DiVerSiTy: verso un pacchetto Python per il campionamento tipologico *bias-sensitive*

Filippo Sergio & Ludovica Pannitto

Alma Mater Studiorum – Università di Bologna

CLUB Day 2025 – 19/05/2025

LABORATORIO SPERIMENTALE

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

- Introduzione
- Riflessione sulla metodologia e studi precedenti
- Il metodo di campionamento
- A che punto siamo?
- Conclusioni

Introduzione

Motivazioni:

- Esigenze dovute alla tesi dottorale (creazione di campione tipologico)
- Alcune riflessioni sulla letteratura inerente alla metodologia della ricerca (con riferimento alla tipologia; vedi Gast & Jędrzejowski 2022, Becker & Guzman Naranjo 2025)

Introduzione

Motivazioni:

- Esigenze dovute alla tesi dottorale (creazione di campione tipologico)
- Alcune riflessioni sulla letteratura inerente alla metodologia della ricerca (con riferimento alla tipologia; vedi Gast & Jędrzejowski 2022, Becker & Guzman Naranjo 2025)

- Più precisamente, rispetto ai concetti di replicazione e replicabilità
  - "[…] replication serves the purpose of consolidating the findings, as they are arguably more robust when being reproduced […]

Introduzione

Motivazioni:

- Esigenze dovute alla tesi dottorale (creazione di campione tipologico)
- Alcune riflessioni sulla letteratura inerente alla metodologia della ricerca (con riferimento alla tipologia; vedi Gast & Jędrzejowski 2022, Becker & Guzman Naranjo 2025)

- Più precisamente, rispetto ai concetti di replicazione e replicabilità
  - "[…] replication serves the purpose of consolidating the findings, as they are arguably more robust when being reproduced […]
  - "[…] replication has not played a very important role in language typology so far […]" (Becker & Guzman Naranjo 2025: 1-2)

Introduzione

Motivazioni:

- Esigenze dovute alla tesi dottorale (creazione di campione tipologico)
- Alcune riflessioni sulla letteratura inerente alla metodologia della ricerca (con riferimento alla tipologia; vedi Gast & Jędrzejowski 2022, Becker & Guzman Naranjo 2025)

- Più precisamente, rispetto ai concetti di replicazione e replicabilità
  - "[…] replication serves the purpose of consolidating the findings, as they are arguably more robust when being reproduced […]
  - "[…] replication has not played a very important role in language typology so far […]" (Becker & Guzman Naranjo 2025: 1-2)
  - "While hardly carried out in practice besides as part of reviewing, exact replications are highly important theoretically and correspond to the minimal requirements of replicability of an empirical study." (Becker & Guzman Naranjo 2025: 3)

Alcune riflessioni sulla metodologia

Necessità di quattro concetti/pratiche fondamentali:

Alcune riflessioni sulla metodologia

Necessità di quattro concetti/pratiche fondamentali:

1. Falsificabilità
2. Riproducibilità
3. Replicazione e replicabilità
4. Robustezza

Alcune riflessioni sulla metodologia

1. Falsificabilità
    - Perché una ricerca e i suoi prodotti siano falsificabili, le procedure metodologiche devono essere rese esplicite e chiare. Tutte le procedure devono essere *operazionalizzate*. (es. anche per i concetti comparativi; "Operationalization is “the construction of actual, concrete measurement techniques” for theoretical constructs"; Gast & Jędrzejowski 2022:598 citando Babble 1985: 6)

2. Riproducibilità
    - Le operazionalizzazioni devono essere descritte nei minimi dettagli . (≠ da falsificabilità. La falsificabilità è a monte, è il principio per cui si operazionalizza)

Alcune riflessioni sulla metodologia

3. Replicabilità e replicazione

- Replicability corresponds to the potential of exact replication. It guarantees that another independent scientist can use the same data and follow the same procedure as in the original study, obtaining the same results.
- Replication is a methodological tool based on a repetition procedure that is involved in assessing or reducing the amount of uncertainty regarding previous research results. In doing so, it can be used to establish a piece of knowledge of our world.

4. Robustezza

- Robustness refers to the stability of experimental conclusions to variations in either baseline assumptions or experimental procedures. (Goodman et al. 2016: 4 citati in Becker & Guzman Naranjo 2025: 3)

Studi precedenti

Campioni tipologici:

- Non hanno avuto un trattamento unitario e sistematico nella letteratura tipologica (vedi però Windmann & Bakker 2006, Bakker 2010, Miestamo et al. 2016, Song 2018)
- Confronti fra pratiche di campionamento rari (mancanza di studi sulla replicabilità e replicazione)
- In termini assoluti, non esiste un tipo di campione, metodo di campionamento e pratica di costruzione migliore di un altro. È parte integrante della ricerca tipologica ideare delle procedure di campionamento più adatte (alle esigenze del tipologo, della ricerca, dell'oggetto di ricerca)

Studi precedenti

Cosa emerge dalla letteratura sul campionamento?

(Bakker & Windmann 2006, Miestamo et al. 2016)

1. Usare un campione bilanciato è meglio che usare un campione di convenienza (= il campione, che non mira ad essere bilanciato e rappresentativo, corrisponde alla quantità/qualità di dati del tipologo)

2. I campioni randomici performano meglio dei campioni di convenienza ma peggio dei campioni bilanciati, soprattutto se stratificati

3. I campioni stratificati performano meglio dei campioni randomici. I risultati, tuttavia, dipendono da:
   - Qualità quantità/qualità di stratificazioni (es. areale, genetica)
   - Dall'ampiezza del campione
   - Dalla classificazione utilizzata (punto problematico)

Il metodo in breve

Il metodo in breve

1. Tipo di campione: campione di varietà
   - (≠ campione di probabilità)
   - Campione di dimensioni medio-grandi (350/+500 lingue)
   - Le variabili possono essere già note, ma di solito sono sconosciute. Lo scopo è far emergere, e quindi catturare, il massimo grado di diversità presente a livello interlinguistico.
   - Tipico di studi iniziali, esplorativi e destinati ad indagare un fenomeno di cui non si conosce la gamma completa di caratteristiche e tratti strutturali (forme, funzioni, distribuzione inter-linguistica, …).
   - Bilanciamento ‘sacrificato’ a favore della rappresentatività della varietà interlinguistica. Spesso esposto a distorsioni. Sacrifica l’indipendenza fra casi a favore della rappresentatività.

