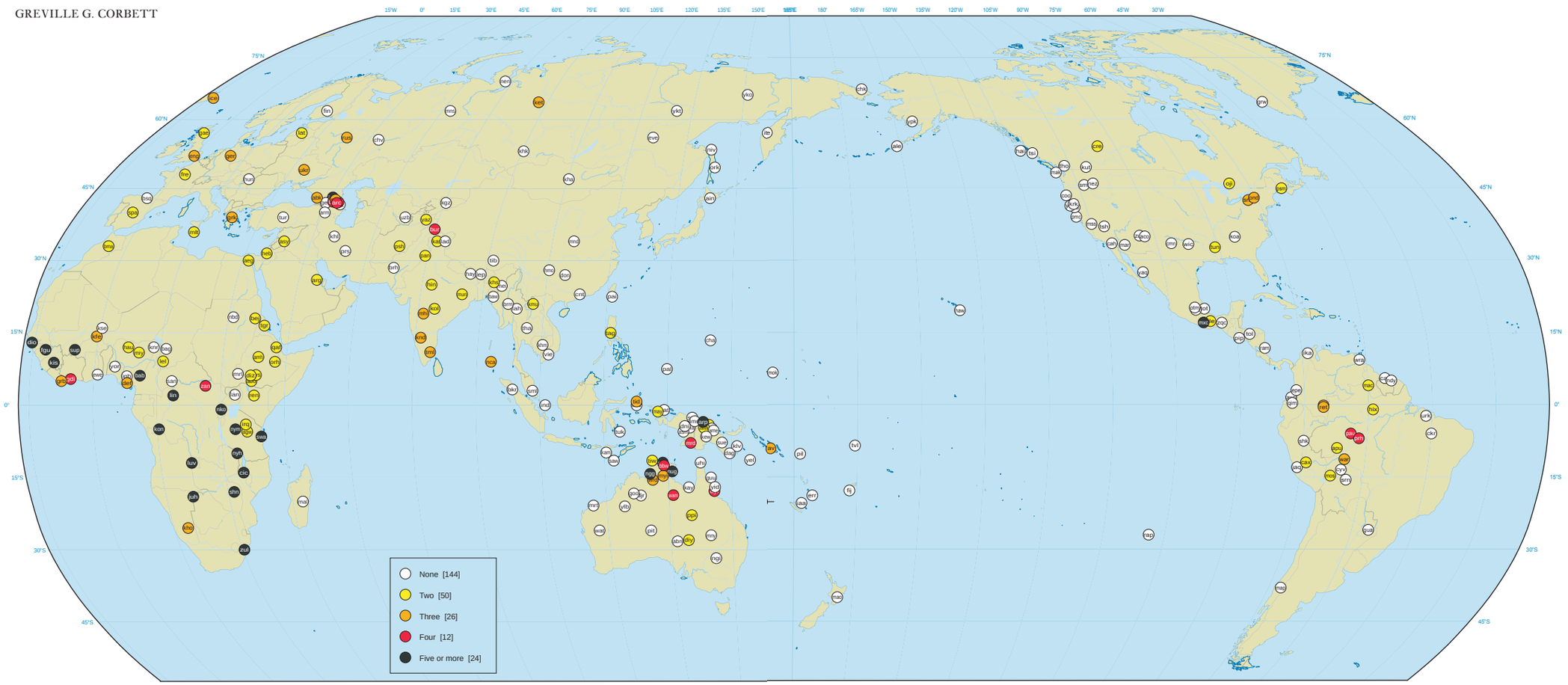
On the typological distribution of rare characteristics

Michael Cysouw
Max Planck Institute for Evolutionary Anthropology

World Atlas of Language Structures (WALS)

- 140 world-maps showing the distribution of linguistic features
- Information included about 2600 languages, though only a few hundred have a good coverage
- A total of 56,000 datapoints!

An example:



Haspelmath, Martin & Dryer, Matthew & Gil, David & Comrie, Bernard (eds.) 2005. *The World Atlas of Language Structures.* Oxford: Oxford University Press.

The basic idea

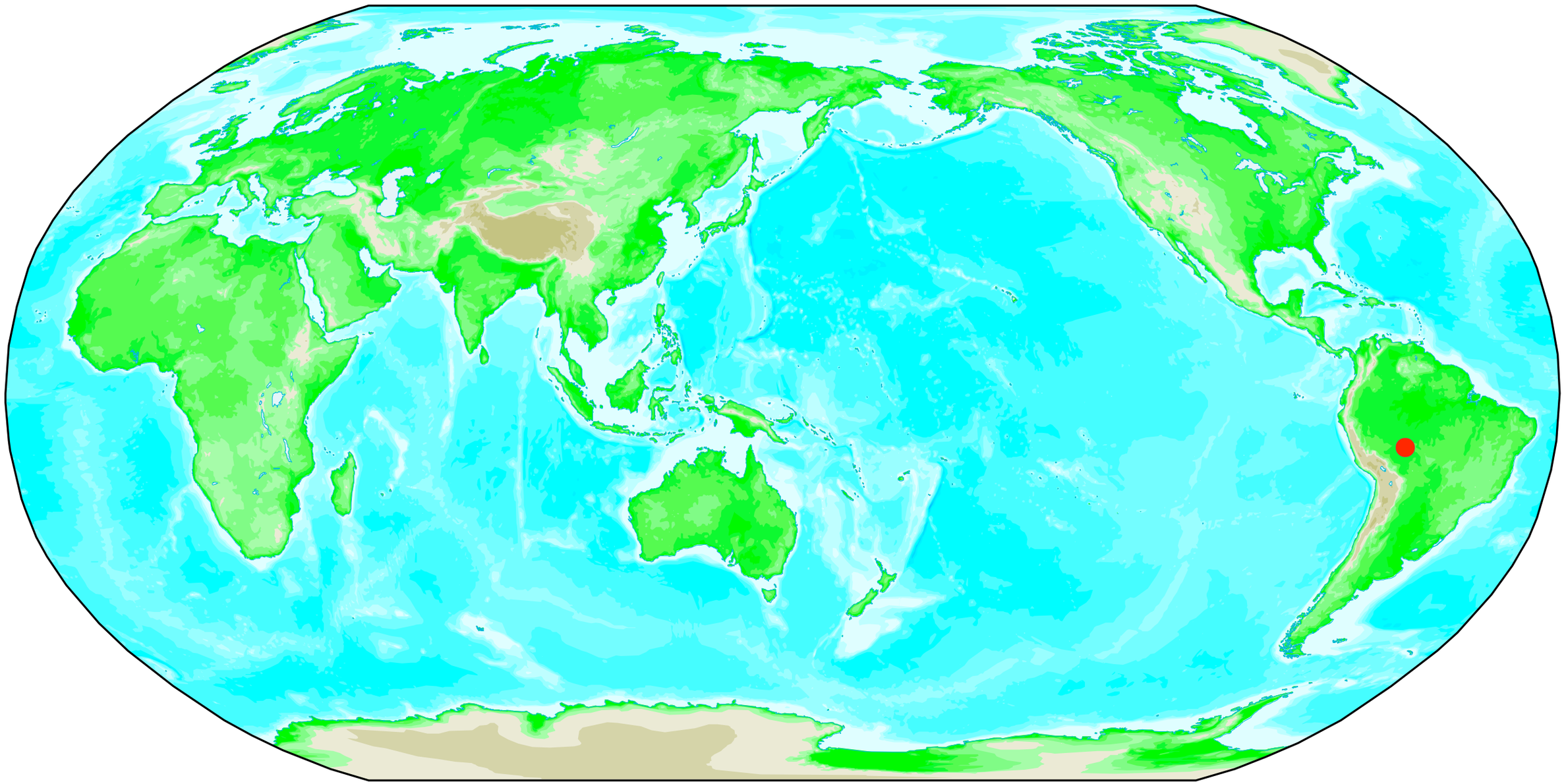
- Use the WALS data for 'holistic' typology
- Not look at the content of the features, but at their relative ubiquity
- Are there languages/families/areas that have more rare features than other?

