Areal patterns in the

World Atlas of Language Structures

Balthasar Bickel & Johanna Nichols

U Leipzig UC Berkeley

www.uni-leipzig.de/~autotyp

AUTOTYP

Theoretical assumptions

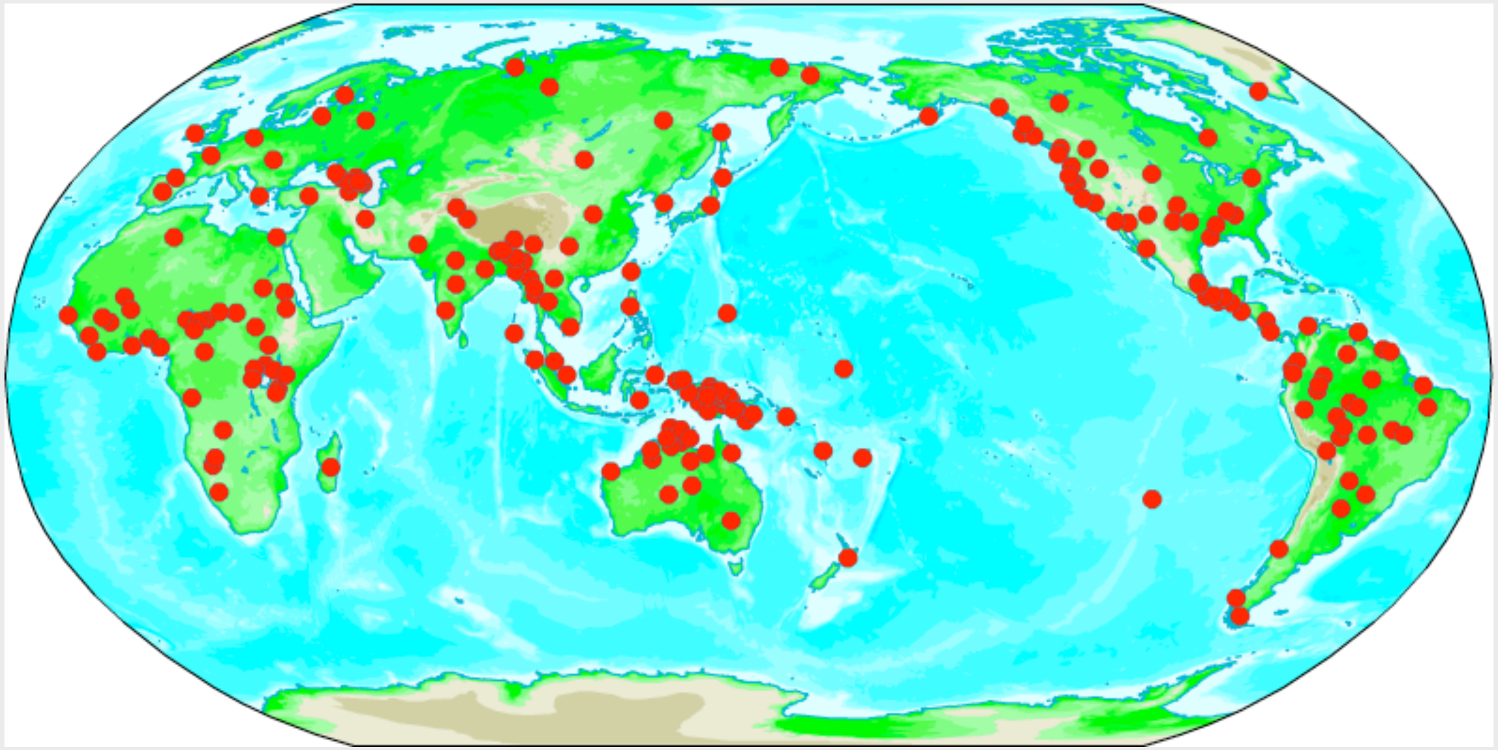
- ‘Areal pattern’ = shared history of contact and inheritance beyond demonstrable genealogical relations
- The quantitative assessment of areal patterns is grounded in a theory of population history, i.e. a theory of large-scale population and/or language movements — not on *visual* impressions
- Because each variable has its own history of inheritance and contact, and its own intrinsic stability degree, we do not necessarily expect clustering/isoglosses
- Instead, each variable can reflect areal factors on its own terms
- If we do find isoglosses, they can arise from
  - structural dependence between variables
  - similar historical stability degrees of variables
  - intensely shared history, pointing to a single (but not reconstructable) stock or a uniform *Sprachbund* in the extreme case

Sample

- The printed maps in WALS are not systematically sampled. Wouldn't they approximate the statistician's notion of a 'random sample'?
- … perhaps, but the statistical standard is random sampling, ***plus control variables (strata)***
- But we cannot control for stock effects by stratification (too many, too few datapoints within the stocks)
- Therefore, sample at a genealogically higher level than individual languages (cf. Dryer 1989)
- Pick one language per stock, more if geographically *and* genealogically far-flung (e.g., Indoeuropean, Pama-Nyungan, Sino-Tibetan)

Sample: WALSG

- Approximate this by taking the WALS sample and reducing overrepresented families (e.g. Bantu, Germanic, etc.), and adding some isolates: the **WALSG** sample (G for ‘genealogy’), *N* = 189