Il metodo in breve

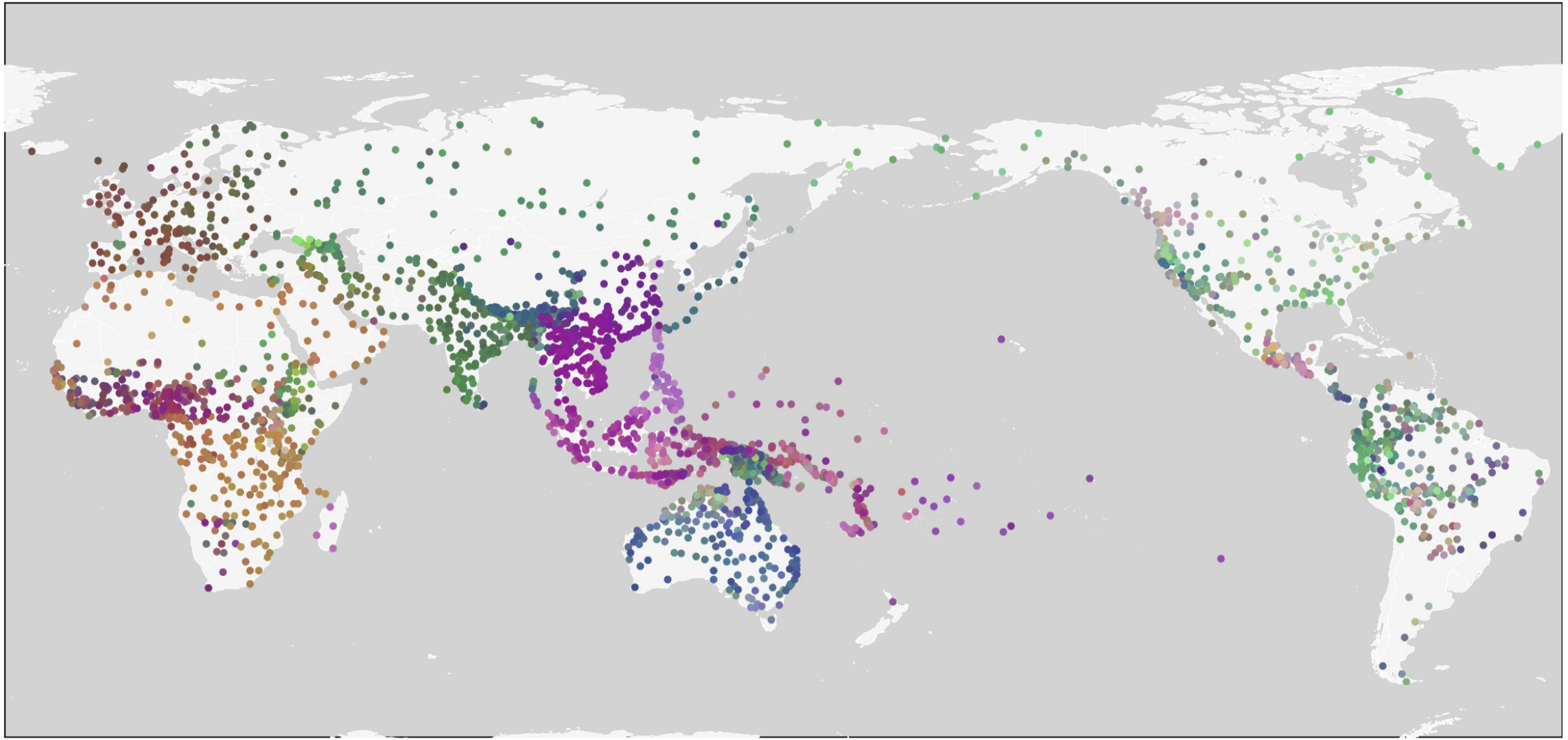
2. Metodo di campionamento: Diversity Value (DV; Rijkhoff et al. 1993)
    - Unico metodo di campionamento per campioni di varietà pienamente formalizzato
    - Permette di ampliare proporzionalmente il campione se necessario
    - Usa un parametro extra-linguistico (la complessità degli alberi genealogici delle famiglie linguistiche) per stimare la diversità interlinguistica
        - Punto a favore, molti metodi per campioni sia di varietà che di probabilità implementano proprozioni basate sul numero di lingue
    - Classificazione basata su Glottolog (Hammarström et al. 2024, v. 5.1)

Il metodo in breve

3. Stratificazioni:
    1. Stratificazione genetica (garantita dal DV)

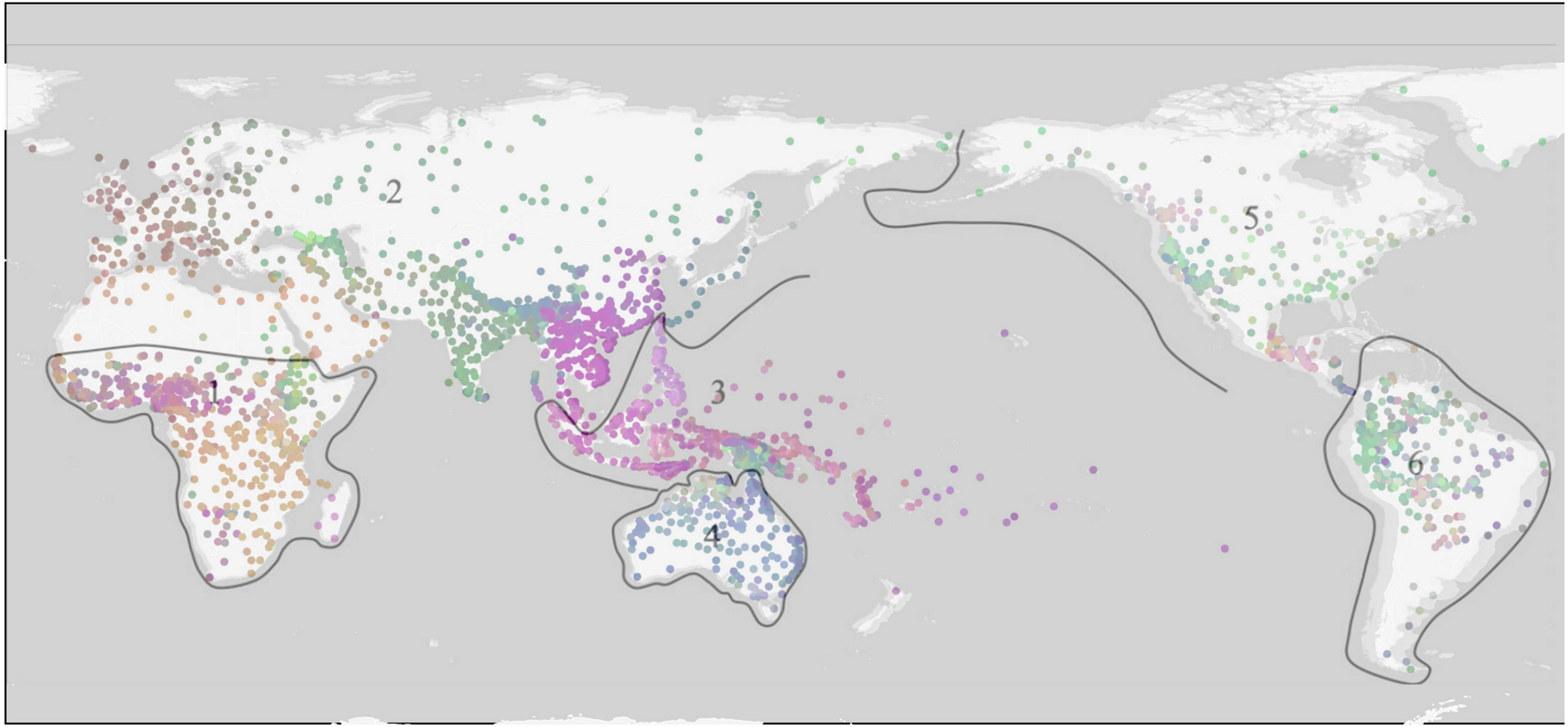
Il metodo in breve

3. Stratificazioni:
    1. Stratificazione genetica (garantita dal DV)
    2. Stratificazione areale (idealmente di due livelli, ma aperta a possibili miglioramenti)