And the winners are:

In the category:

‘Most Unusual Individual Language’

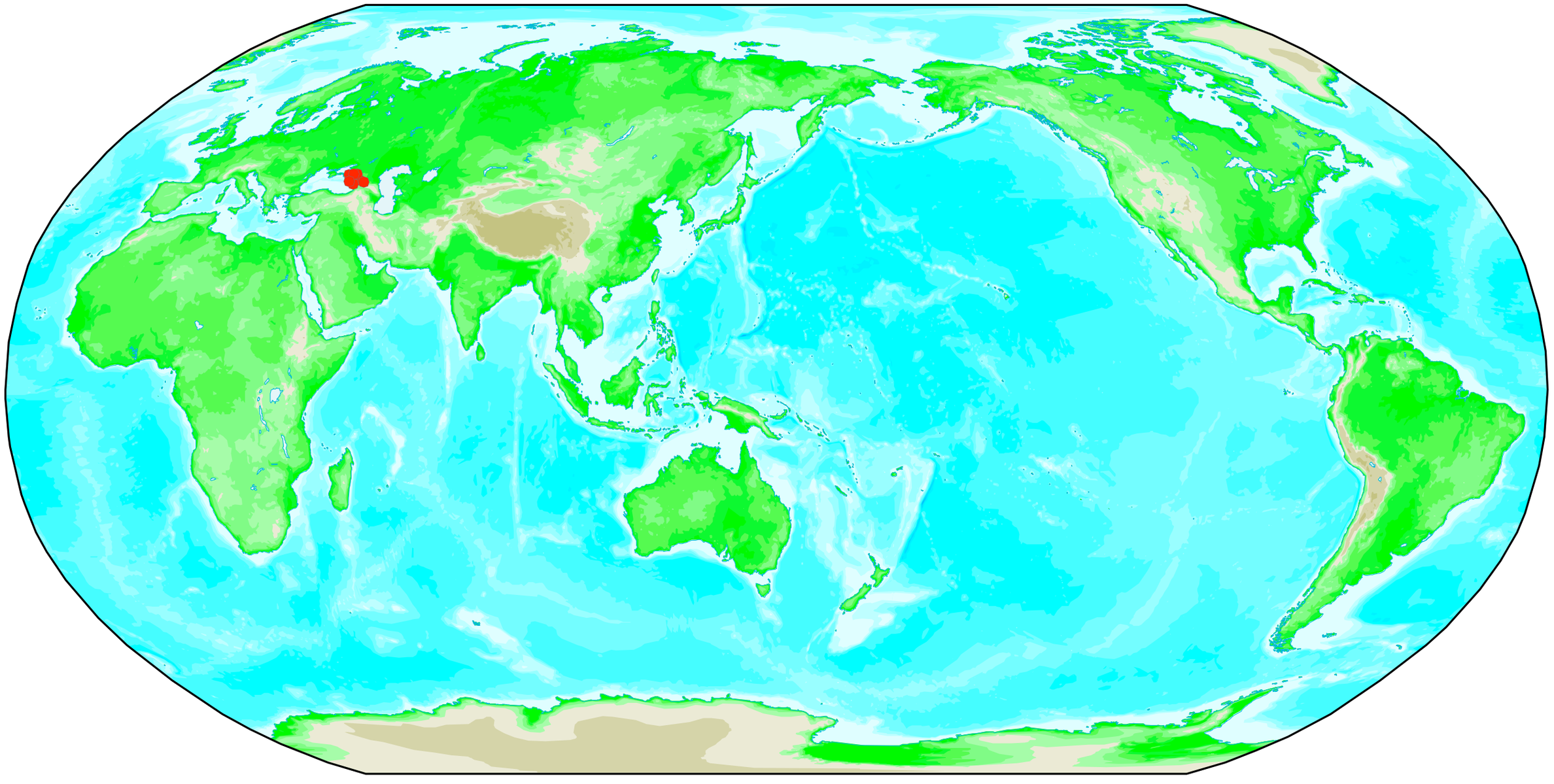
Wari’



In the category:

‘Most Unusual Genealogical Group’