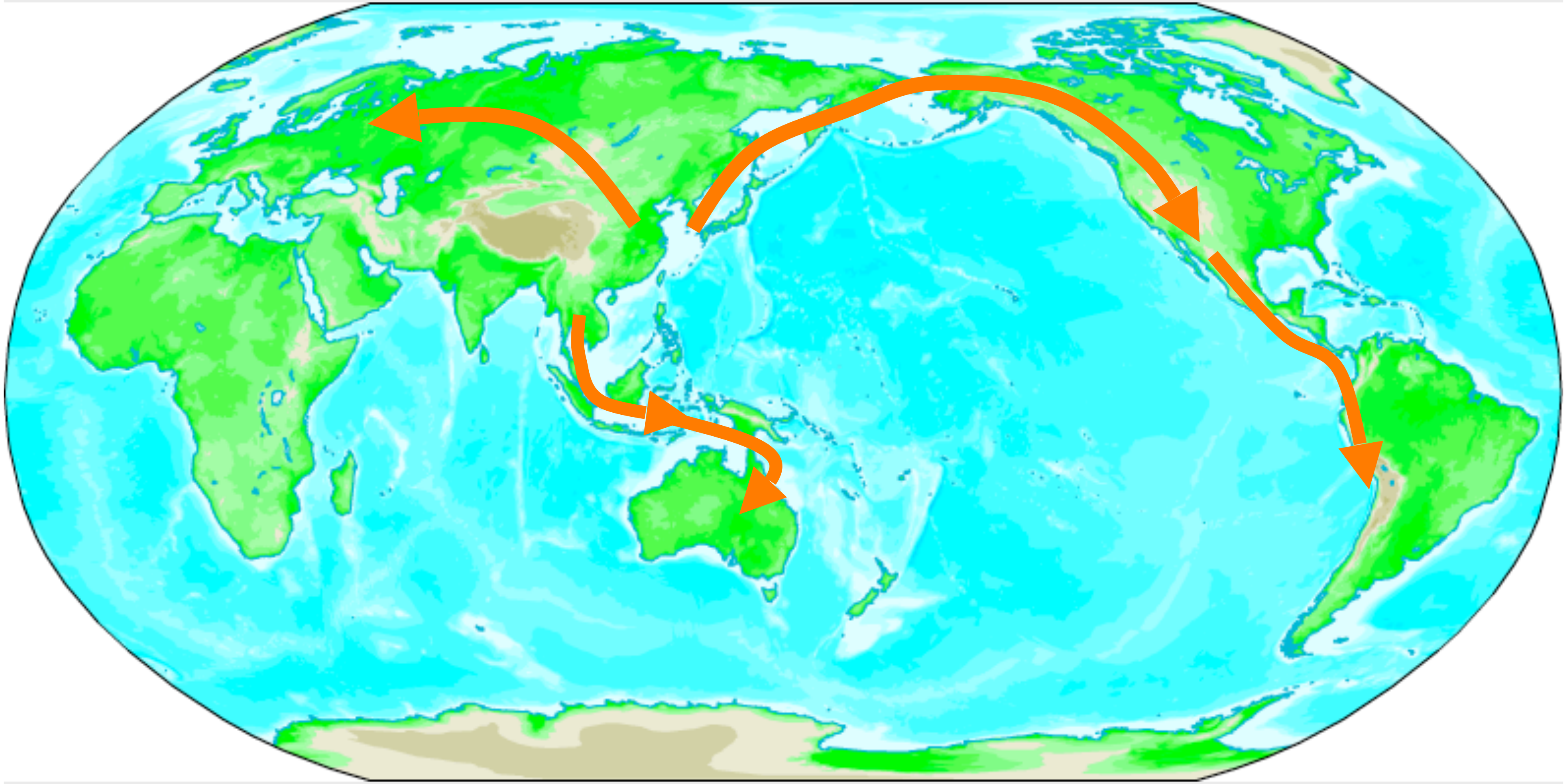
Method (Janssen, Bickel & Zúñiga 2005)

- our sample = population (*all* stocks for which we have data, and *not* sampled from an assumed pan-chronic population we have no way of knowing about…)
- therefore, all distribution-based statistics (including nonparametric statistics) is mathematically meaningless
- therefore, permutation tests (= exact and Monte-Carlo methods)
- advantages:
  - can also handle very heterogenous factor levels
  - if exact, p-value reflects strength of association (Gries & Stefanowitsch 2003)
- consequence: all inference to underlying populations (the populations that cause the observed distribution) is a theoretical, not a statistical issue

