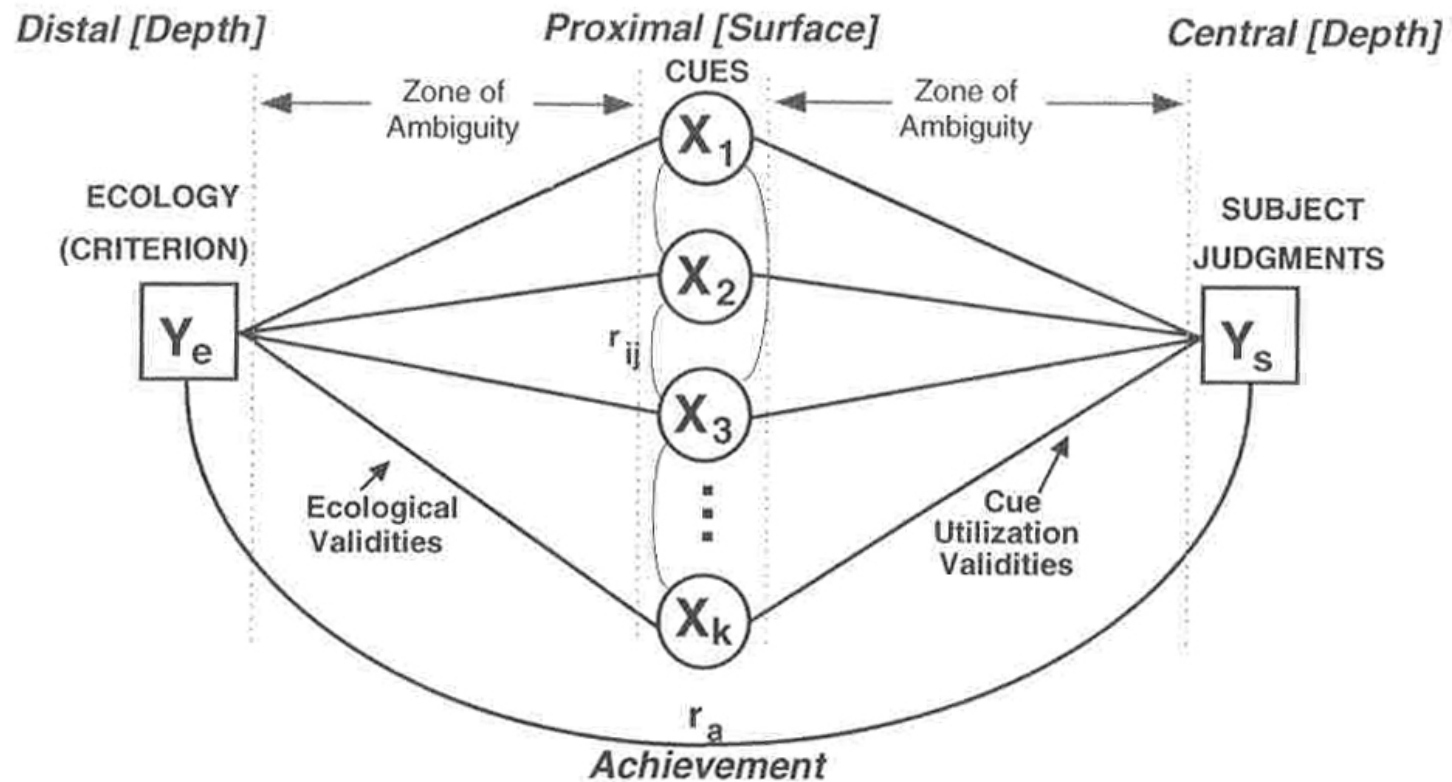
Initial work: develop analytical tools to provide precise comparisons of the accuracy of onomastic categorization.

This kind of precise analysis lends itself well to making cross-language and cross-cultural onomastic comparisons.

One particularly useful analytical tool for these purposes is the Brunswick Lens Model.