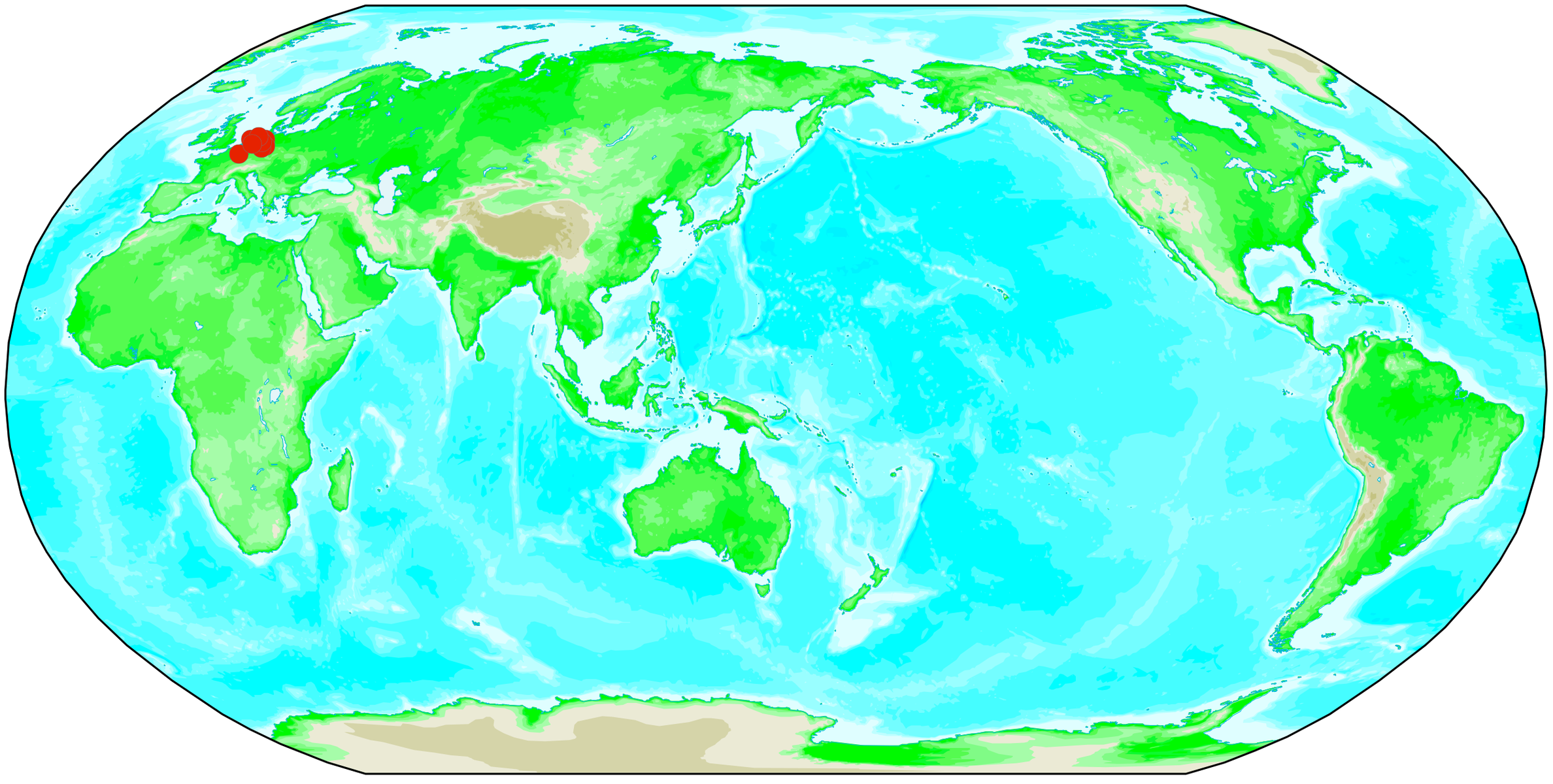
Northwest Caucasian



In the category:

‘Most Unusual Geographical Area’