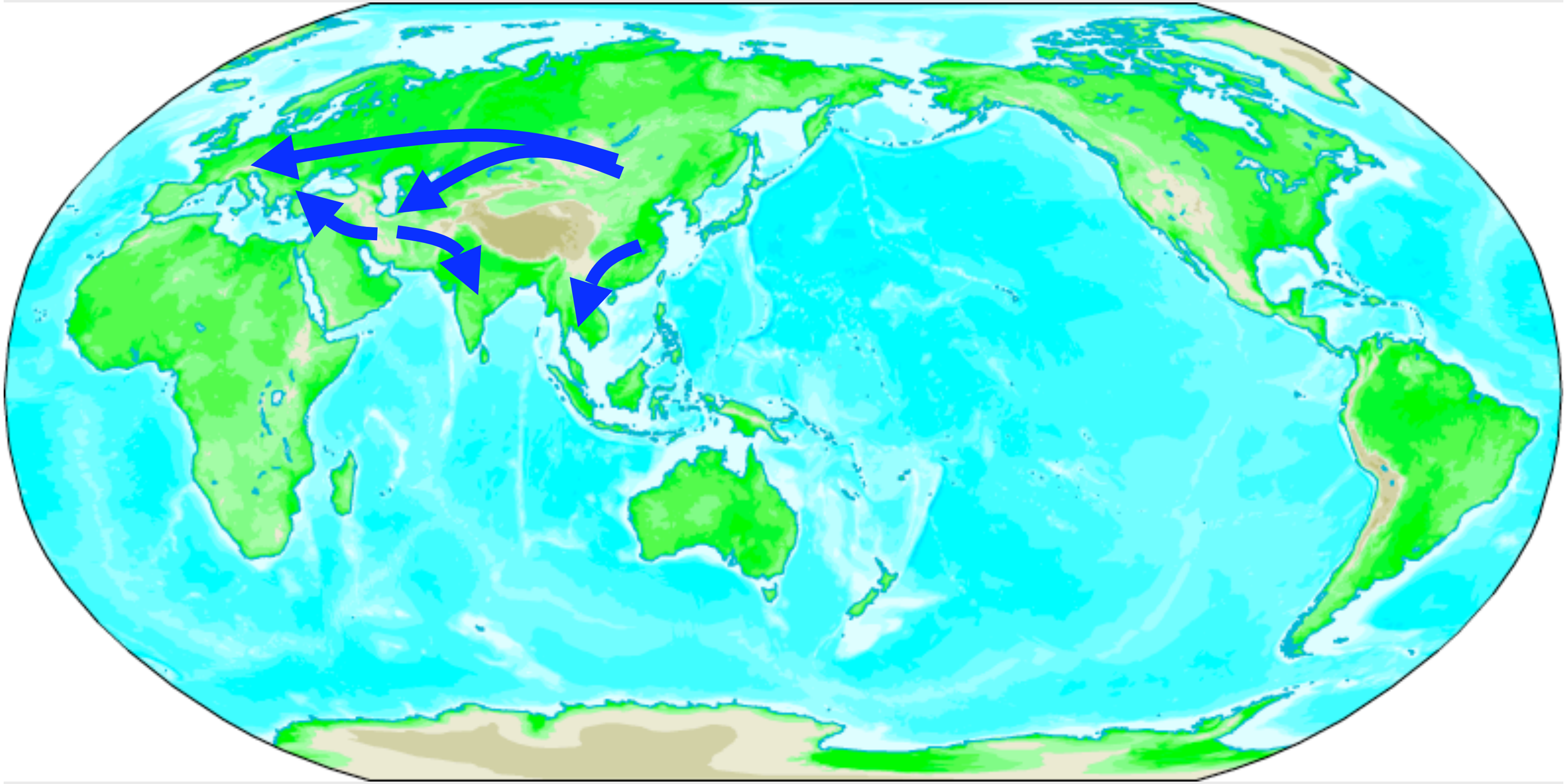
Areal factors: theory and hypotheses

- Four major classes of events (Bickel & Nichols 2003, 2005):
  - Circumpacific spreads
  - Eurasian (chiefly northern, southwestern and southeastern) spreads
  - Enclave effects (Himalayas, Caucasus)
  - Fringe effects (Europe)