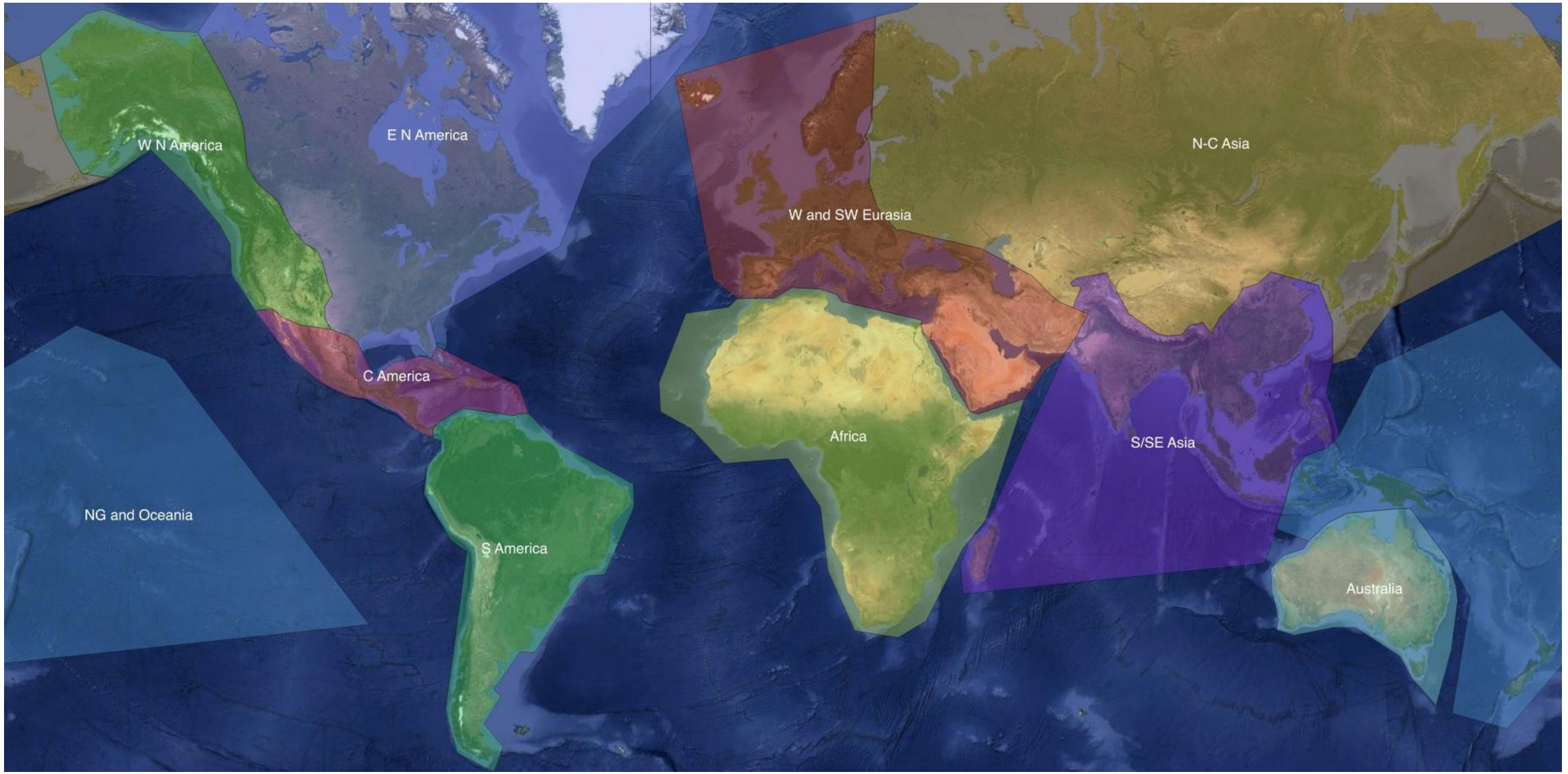
Kylan & Donohue (2023; Supplementary materials). Mappa della diversità linguistica a partire da aggregati di feature morfosintattiche collezionate da diversi database, es. WALS, GramBank, AUTOTYP, etc per ~3000 delle +7500.

Primo livello di stratificazione: per macroaree (Hammarström & Donohue 2014)



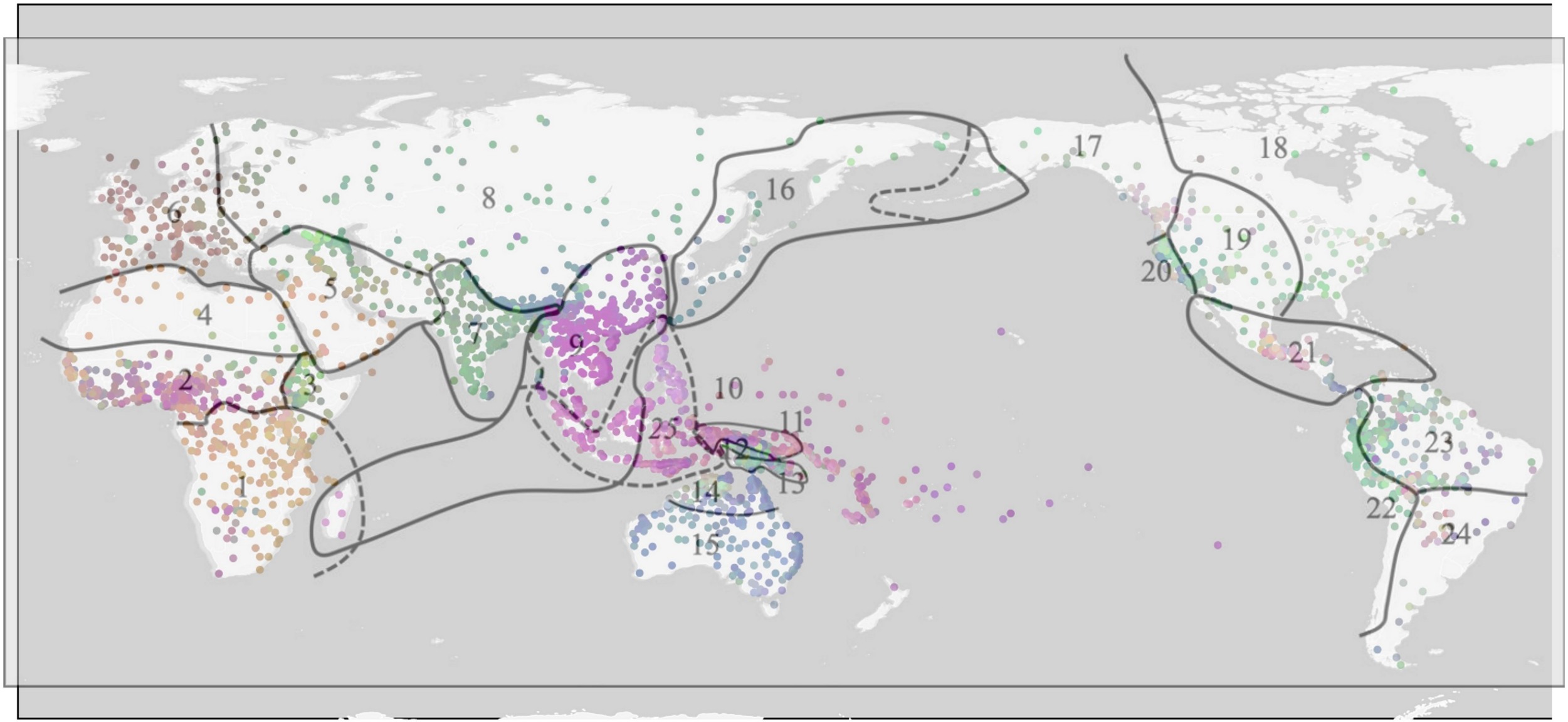
Kylan & Donohue (2023; Supplementary materials). Mappa della diversità linguistica a partire da aggregati di feature morfosintattiche collezionate da diversi database, es. WALS, GramBank, AUTOTYP, etc per ~3000 delle +7500.