Northwest Continental Europe



Rarity Index R_i

n = number of feature values

f_i = frequency of feature value i

f_{tot} = total number of languages included

$$R_{f_i} = n \cdot \frac{f_i}{f_{tot}}$$

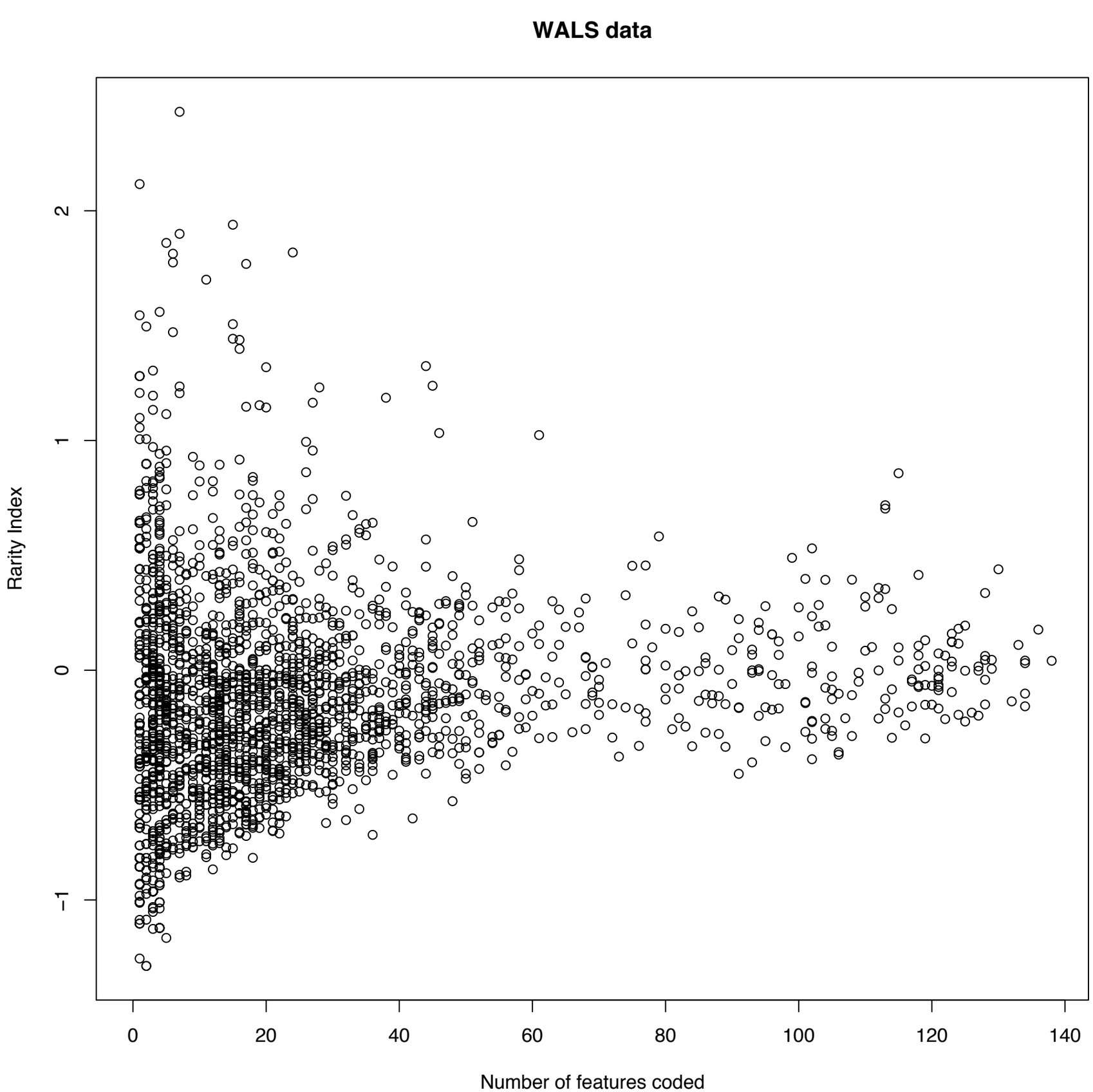
Inverse Index

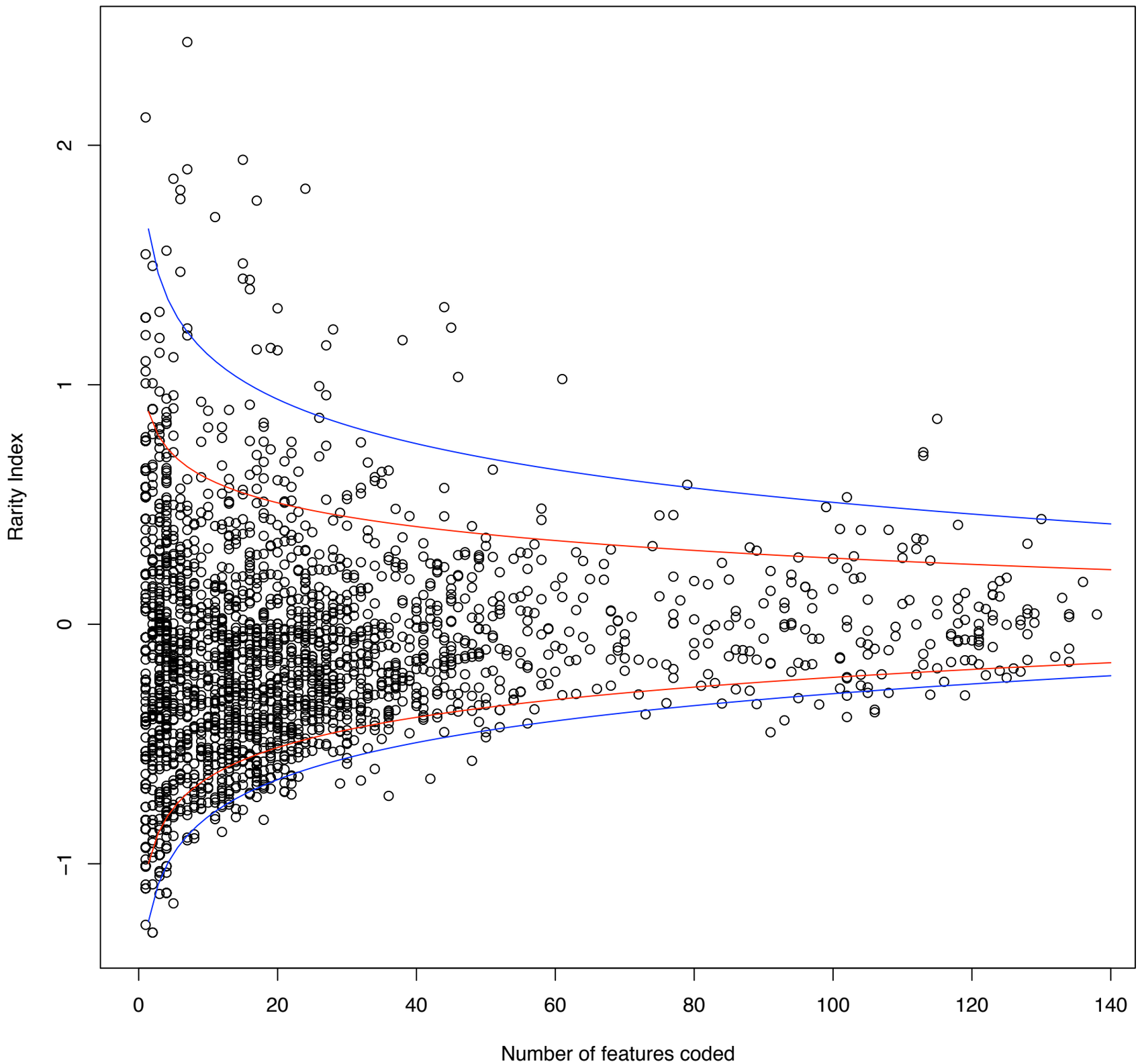
I used the inverse instead:

$$R_i = \frac{f_{tot}}{n \cdot f_i}$$

Because the mean of all R_i values is one:

$$\frac{\sum_{i=1}^{n} (R_i \cdot f_i)}{f_{tot}} = 1$$

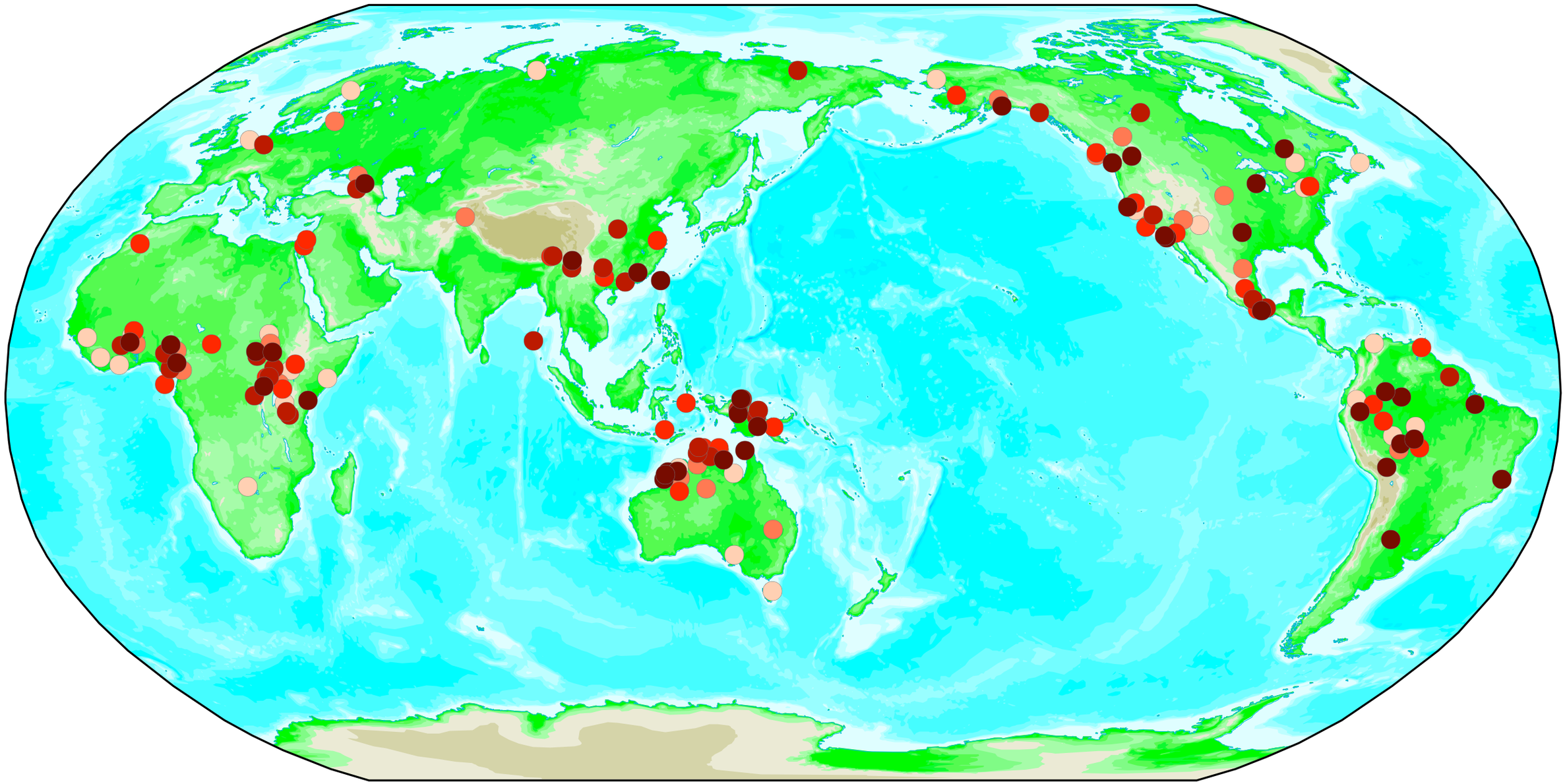

WALS data
Rarity Index
2
1
0
−1
0
20
40
60
80
100
120
140
Number of features coded


WALS data with 1% and 5% extremes
Rarity Index
2
1
0
−1
0
20
40
60
80
100
120
140
Number of features coded