Geographical factors: theory and hypotheses

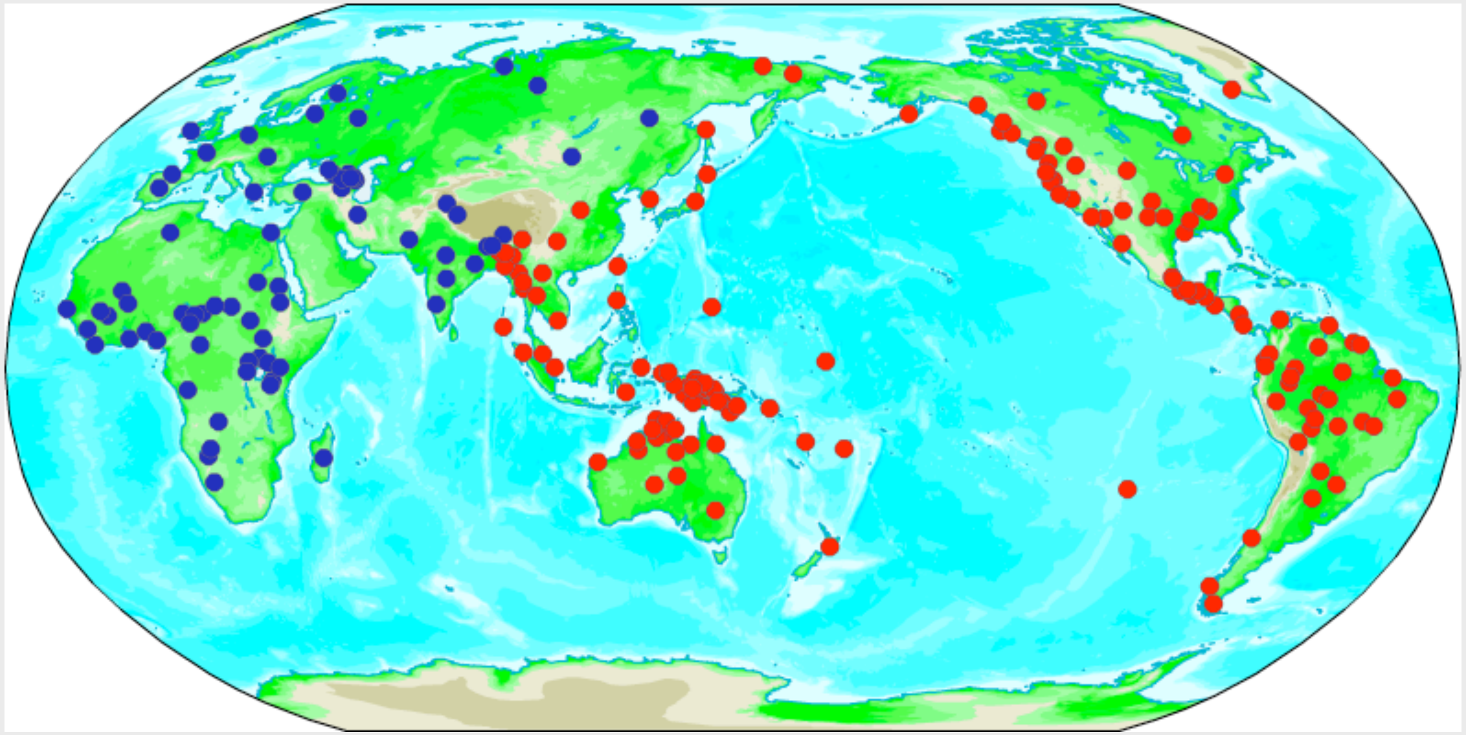


Geographical factors: theory and hypotheses



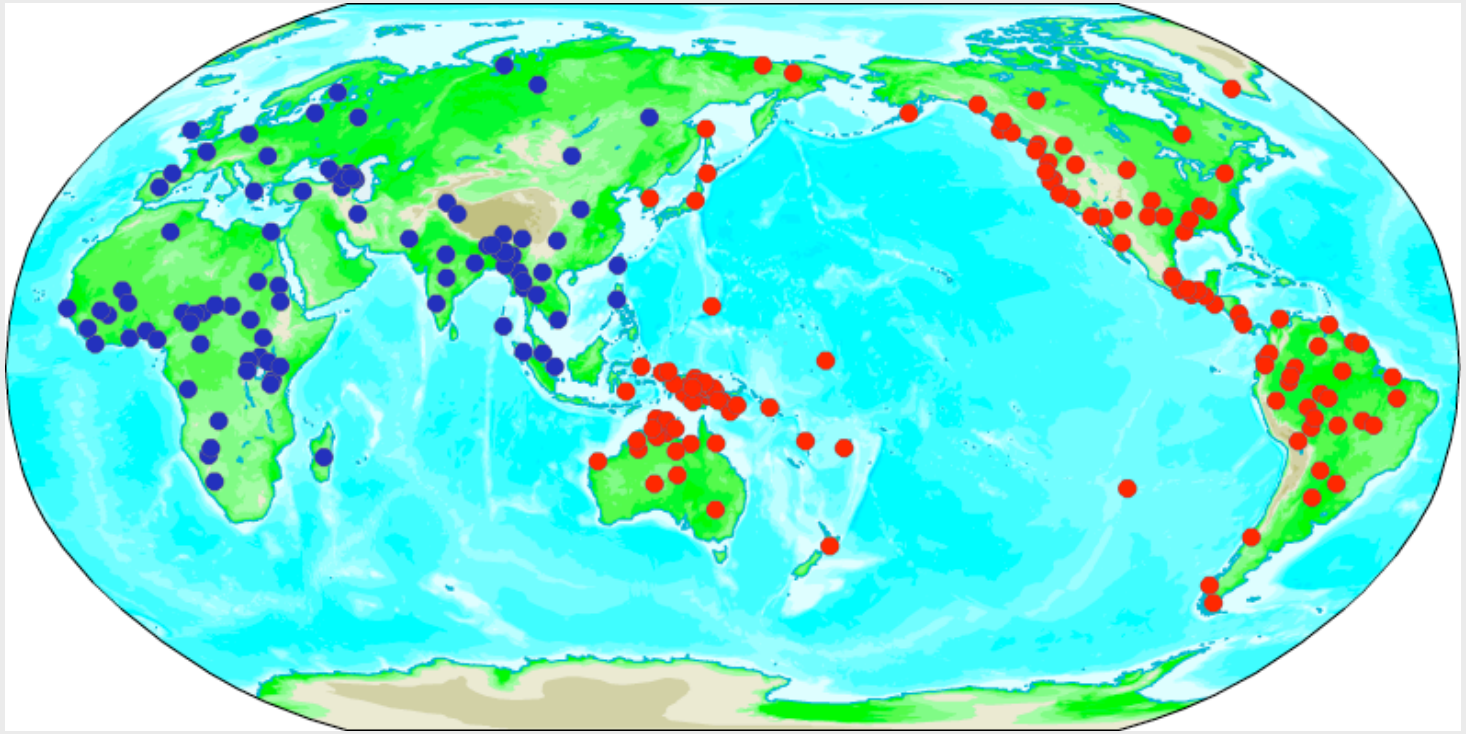
Areal factors: theory and hypotheses

- Main factors to test (a regular part of the AUTOTYP database system)
  - Circumpacific (CP) vs. rest of the world (CP_Rest)
  - Ambiguous position of SEA, yielding two factor sets
    - o SEA belongs to the CP (East)



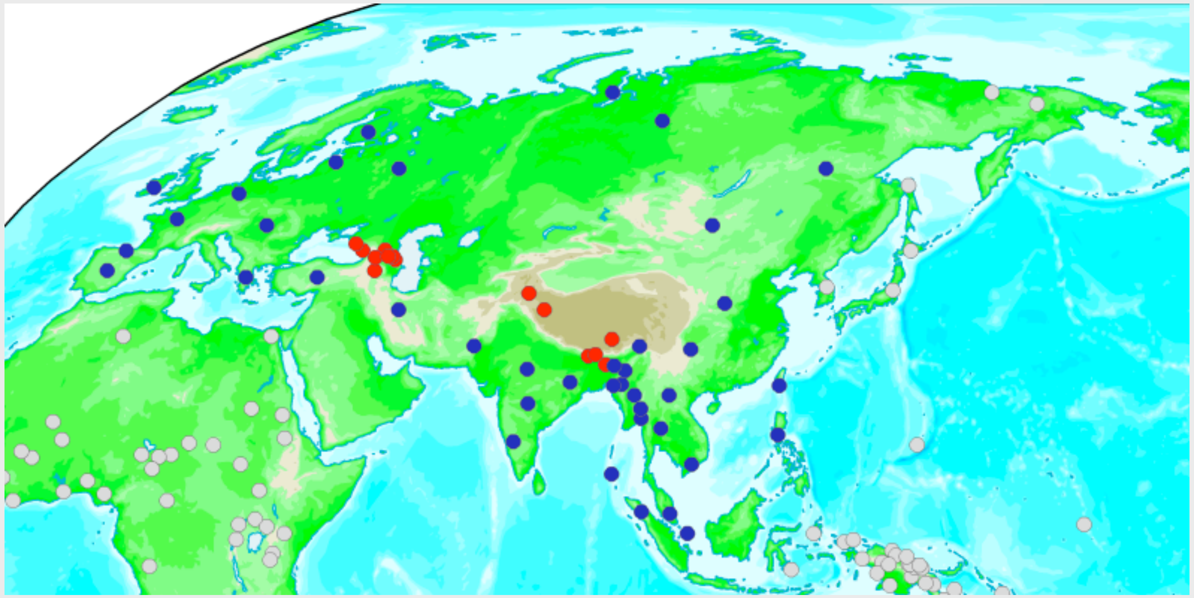
Areal factors: theory and hypotheses

- Main factors to test (a regular part of the AUTOTYP database system)
  - Circumpacific (CP) vs. rest of the world (CP_Rest)
  - Ambiguous position of SEA, yielding two factor sets
    - o SEA belongs to Eurasia (West)