Secondo livello di stratificazione: per sub-aree (Bickel & Nichols 2013)



Kylan & Donohue (2023; Supplementary materials). Mappa della diversità linguistica a partire da aggregati di feature morfosintattiche collezionate da diversi database, es. WALS, GramBank, AUTOTYP, etc per ~3000 delle +7500.

Terzo livello di stratificazione (ipotetico): per aree culturali (Bickel & Nichols 2013 a partire da Nichols 1992)



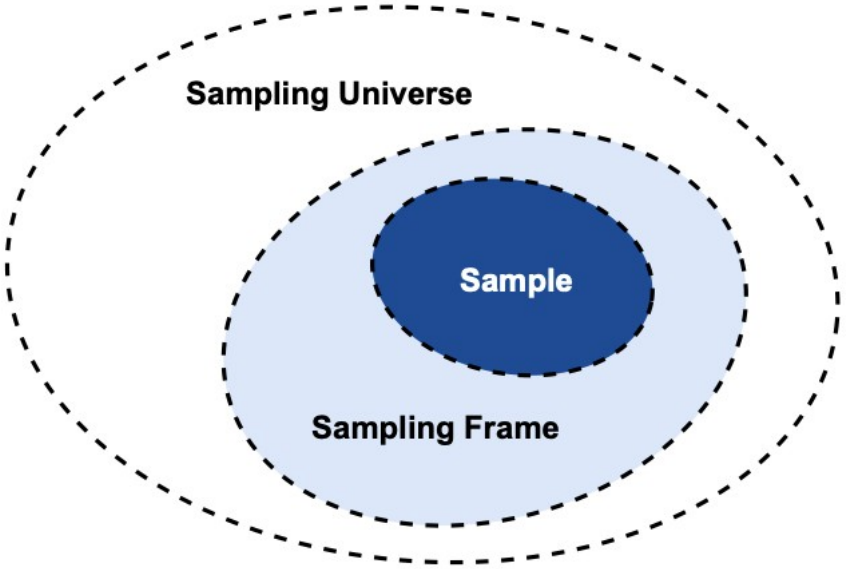
Kylan & Donohue (2023; Supplementary materials). Mappa della diversità linguistica a partire da aggregati di feature morfosintattiche collezionate da diversi database, es. WALS, GramBank, AUTOTYP, etc per ~3000 delle +7500.

Il metodo in breve

3. Stratificazioni:
    1. Stratificazione genetica (garantita dal DV)
    2. Stratificazione areale (idealmente di due livelli, ma aperta a possibili miglioramenti)
    3. Stratificazione "bibliografica"

Il metodo in breve

3. Stratificazioni:
    1. Stratificazione genetica (garantita dal DV)
    2. Stratificazione areale (idealmente di due livelli, ma aperta a possibili miglioramenti)
    3. Stratificazione "bibliografica"

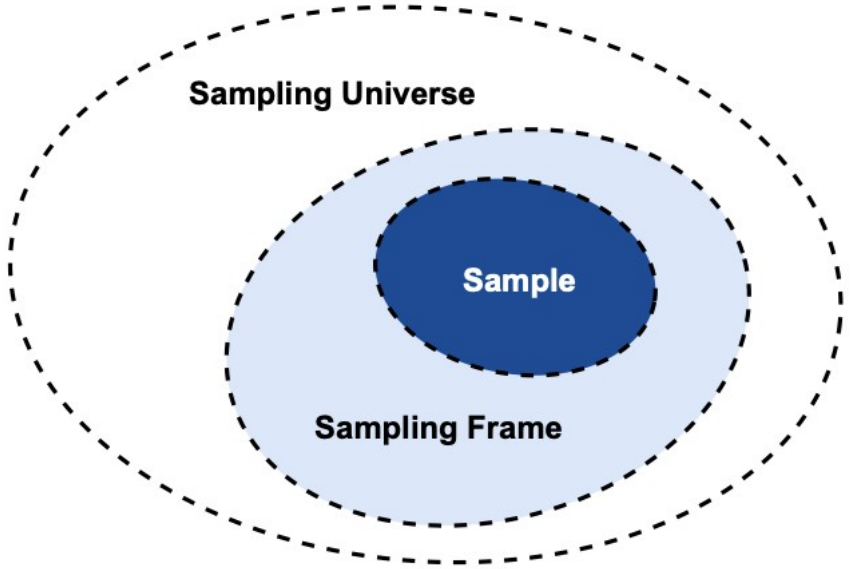


Concetti di universo, frame e campione come proposti da Bell (1978) e discussi in Ploeger et al. 2025

Il metodo in breve

3. Stratificazioni:

1. Stratificazione genetica (garantita dal DV)
2. Stratificazione areale (idealmente di due livelli, ma aperta a possibili miglioramenti)
3. Stratificazione "bibliografica"



Secondo Sjöberg (2023), il Sampling Frame andrebbe ulteriormente distinto in:

1. *Catalogue frame*: insieme delle lingue che sappiamo esistere
2. *Corpus frame*: insieme delle lingue per cui esistono sufficienti materiali descrittivi (nel nostro caso, grammatiche)

Concetti di universo, frame e campione come proposti da Bell (1978) e discussi in Ploeger et al. 2025

Come si trova il campione?

- Supponiamo di voler ottenere un campione di X lingue da un certo numero F di famiglie linguistiche
- Il metodo impone di calcolare il Diversity Value (DV) attribuibile ad ogni famiglia
- Decidiamo poi quante lingue campionare da ogni famiglia a seconda del peso relativo del DV attribuito a quella famiglia

- L'obiettivo finale è di utilizzare l'algoritmo su glottolog ma:
  - Il paper descrive l'esito del sampling su tre classificazioni generiche: Voegelin&Voegelin 1977, **Ruhlen 1987**, Grimes 1997
  - Nessuna di queste è reperibile in formato machine-readable

E cos'è il Diversity Value?

$$DV = \frac{1}{L} \sum_{i=1}^{L} C_i$$

ESKIMO-ALEUT [9]:
I ALEUT [1]: *Aleut
II ESKIMO [8]:
A INUIT [3]: Alaskan Inuit, *Canadian Inuit, *Greenlandic
B YUPIK [5]:
1 ALASKAN [2]: Alutiiq, *Central Yupik
2 SIBERIAN [3]:
[1]: Sirenik
a CHAPLINO-NAUKAN [2]: *Chaplino, *Naukan

- Formalmente, viene definito come la **media** del valore di una funzione C calcolata su **ogni livello** dell'albero filogenetico.