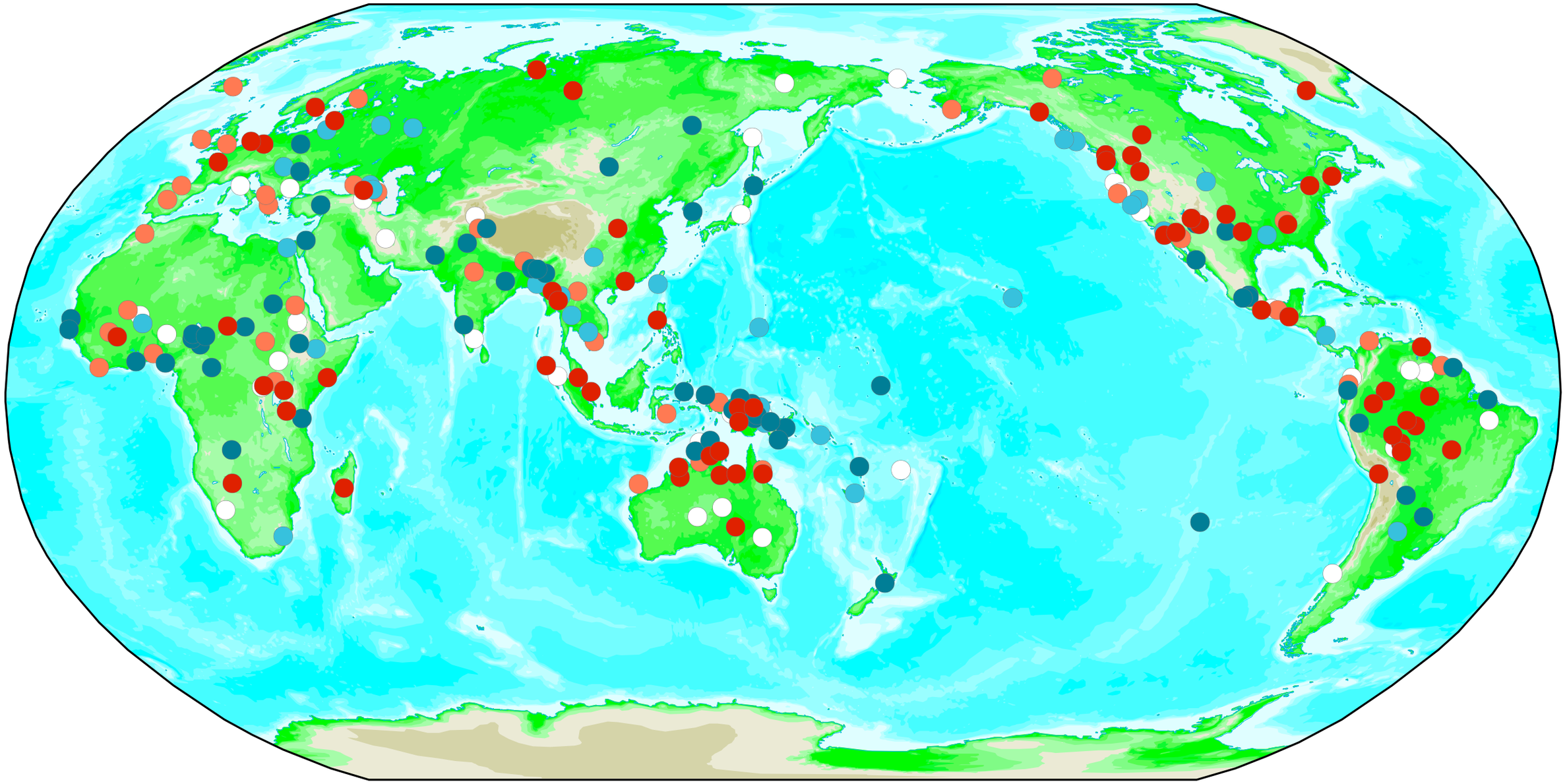
Highest *Mean Rarity*

Language	Family	Genus	Features Coded	Mean Rarity	%
Wari’	Chapacura-Wanhan	Chapacura-Wanhan	115	2.36	99.9
Dinka	Nilo-Saharan	Nilotic	45	3.45	99.9
Tiipay (Jamul)	Hokan	Yuman	44	3.76	99.9
Nuer	Nilo-Saharan	Nilotic	28	3.42	99.9
Karó (Arára)	Tupian	Tupi-Guaranì	24	6.16	99.9
Winnebago	Siouan	Siouan	7	11.37	99.9

Top 5% by *Mean Rarity*



All languages with more than 60 features coded for (red = rare, blue = common)



Rarity Index for a group

n = number of languages in a group

L_i = number of features coded for language i

$\%R_i$ = Relative position of Rarity Index for language i (in percentage)

Weighted mean of $\%R_i$ by number of features coded:

$$\frac{\sum_{i=1}^{n} \log(L_i) \cdot (\%R)_i}{\sum_{i=1}^{n} \log(L_i)}$$

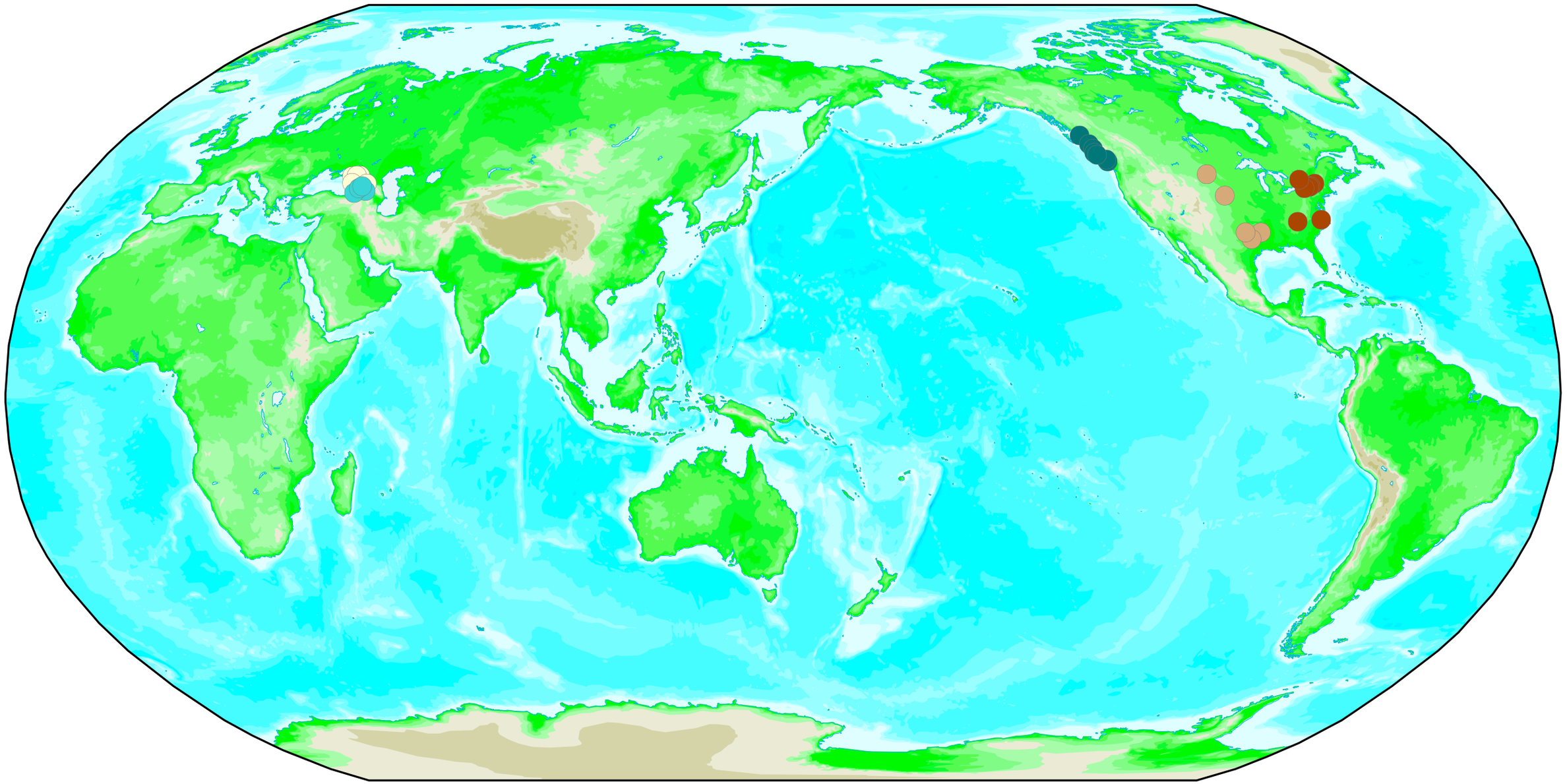
Genealogical groups

- Compute *Group Indices* for all Families and Genera as coded in the WALS
- Only groups with more than three members are shown, to be sure to get a group measure, and not an effect of an individual language

Top 5 Families

Family	Languages	%
NorthwestCaucasian	7	87.8
Kartvelian	4	83.7
Caddoan	5	82.2
Wakashan	7	80.2
Iroquoian	8	76.3