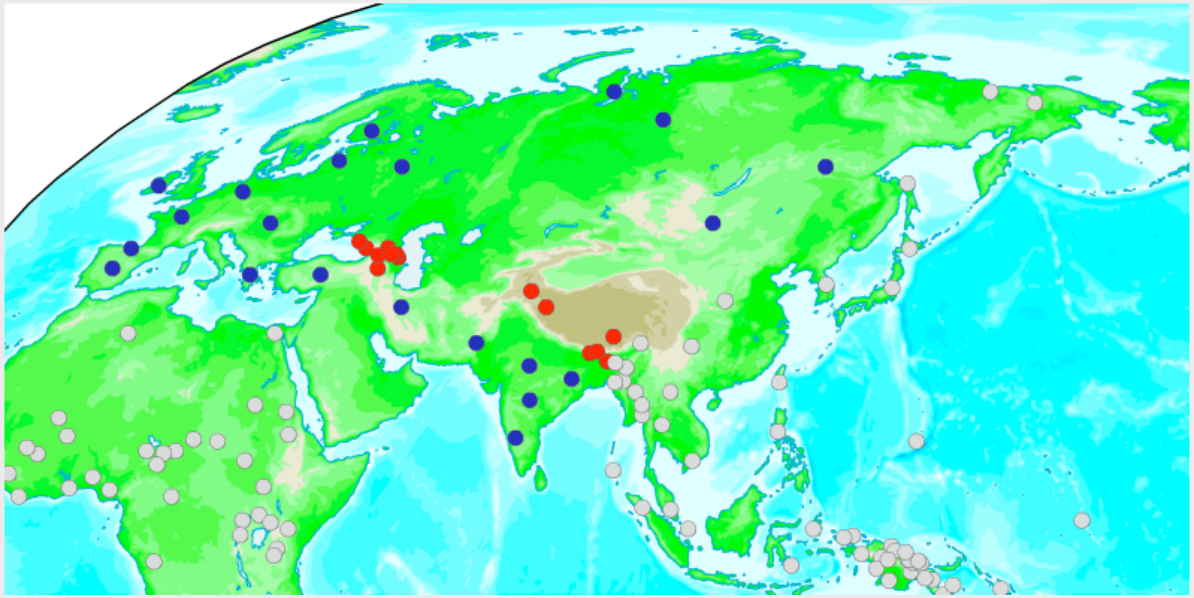
Areal factors: theory and hypotheses

- Main factors to test (a regular part of the AUTOTYP database system)
  - Circumpacific (CP) vs. rest of the world (CP_Rest)
  - Ambiguous position of SEA, yielding two factor sets
  - Eurasia vs. rest of the world (Eur_Rest W, E)
  - Europe vs. rest of the world (europ)
  - Enclaves in Eurasia (both at once; EEn2_0, W, E)



Areal factors: theory and hypotheses

- Main factors to test (a regular part of the AUTOTYP database system)
  - Circumpacific (CP) vs. rest of the world (CP_Rest)
  - Ambiguous position of SEA, yielding two factor sets
  - Eurasia vs. rest of the world (Eur_Rest W, E)
  - Europe vs. rest of the world (europ)
  - Enclaves in Eurasia (both at once; EEn2_0, W, E)



Areal factors: theory and hypotheses

- Main factors to test (a regular part of the AUTOTYP database system)
  - Circumpacific (CP) vs. rest of the world (CP_Rest)
  - Ambiguous position of SEA, yielding two factor sets
  - Eurasia vs. rest of the world (Eur_Rest W, E)
  - Europe vs. rest of the world (europ)
  - Enclaves in Eurasia (both at once; EEn2_0, W, E)
- Coding: assign WALS languages to these factors using *ArcView*

WALS Variables

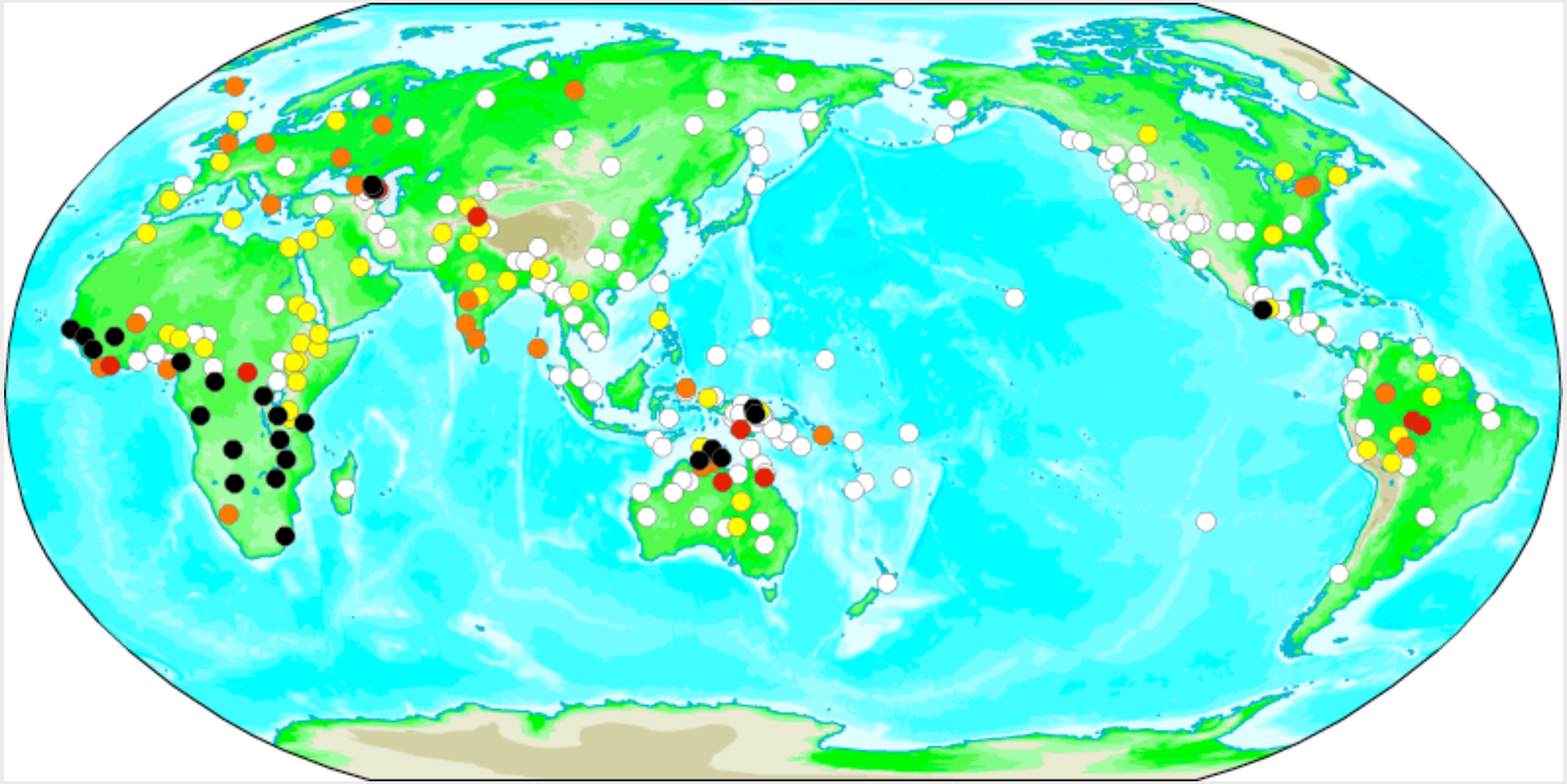
- Test the variables with no more than 30% missing values (i.e. at least 126 datapoints) in WALSG
- This is 33 (out of 142) variables (WALS chapters).
- Recoding of most variables in order to increase sample density, e.g. in COMALN5, collapse marked and unmarked nominative
  - Take out zeros, e.g. ‘no adpositions’ in BAKADP
  - Take out cases with no dominant pattern (e.g. word order)
- Applying the most obvious and linguistically meaningful re-codings yields 68 variables
- Test our own variables using the larger GEN sample from AUTOTYP (synthesis, *N*=202; possessive classes, *N*=236; locus, *N*=245)
- For practical reasons, we limit tests to about half of the variables, about one per chapter