Table 6. *The distribution of languages over language families in a 100-language sample (Ruhlen's classification)*

Family	DV	Phase 1	Phase 2		Phase 3
Afro-Asiatic	55.53	0	6		6
Altaic	14.79	0	1	→	2
Amerind	178.44	0	18		18
Australian	67.58	0	7		7
Austric	137.41	0	14		14
Caucasian	8.54	0	1		1
Chukchi-Kamchatkan	2.47	1	1		1
Elamo-Dravidian	7.43	1	1		1
Eskimo-Aleut	3.34	1	1		1
Indo-Hittite	39.71	0	4		4
Indo-Pacific	123.39	0	12	→	13
Khoisan	6.97	1	1		1
Language Isolates:					
Basque	1.00	1	1		1
Burushaski	1.00	1	1		1
Etruscan	1.00	1	1		1
Gilyak	1.00	1	1		1
Hurrian	1.00	1	1		1
Ket	1.00	1	1		1
Meroitic	1.00	1	1		1
Nahali	1.00	1	1		1
Sumerian	1.00	1	1		1
Na-Dene	9.44	0	1		1
Niger-Kordofanian	90.38	0	9		9
Nilo-Saharan	42.18	0	4	→	5
Pidgins & Creoles	13.47	0	1	→	2
Sino-Tibetan	38.52	0	4		4
Uralic-Yukaghir	4.93	1	1		1
Totals	853.20	14	96		100

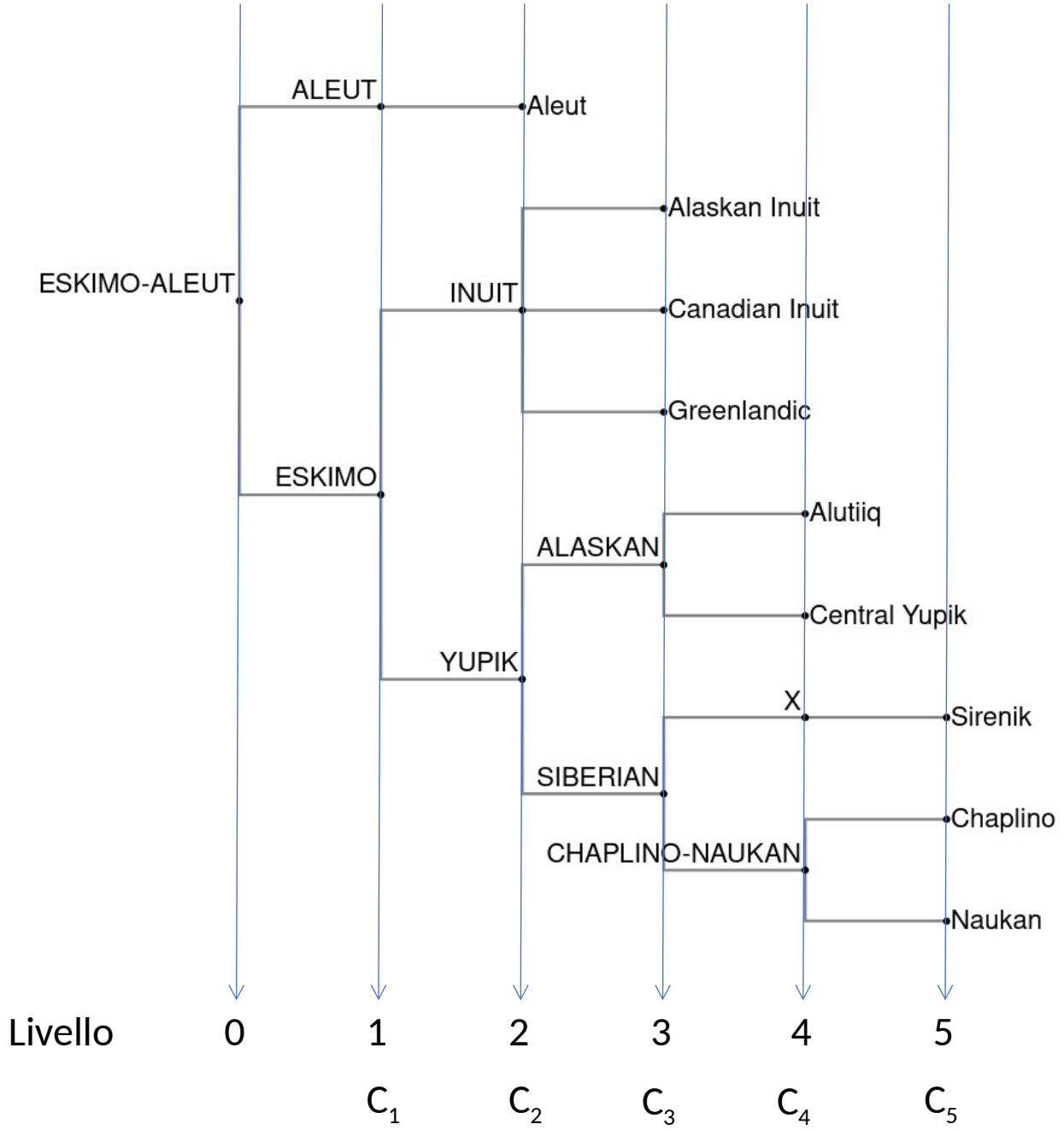
Creazione del campione:

- [Fase 1] Se il DV della famiglia produce un peso < 1, almeno una lingua viene selezionata
- [Fase 2] Quanto resta del campione viene selezionato proporzionalmente sul resto dei DV
- [Fase 3] Il campione viene aggiustato in eccesso o in difetto per via delle approssimazioni fatte nelle fasi 1 e 2

- Il livello è definito come la distanza dalla radice dell'albero, ovvero dal nodo che corrisponde all'intera famiglia.
- In generale C_i è definito come segue:

$$C_i = \begin{cases} N_i & : \ i = 1 \\ C_{i-1} + (N_i - N_{i-1}) \times \frac{N_{\max} - (i-1)}{N_{\max}} & : \ i > 1 \end{cases}$$

- L'idea è che dipenda dal valore del livello precedente più un fattore (pesato) che dipende dal *branching al livello i*
- Resta da scoprire che funzione è N e chi è N_{max}