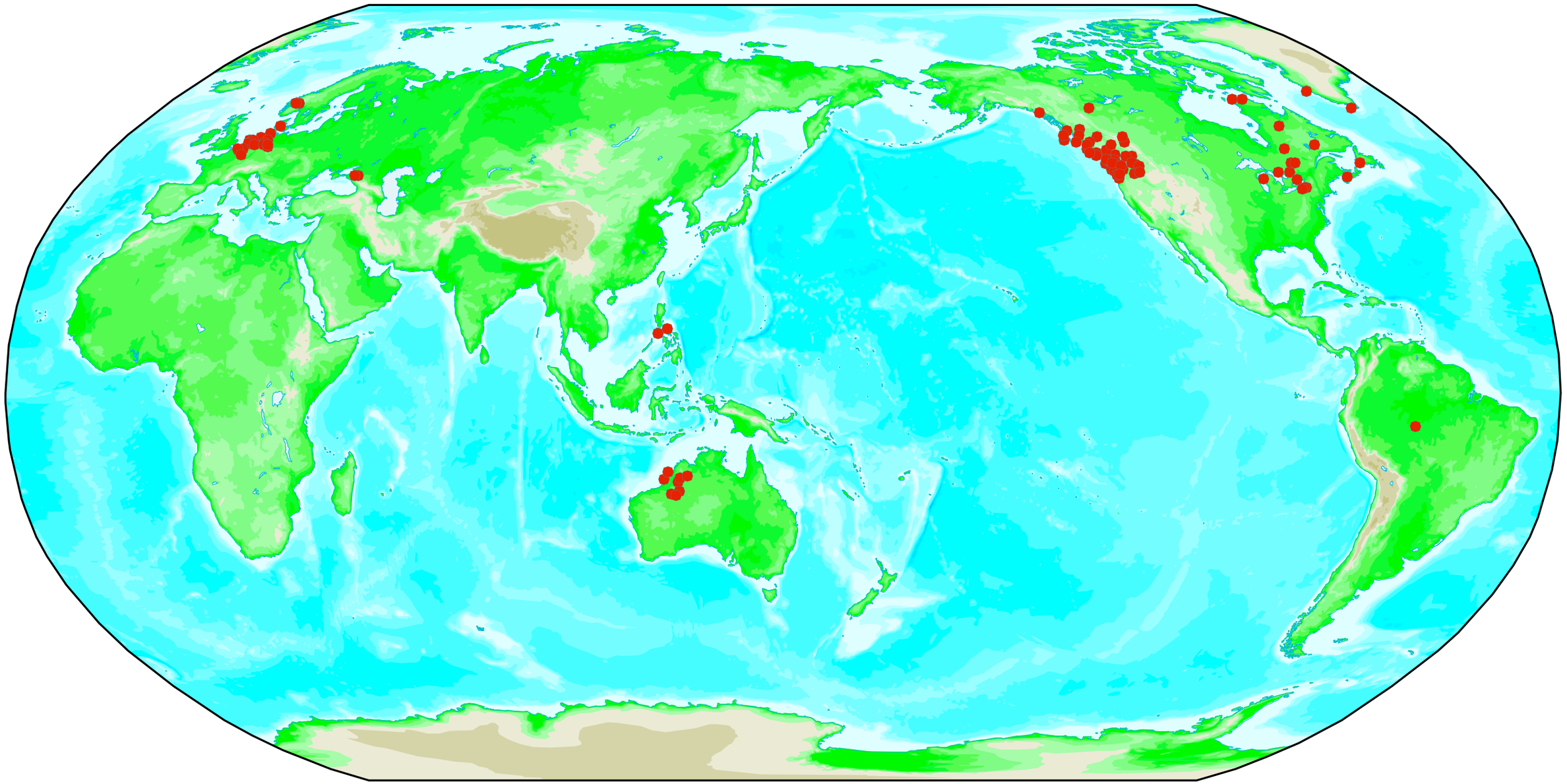
Top 5 Families



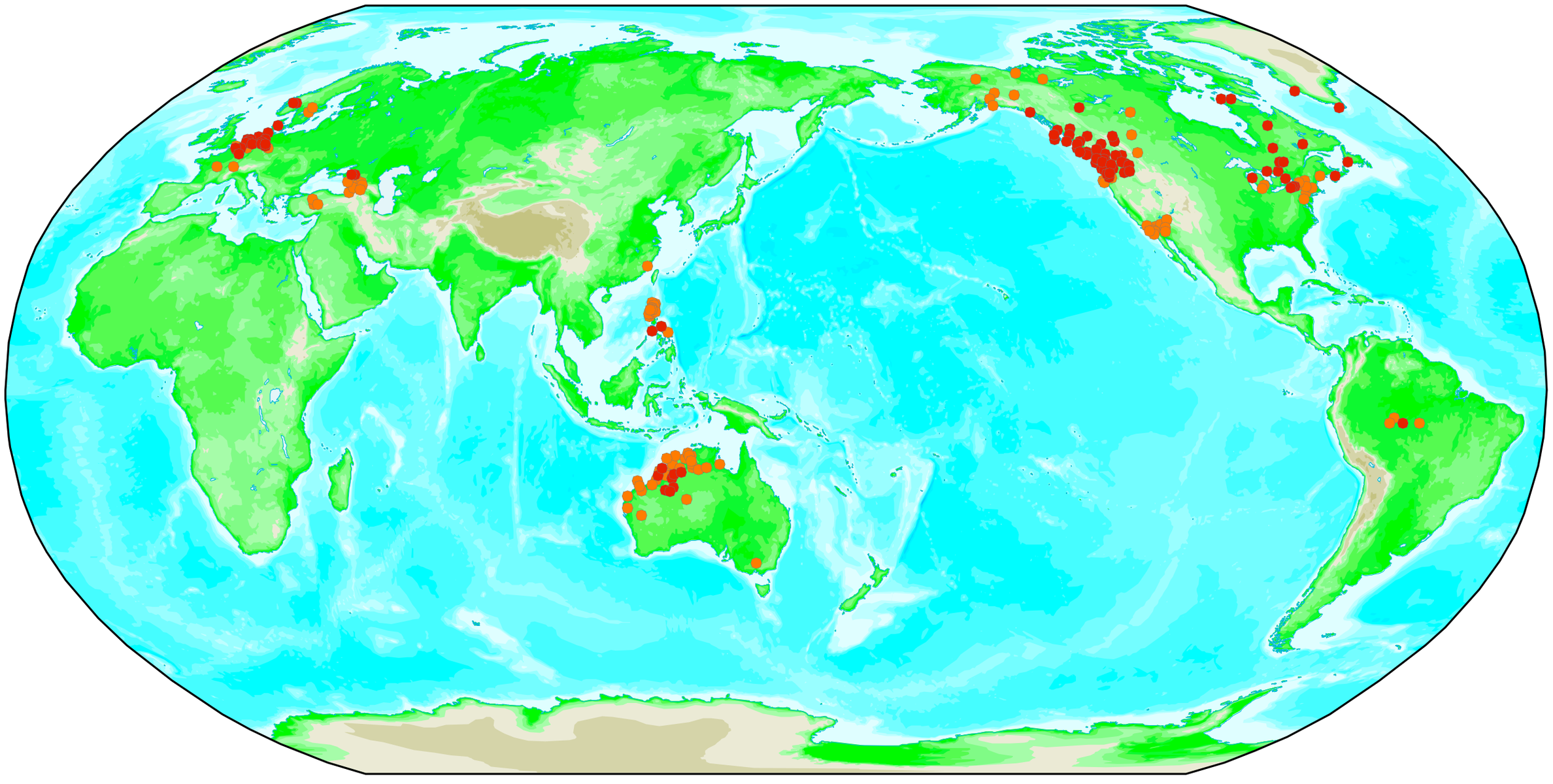
Areal groups

- For each language, take the 30 geographically nearest languages
- Compute *Group Indices* for the surrounding area of each language
- Such a measure should be definition be areally consistent, but it can indicate geographical centers of ‘rarity’