Results

Variable	CP_RestW	CP_RestE	Eur_RestW	Eur_RestE	europ	NW_OW	EEn2_0W	EEn2_0E
ANDANG	7.44,p=.024	ns	5.20,p=.081	8.19, p=.0164	6.45, p=.0442	34.96, p=1e-04	ns	ns
AUWEPI	45.56,p=1e-04	38.79,p=1e-04	21.31, p=1e-04	19.33,p=2e-04	32.16,p=1e-04	23.48,p=1e-04	ns	ns
AUWHOR	7.01,p=.074	11.94,p=.006	ns	ns	ns	ns	ns	ns
AUWIMP	ns	ns	ns	ns	12.067,p=.021	ns	ns	ns
AUWIMP2 (±dedicated IMP)	ns	ns	ns	4.97, p=.016	ns	ns	ns	ns
AUWPRH	ns	ns	ns	ns	9.72,p=.022	ns	ns	ns
AUWPRH22 (±imp mph)	ns	ns	ns	ns	ns	ns	ns	ns
BAECSY01 (without 0)	9.40,p=.009	10.99,p=.004	10.61, p=.004	13.22,p=.002	10.35,p=.0001	5.121,p=.0885	ns	ns
BAKADP01 (without 0, 3=4)	ns	ns	5.53, p=.013	ns	ns	ns	ns	ns
COMALN5 (collapse ACC)	14.59,p=.003	14.14,p=.004	15.11,p=.003	20.25,p=.001	16.38,p=.024	ns	17.75,p=5e-04	13.91,p=.002
CORNUM (scalar)	F=3.46, p=.080	F=7.92, p=.005	ns	ns	ns	F=3.56, p=.069	F=6.44 ,p=.013	ns
CORSEX01 (±gender)	p=.023	p=.003	ns	ns	ns	p=.089	ns	ns
CYSIND2 (±incl)	p=.050	p=.006	ns	p=.032	p=.008	ns	ns	ns
DOBOPT	ns	ns	p=.0003	p=.0003	ns		p=0.010	p=.031
DRYNEG2 (±double neg)	ns	ns	ns	ns	ns	ns	ns	ns
DRYPOS2 (±poss. affixes)	p=.009	ns	p=.014	ns	p=.039	p=.013	ns	ns
DRYRAO0 (w/o free order)	14.08,p=.002	9.78,p=.020	10.63,p=.015	27.58,p=1e-04	8.41,p=.034	9.98,p=.017	11.79,p=.009	8.22,p=.042
DRYSBV (w/o free order)	p=.013	p=.016	p=.030	p=.018	ns	p=.0002	ns	ns
DRYTAA2 (±TA infl)	ns	ns	ns	p=.009	ns	ns		all have TA
HAAEVD2 (±evidentials)	p=.002	p=.019	ns	ns	ns	p=3.4e-05	ns	ns
HAJNAS	ns	ns	ns	ns	ns	p=.005	ns	ns
IGGNUM0 (scalar; w/o 'none')	ns	ns	ns	F=7.05, p=.008	ns	ns	F=4.25, p=.047	ns
LOCUS_P (w/o NA; GEN sample)	p=9.752e-09	p=5.557e-06	p=4.159e-06	p=.001	ns	ns	ns	ns
LOCUS POSS (w/o NA;GEN sample)	p=2.484e-09	p=1.706e-06	p=3.505e-05	p=.002	p=.020	ns	ns	ns
MADFRV2 (± front rd V)	p=.010	p=.008	p=8.7e-05	p=4.6e-05	p=2.56e-05	ns	ns	ns
MADLAT2 (±laterals)	p=4.0e-06	p=6.8e-05	p=.0003	p=.003	ns	p=.003	ns	ns
MADTON02 (±tone)	p=.003	p=.058	ns	p=.0001	ns	ns	p=.084	ns
MADUVU2 (±uvulars)	ns	ns	p=.028	p=0.009	ns	ns	p=.002	p=.005
MADVOI2 (±voicing)	p=1.4e-11	p=1.1e-14	p=1.62e-08	p=5.2e-07	p=.014	p=.002	ns	all voiced
POSSCL (±posscl; GEN sample)	p=9.7e-07	p=.001	p=1.5e-11	p=4.6e-07	p=.0006		p=0.006	p=.022
SIEPAS (±passive)	12.47,p=4e-04	15.81,p=3e-04	6.47,p=.009	11.11,p=2e-04	11.57,p<.0001	.04,ns	3.27,p=.067	p=.001
SIEZER2 (±Sagr)	p=.001	ns	p=.068	ns	ns	p=.016	ns	p=.004
SYN (scalar; fpw+cpw; GEN sample)	F=23.81, p=1e-04	F=11.72, p=6e-04	F=15.05, p=2e-04	F=6.18, p=.015	F=5.89, p=.016		F=8.39, p=.006	F=8.26, p=.004
SONNON2 (±periphr caus)	ns	ns	ns	ns	ns	p=.037	ns	ns