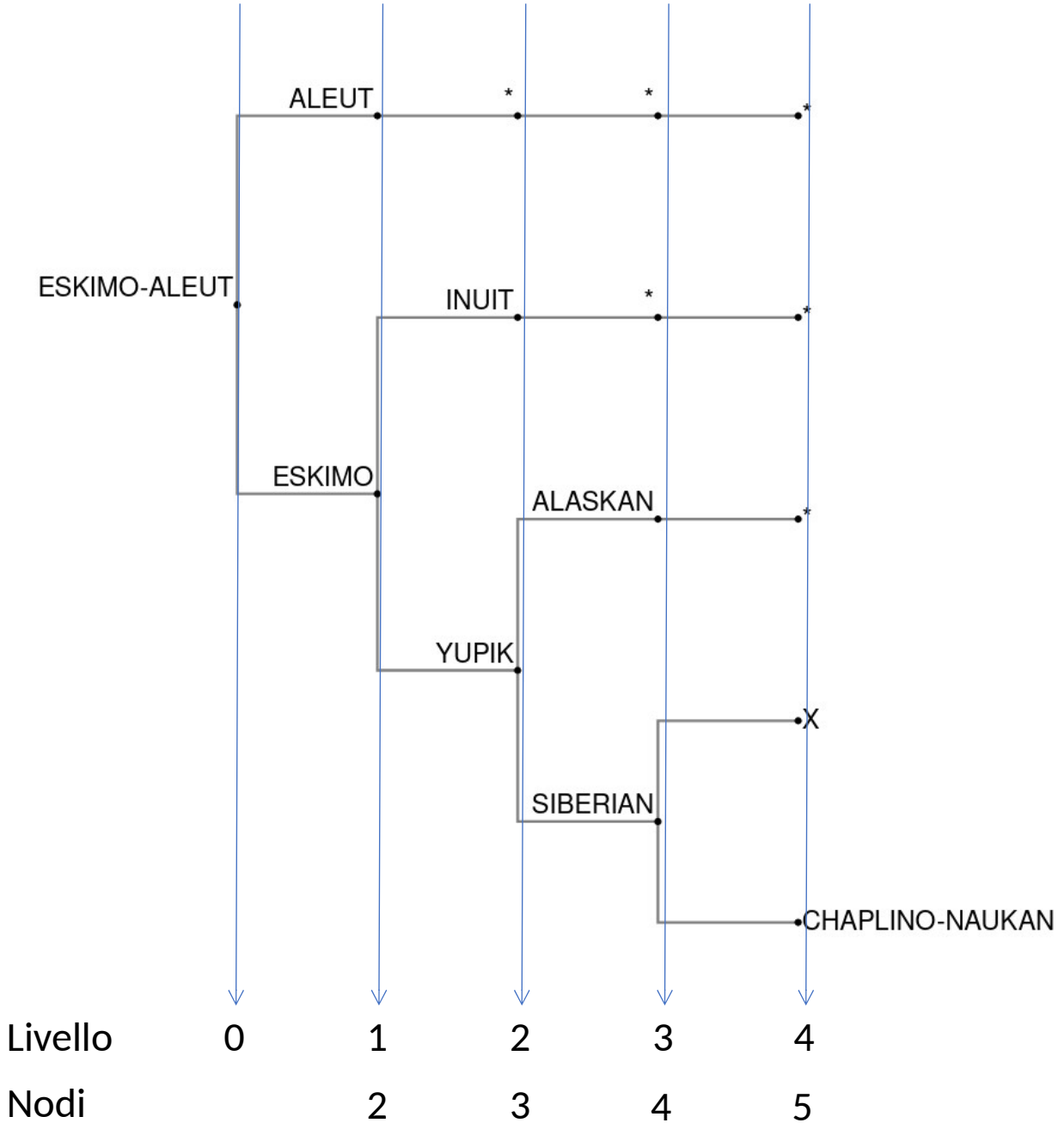
Non viene in realtà considerato l'albero completo ma un suo derivato, dove:

- Le lingue (foglie dell'albero) vengono eliminate
- I nodi i cui figli sono solamente foglie vengono "proiettati" fino al massimo livello

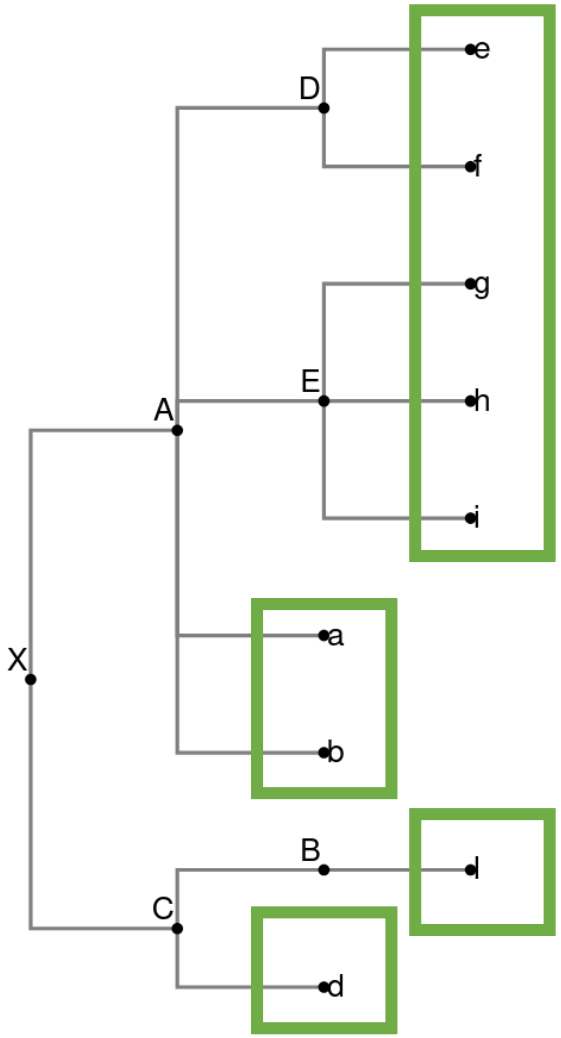
Nella nostra implementazione, chiamiamo questo albero **RB-equivalente** al precedente

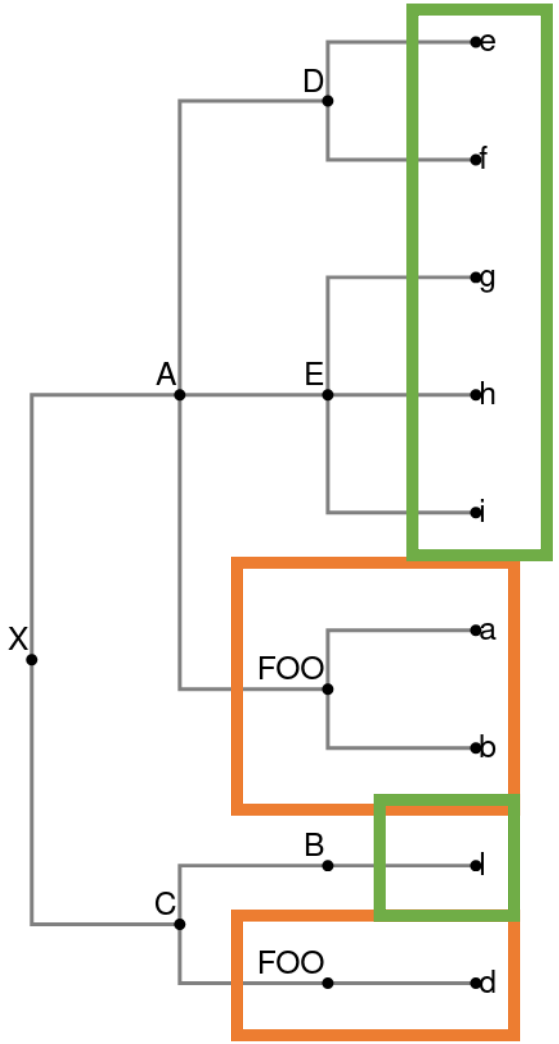
N(i) è la funzione che conta il numero di nodi di livello *i* nell'albero.