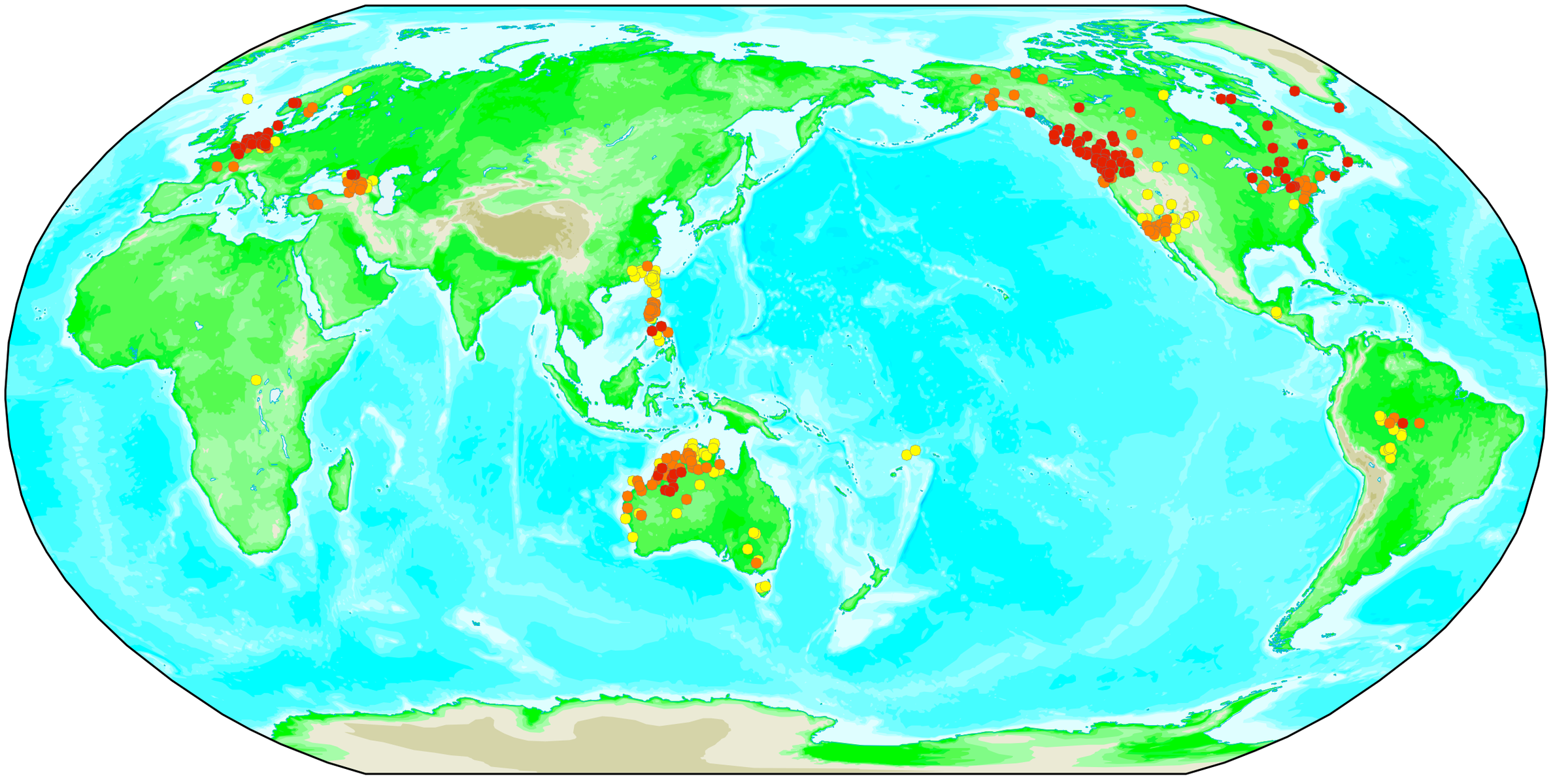
Top 100



Top 200

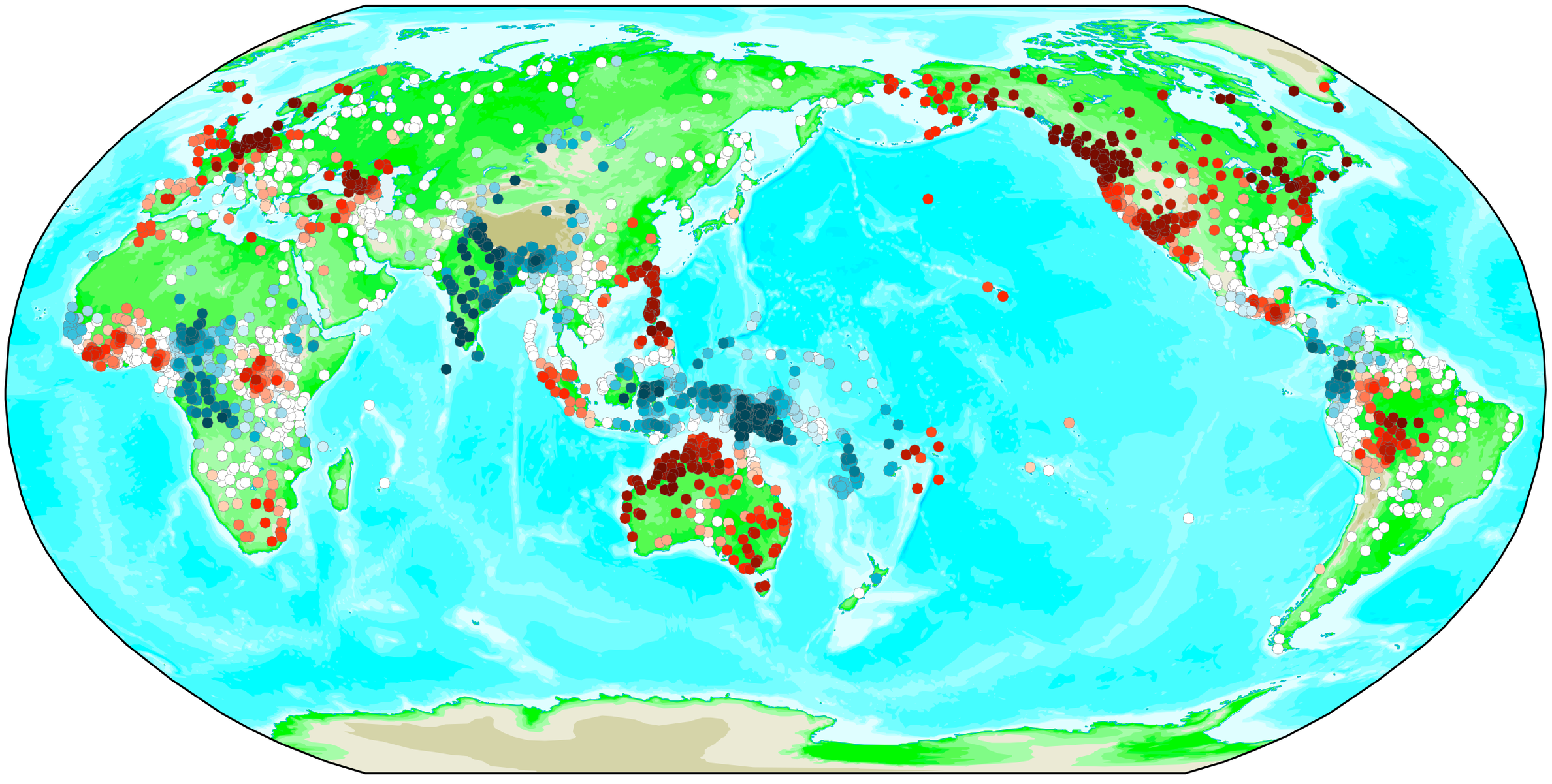


Top 300

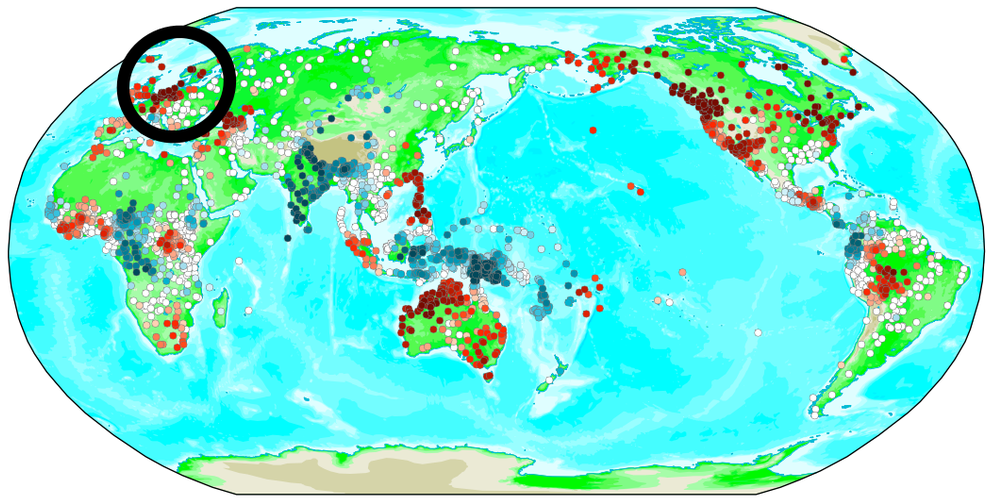


All languages

(red = rare, blue = common)

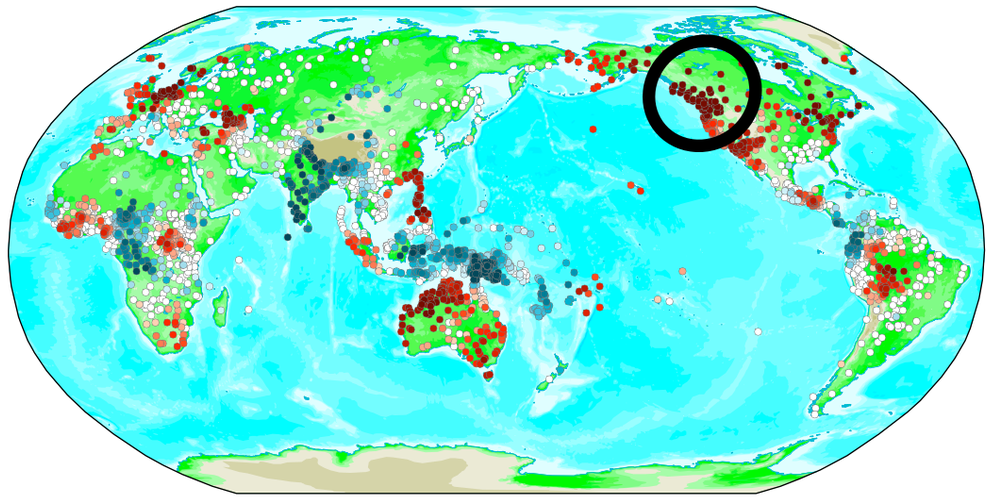


Northwest Europa



- Polar question are marked by word order only
- Some word order variability (object/verb and adjective/noun)
- Some unusual phonological characteristics (uvular continuants, front rounded vowels)
- ‘have’-perfect and tense suppletion
- The existence of relative pronouns
- No distance contrast in demonstrative pronouns

Northwest America



- Many unusual phonological characteristics: ejectives, lateral obstruents, uvular stop/continuants, absence of nasals
- Various forms of clitics and case prefixes
- VS order, but otherwise very flexible in word order
- Four way demonstrative contrast
- Complex verbal morphology

The future ?!

- These first results are nice, but ...
- The investigation of such large datasets in typology is still only in its infancy
- Many dependencies in the data
- No really applicable statistical tests exist

The End