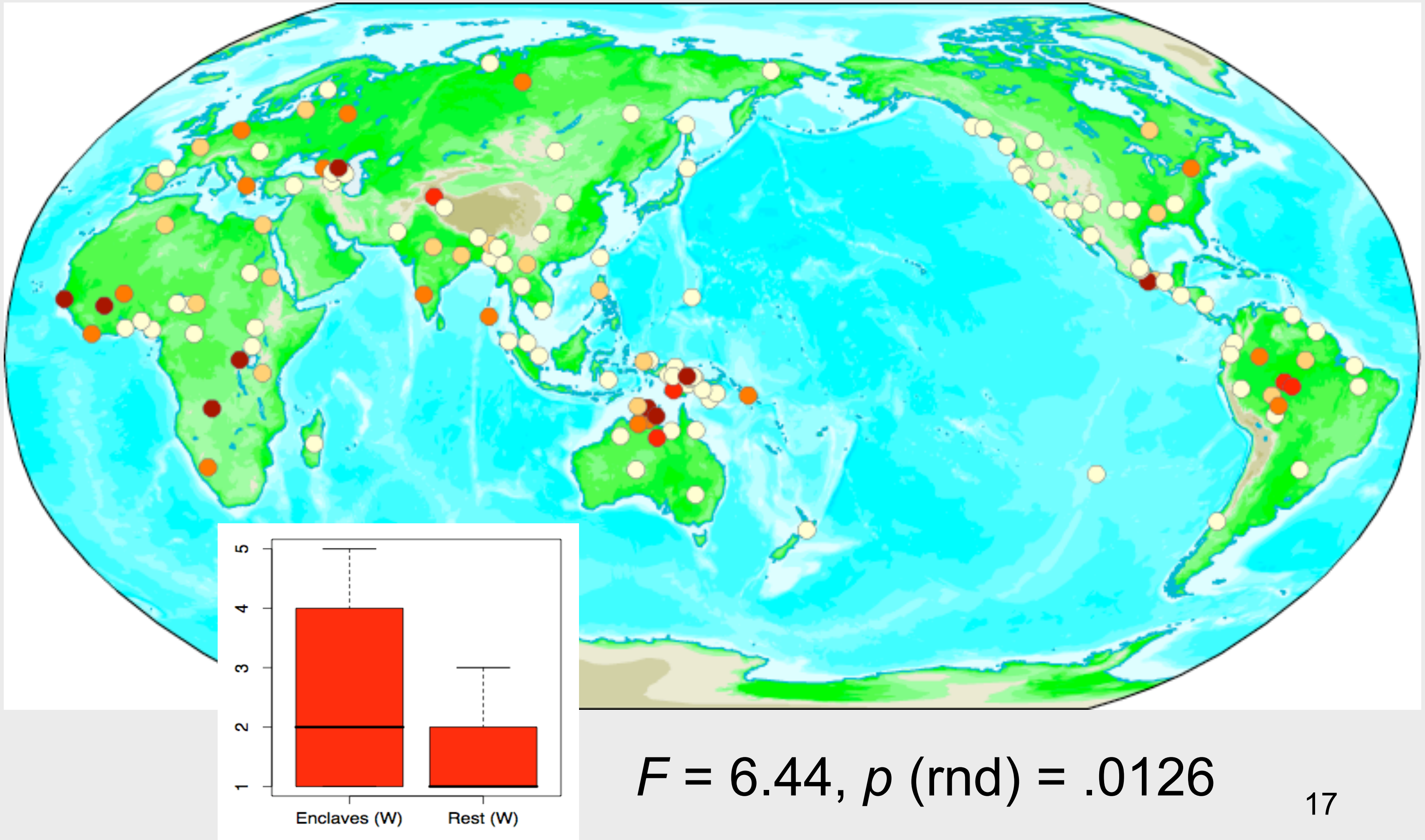
Result: examples

CORNUM unsampled, from WALS (*N*=256)