Study 3. Statistical Analysis and Comparison

The Lens Model of Proximal Cues as the mediators of accurate subjective judgments:



The Lens Model Equation:

$$r_a = GR_e R_s + C \sqrt{1 - R_e^2} \sqrt{1 - R_s^2}$$

- r_a = Correlation of the subjects' judgments with the distal variable
- G = Correlation of predicted scores from the two models
- R_e = Multiple correlation of the distal variable and cues
- R_s = Multiple correlation of subjects' judgments and cues
- C = Correlation of the residuals from the two models

Figure 1. Cross Tabulation of the Thirty-Four Most Common Han Names in the 2180 Dataset Crossed with Geo-Location, the Thirty-Two Provinces of China

		northeast										east				south																
		a1 Heilongjiang	a2 Jilin	a3 Liaoning	b1m Beijing	b2m Tianjin	b3 Hebei	b4 Shandong	b5 Henan	b6 Shanxi	c1 Shaanxi	c2 Gansu	c3ar NeiMonggol	c4ar Ningxia	c5 Qinghai	d1 Zhejiang	d2 Jiangsu	d3m Shanghai	d4 Anhui	e1 Hubei	e2 Hunan	e3 Jiangxi	f1 Guangdong	f2 arGuangxi	f3 Fujian	g1 Sichuan	g2m Chongqing	g3 Guizhou	g4 Yunnan	h1 Hainan	i1 Taiwan	
FN3	何	0	1	2	3	2	8	3	2	1	6	0	0	2	0	5	3	1	2	10	6	0	12	0	5	11	0	0	0	1	0	86
FN4	侯	0	0	2	2	0	5	3	3	4	1	0	1	0	0	0	3	1	1	1	1	0	3	0	1	2	0	0	0	0	0	34
FN8	关	0	2	2	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	3	2	0	0	0	0	0	0	0	13
FN11	华	0	0	0	0	1	2	1	0	1	0	0	0	0	0	1	8	0	1	1	0	0	0	1	1	0	0	0	0	0	0	18
FN21	姜	2	1	2	1	1	1	11	0	1	0	0	0	0	0	0	5	1	0	1	1	0	0	0	0	0	0	0	0	0	0	28
FN23	孔	0	0	0	0	0	0	6	0	1	1	0	1	0	0	2	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	15
FN24	季	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	10
FN33	康	0	0	0	0	1	3	2	1	2	1	1	0	0	0	1	0	0	0	1	2	4	0	0	1	2	1	0	1	0	0	24
FN35	惠	0	1	0	0	0	1	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
FN40	景	0	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	5
FN42	柯	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0	0	1	0	1	0	3	0	1	0	0	0	0	0	0	10
FN46	江	0	0	0	0	0	0	3	0	1	2	0	0	0	0	2	8	1	2	3	3	4	0	1	3	2	0	0	0	0	0	35
FN47	洪	0	0	0	0	0	2	0	0	0	0	0	0	2	0	3	2	0	4	0	0	0	3	0	4	0	0	0	0	0	0	20
FN49	焦	0	0	0	0	0	4	0	2	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
FN50	甘	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	0	2	1	2	3	0	3	0	0	0	0	1	17
FN54	管	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
FN55	纪	0	0	2	0	0	1	5	0	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	12
FN60	耿	0	0	1	0	1	2	3	0	2	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	11
FN61	胡	1	0	0	1	0	9	2	3	7	7	0	1	0	0	12	11	7	7	9	12	6	1	0	2	10	0	1	0	1	0	110
FN65	葛	0	0	0	0	0	3	4	1	0	0	0	1	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	13
FN66	蒋	0	0	0	1	0	3	0	0	1	1	0	0	0	0	11	12	4	5	0	3	0	0	0	1	5	0	0	0	0	0	47
FN68	谷	0	0	2	0	0	4	3	0	1	0	0	0	0	0	1	1	0	1	0	4	0	0	0	0	1	0	0	0	0	0	18
FN70	贺	0	0	0	0	0	1	2	4	1	4	0	0	0	0	2	0	0	0	1	7	3	1	0	1	1	0	0	0	0	0	28
FN71	贾	1	1	2	2	0	9	2	2	5	2	1	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	32
FN74	郝	0	0	0	1	0	6	3	1	2	3	0	2	1	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	22
FN75	郭	1	2	3	1	2	18	7	9	12	4	1	0	1	0	5	6	1	1	5	1	1	10	1	2	1	0	1	3	0	2	101
FN76	金	4	8	3	1	1	0	3	2	2	0	1	1	2	0	9	11	5	4	1	2	3	1	1	1	0	0	0	0	0	0	66
FN78	霍	0	0	0	1	0	0	1	0	2	1	1	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	10
FN80	斯	0	0	0	0	0	1	0	2	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6
FN82	韩	2	1	4	0	2	11	7	2	7	2	0	0	2	4	2	5	2	0	3	0	0	2	1	1	1	0	1	0	1	0	63
FN83	顾	0	0	2	1	0	1	0	0	0	0	0	0	0	0	6	16	7	0	0	1	0	0	0	0	0	1	1	0	0	0	36
FN84	高	0	3	13	3	1	9	16	3	0	13	0	0	0	0	6	6	1	2	0	2	0	1	0	2	0	0	0	1	0	0	82
FN85	黄	1	1	2	1	3	6	3	2	4	3	1	0	1	0	9	10	5	6	16	8	21	24	11	22	9	1	1	3	1	2	177
FN87	龚	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	3	3	3	1	2	0	1	0	2	1	0	1	0	0	0	21