N_{max} è invece la massima distanza dalla radice trovata nell'insieme F di famiglie



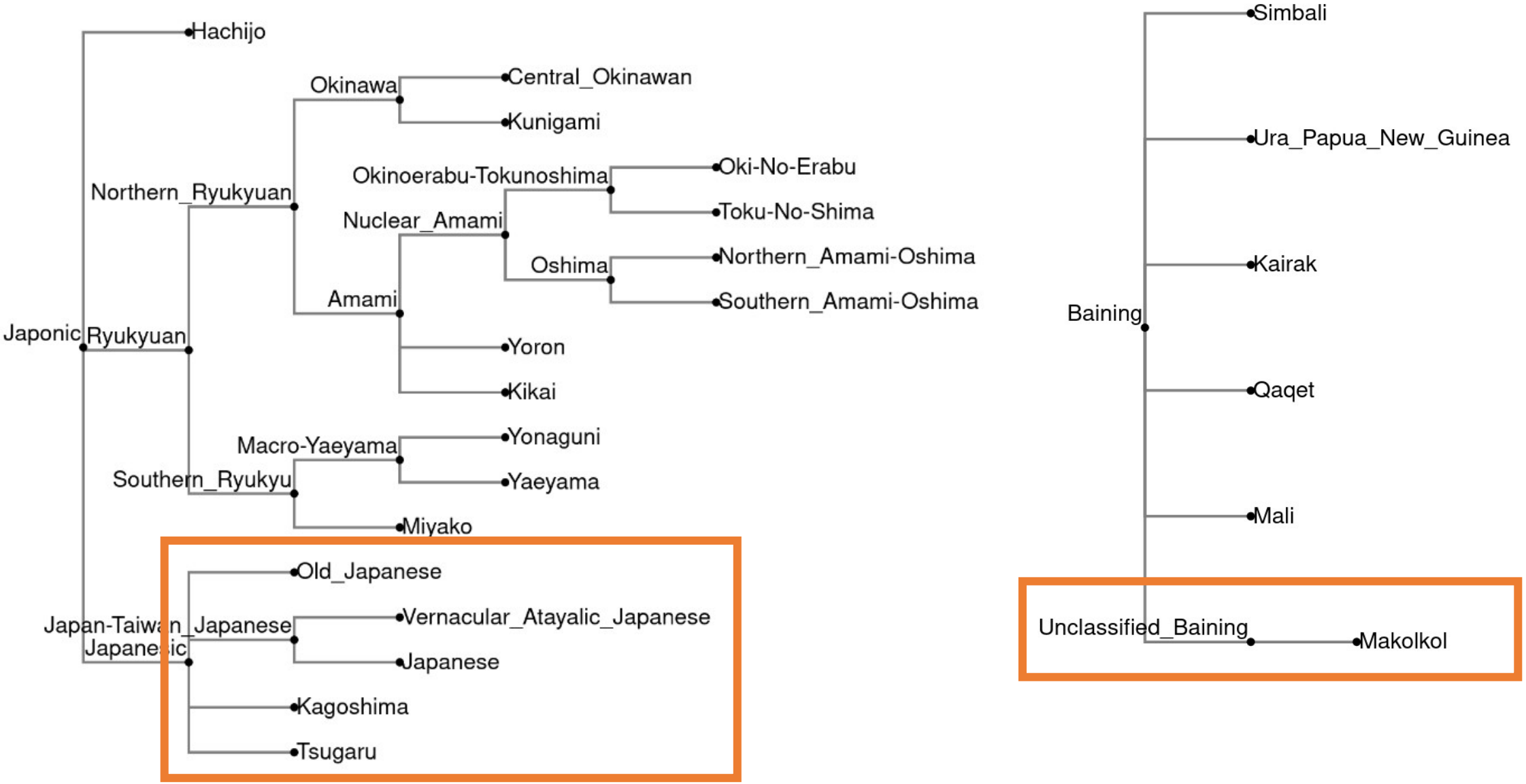
Come ottenere l'albero RB-equivalente?





Per ignorare il numero di lingue (foglie) ma mantenere l'informazione sul *branching* ad ogni livello, vengono in realtà **aggiunti dei nodi intermedi** che raggruppano tutte le foglie ai livelli intermedi dove è presente anche una sottofamiglia

ESKIMO-ALEUT [9]:
I ALEUT [1]: *Aleut
II ESKIMO [8]:
A INUIT [3]: Alaskan Inuit, *Canadian Inuit, *Greenlandic
B YUPIK [5]:
1 ALASKAN [2]: Alutiiq, *Central Yupik
2 SIBERIAN [3]:
[1]: Sirenik
a CHAPLINO-NAUKAN [2]: *Chaplino, *Naukan



La classificazione proposta da glottolog non è consistente su questo aspetto, a volte foglie e nodi intermedi appaiono insieme tra i figli di un nodo, altre volte esistono sottofamiglie con una sola foglia

$$C_i = \begin{cases} N_i & : \ i = 1 \\ C_{i-1} + (N_i - N_{i-1}) \times \frac{N_{\max} - (i-1)}{N_{\max}} & : \ i > 1 \end{cases}$$

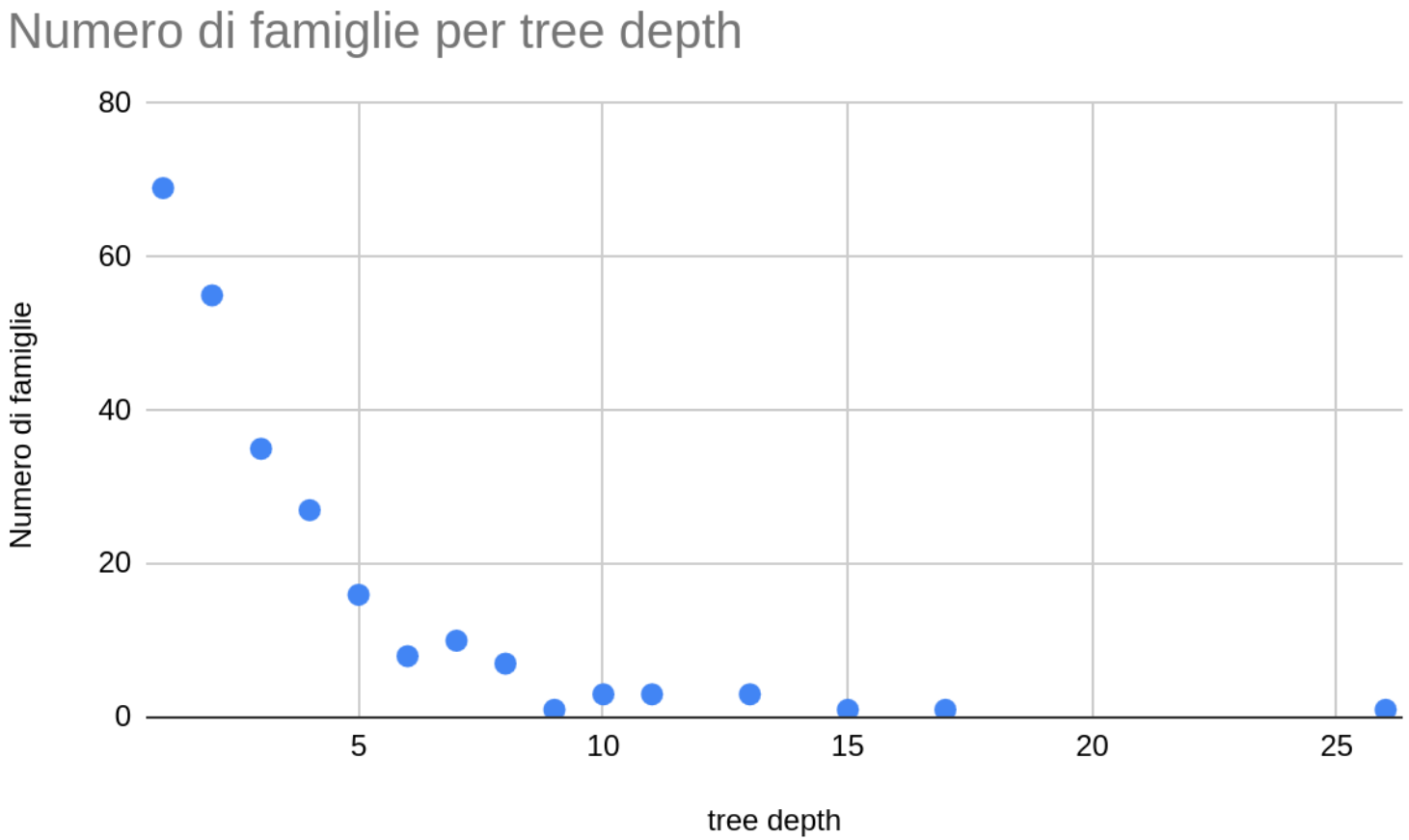
$$C_i = \begin{cases} f(1) & : \ i = 1 \text{ with } N_0 = 0 \\ C_{i-1} + f(i) \times \frac{N_{\max} - (i-1)}{N_{\max}} & : \ i > 1 \end{cases}$$