Figure 2. Geo-Location of the Han Names: The Thirty-two Provinces of China Grouped into Nine Regions

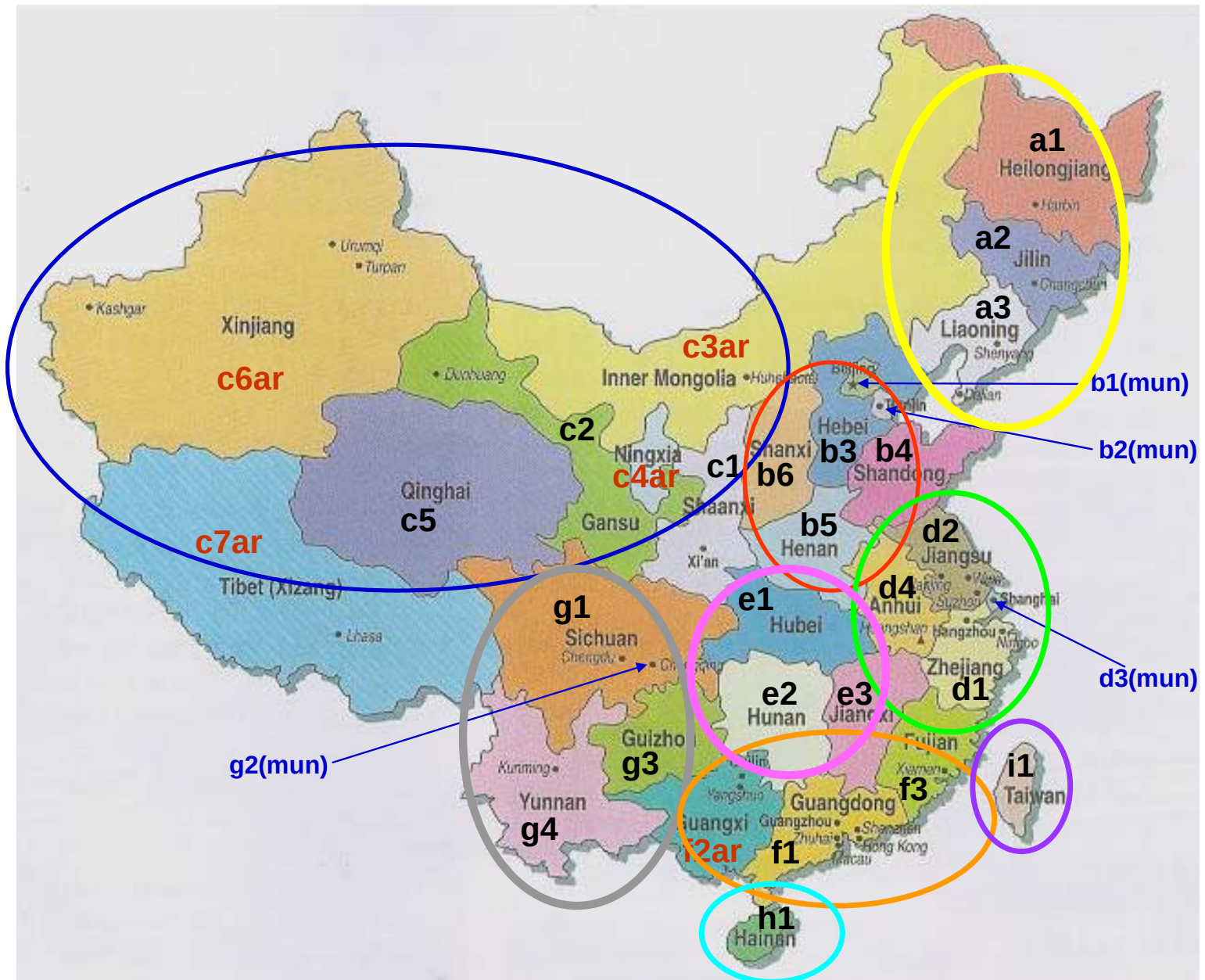


Figure 3. Metrika Vector Plot of Thirty Chinese Provinces in the Anthroponomastic Space of the Thirty-Two Most Common of the 2180 Names, Colored According to Region