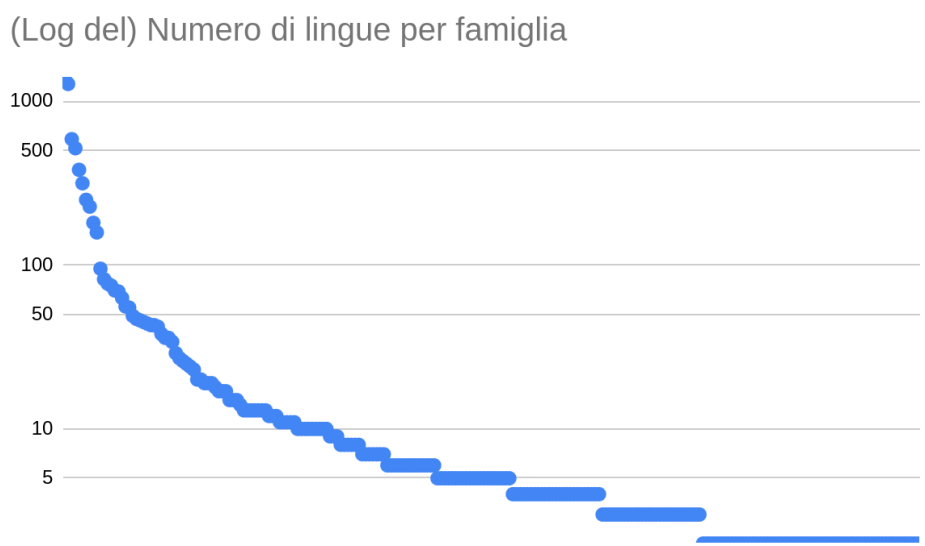
$$C_i = C_{i-1} + f(i) \times \frac{N_{\max} - (i-1)}{N_{\max}} \qquad C_i = C_{i-1} + f(i) \times (\frac{N_{\max}}{N_{\max}} - \frac{(i-1)}{N_{\max}}) \qquad C_i = C_{i-1} + f(i) \times (1 - \frac{(i-1)}{N_{\max}})$$

$$C_i = C_{i-1} + f(i) - f(i)\frac{(i-1)}{N_{\max}}$$

Nel caso di glottolog, la distribuzione delle profondità degli alberi è questa mostrata a sinistra.

In particolare, la mediana delle profondità è 2 e la media è 3,95, il 75° percentile 4…
sebbene **N_{max} sia 26**





$$C_i = C_{i-1} + f(i) - f(i)\frac{(i-1)}{N_{\max}}$$

Per (almeno) 180 famiglie su 240, questo vale al massimo $f(i) \times \frac{(4-1)}{26} = f(i) \times \frac{3}{26} = 0.11 \times f(i)$

Possiamo dimenticarcene e la formula si semplifica: $DV = (N_1 + N_2 + N_3 + N_4)/4$

$$C_1 = f(1) = N_1$$

$$C_2 = C_1 + f(2) = f(1) + f(2) = N_1 + N_2 - N_1 = N_2$$

$$C_3 = C_2 + f(3) = C_1 + f(2) + f(3) = f(1) + f(2) + f(3) = N_1 + N_2 - N_1 + N_3 - N_2 = N_3$$

$$C_4 = C_3 + f(4) = C_2 + f(3) + f(4) = C_1 + f(2) + f(3) + f(4) = f(1) + f(2) + f(3) + f(4) = N_1 + N_2 - N_1 + N_3 - N_2 + N_4 - N_3 = N_4$$

$$DV = (N_1 + N_2 + N_3 + N_4)/4$$

Per 163 delle 180 famiglie precedenti , il numero di lingue nella famiglia è al massimo 9.

Per queste, la somma dei nodi interni è al massimo 6 (nodi propri) + 8 (nodi proiettati) = 14

Quindi DV è al massimo poco meno di 4.

Family	Number of languages
Afro-Asiatic	258
Altaic	66
Amerind	854
Australian	262
Austric	1186
Caucasian	38
Chukchi-Kamchatkan	5
Elamo-Dravidian	29
Eskimo-Aleut	9
Indo-Hittite	180
Indo-Pacific	748
Khoisan	33
Na-Dene	41
Niger-Kordofanian	1068
Nilo-Saharan	138
Sino-Tibetan	268
Uralic-Yukaghir	27
Language Isolates	9
Pidgins and Creoles	38
Unclassified languages	16
Total	5273

Table 5. *Diversity Values (DVs) of Ruhlen's language families*

Family	DV	Family	DV
Afro-Asiatic	55.53	Nilo-Saharan	42.18
Altaic	14.79	Pidgins and Creoles	13.47
Amerind	178.44	Sino-Tibetan	38.52
Australian	67.58	Uralic-Yukaghir	4.93
Austric	137.41	Basque (LI)	1.00
Caucasian	8.54	Burushaski (LI)	1.00
Chukchi-Kamchatkan	2.47	Etruscan (LI)	1.00
Elamo-Dravidian	7.43	Gilyak (LI)	1.00
Eskimo-Aleut	3.34	Hurrian (LI)	1.00
Indo-Hittite	39.71	Ket (LI)	1.00
Indo-Pacific	123.39	Meroitic (LI)	1.00
Khoisan	6.97	Nahali (LI)	1.00
Na-Dene	9.44	Sumerian (LI)	1.00
Niger-Kordofanian	90.38	Total	853.20

E quindi?

1. La classificazione di Ruhlen, su cui il metodo è stato sviluppato, prevedeva 20 famiglie (compresi anche pidgin, lingue non classificate etc…). Glottolog ne ha 240.
2. Il metodo assicura di campionare almeno una lingua per famiglia.
3. Immaginiamo di aver fissato un campione di 500 lingue (*è realistico?*), 180 di queste verranno pescate dal 75% delle famiglie "povere" di glottolog (DV poco meno di 4 determinerà la scelta di una lingua), quindi di fatto stiamo campionando il restante 25% (circa 60 famiglie) per rappresentare il 64% (320 lingue) del campione…

Intanto…

- Stiamo ancora tentando di riprodurre del tutto i risultati di Rijkoff&Bakker, poiché le classificazioni genealogiche non sono disponibili in formato elettronico (> le stiamo producendo)

- Lo script ad oggi permette di determinare sia la distribuzione di lingue per famiglia che all'interno di ognuna di esse

Conclusioni

1. Importanza crescente di aderire e adottare procedure metodologiche che garantiscano
    1. Falsificabilità
    2. Riproducibilità
    3. Replicabilità e replicazione
    4. Robustezza dei risultati

    5. Favorire il riuso (e il controllo!) non solo dei dati ma anche degli strumenti creati e utilizzati per una ricerca
    6. Necessità di studi comparativi sulle performance di campioni studio-specifici o versatili/universali (?)