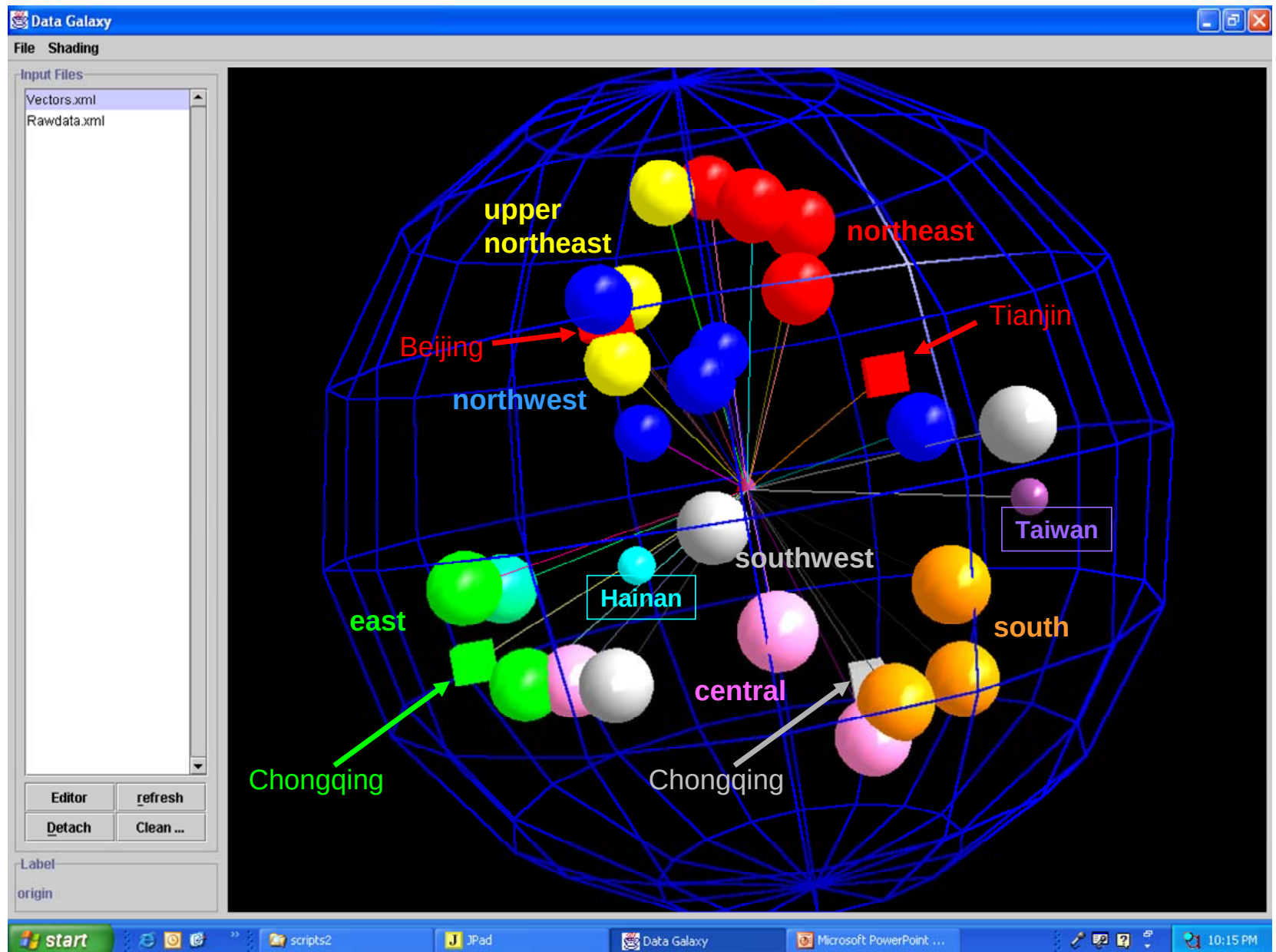
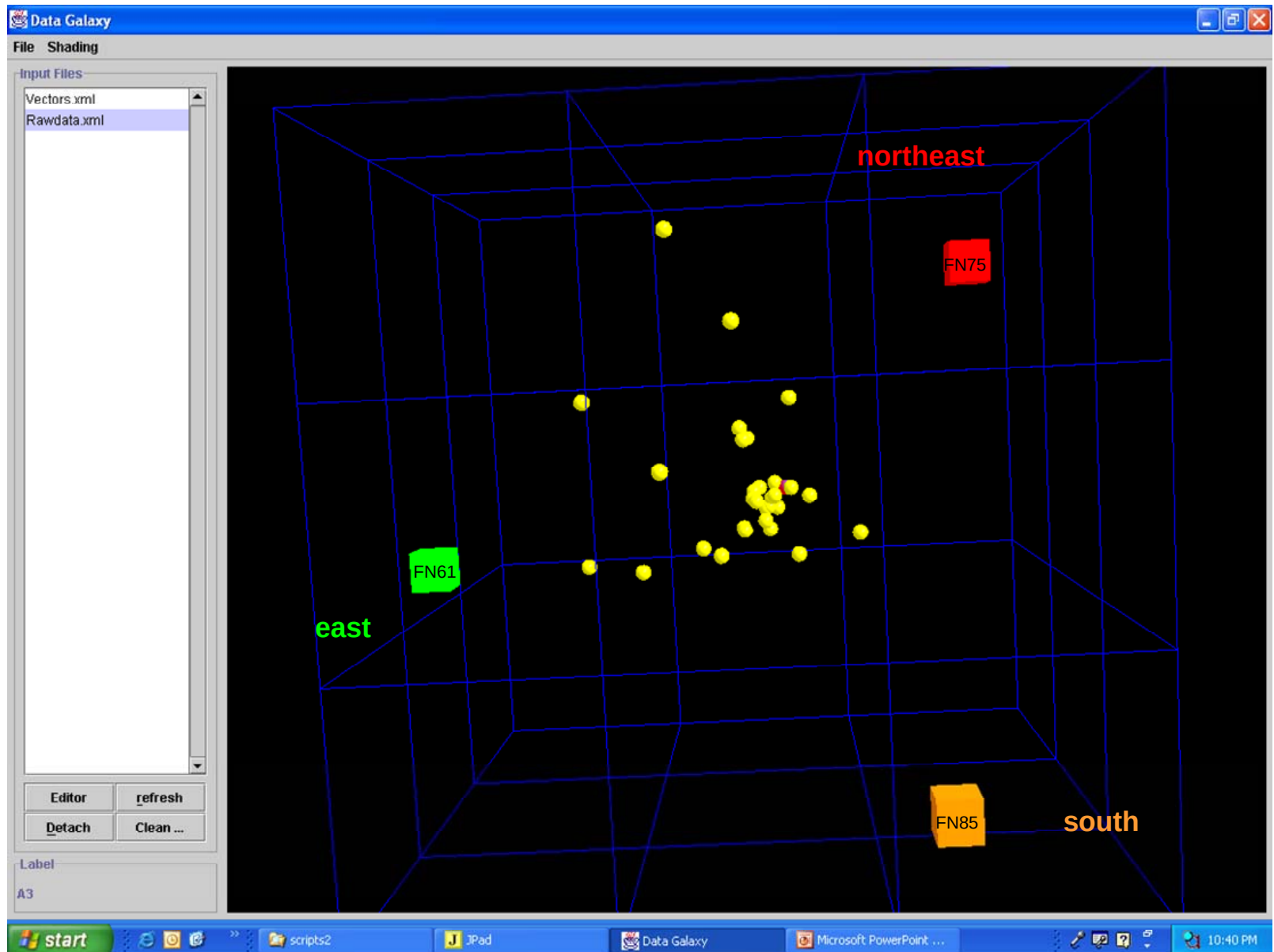


Figure 4. Metrika Datapoint Plot of the Thirty-Four Most Common Names, Plotted in the Same Anthroponomastic Space as Figure 3, with Three Names in Extreme Positions Labeled



Conclusions

- Gender: Chinese given names less informative than American given names.
- Geo-Location: Chinese family names well beyond chance level (contra experts' view).
- Similar studies are being conducted for six national groups: India, Korea, Jordan, Russia, Singapore, and Nepal.