Result: examples

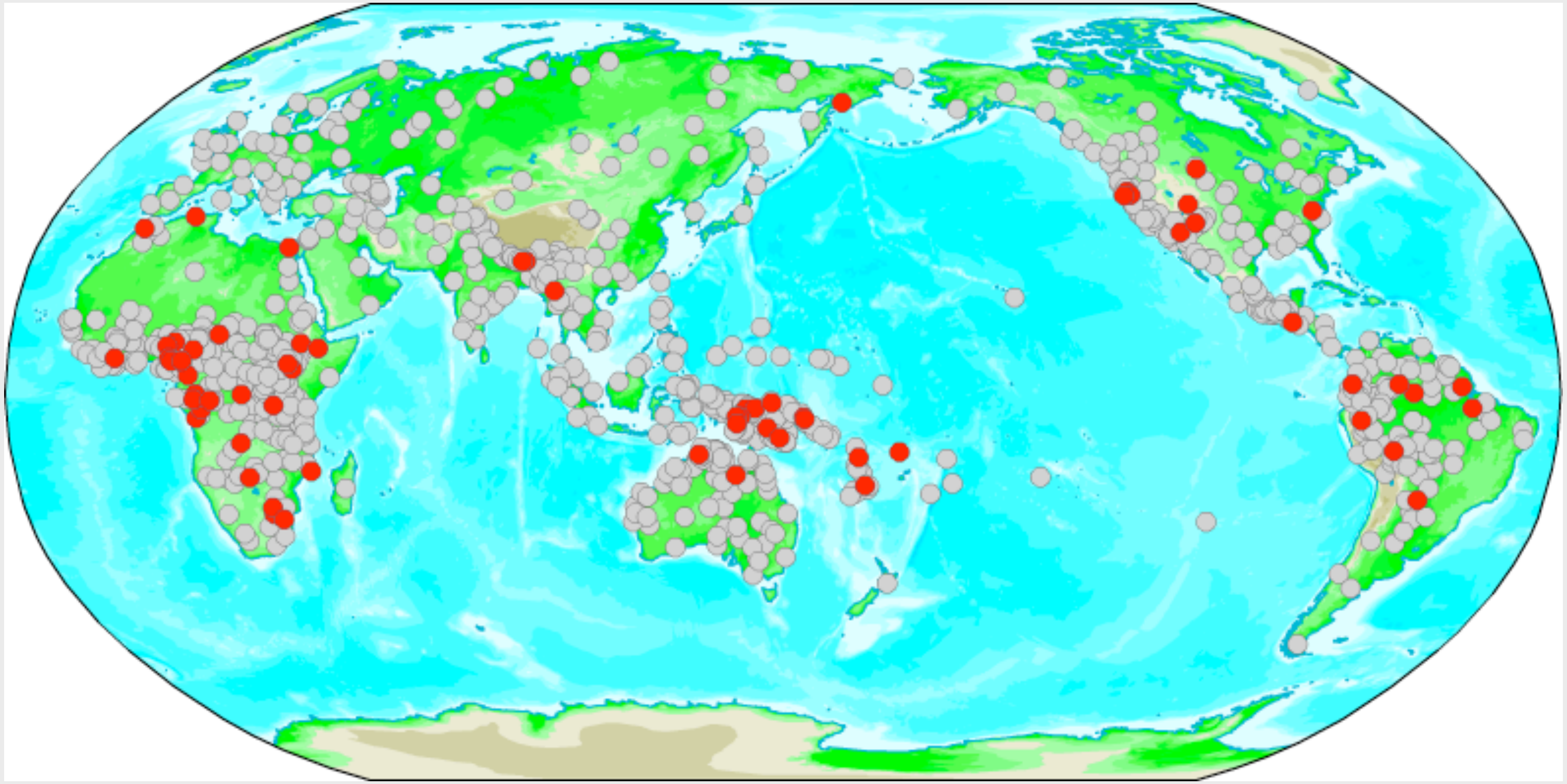
CORNUM in WALSG sample (*N*=143)



F = 6.44, *p* (rnd) = .0126

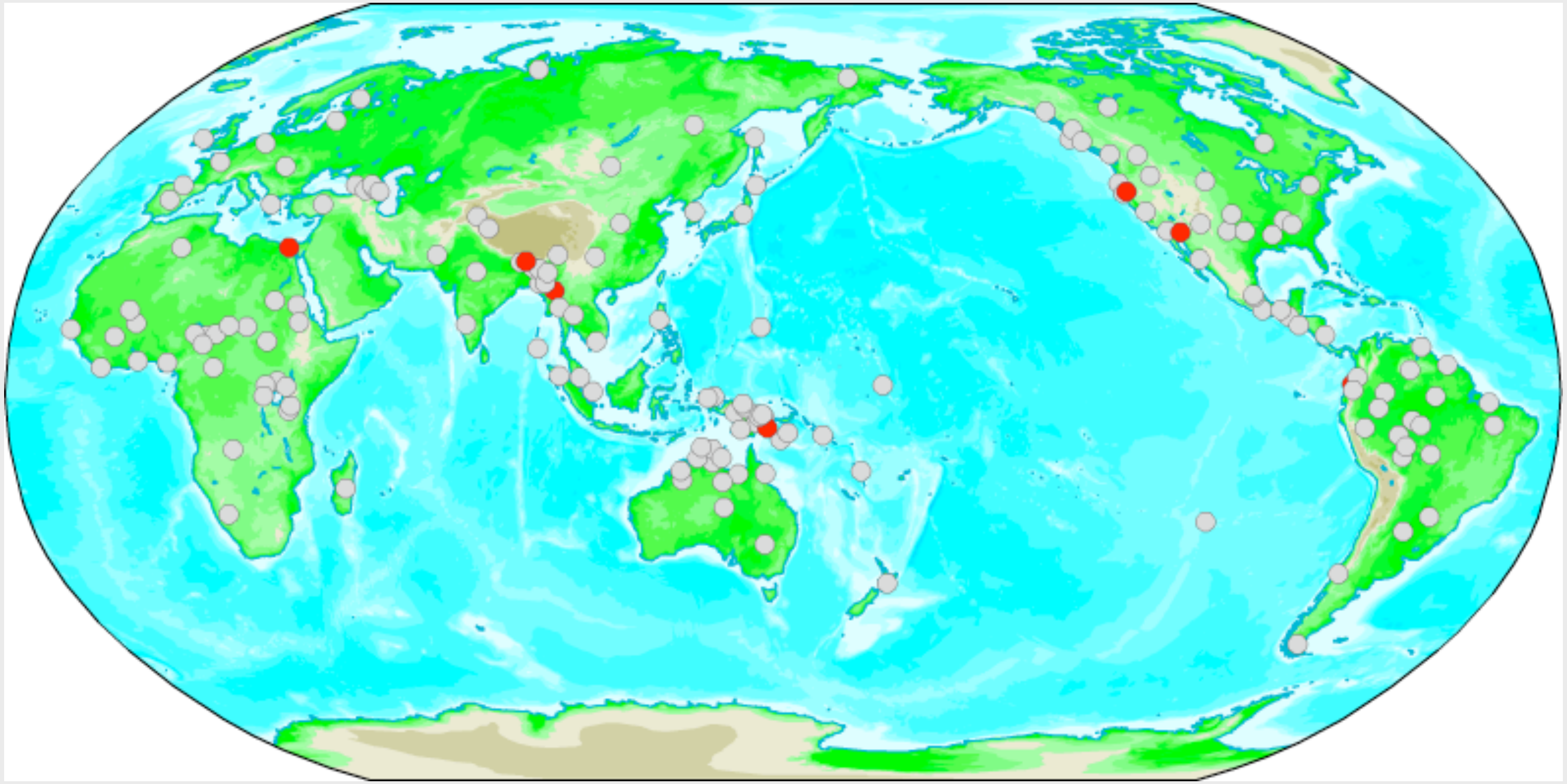
Result: examples

DRYNEG2 unsampled, from WALS (*N*=1011)



Result: examples

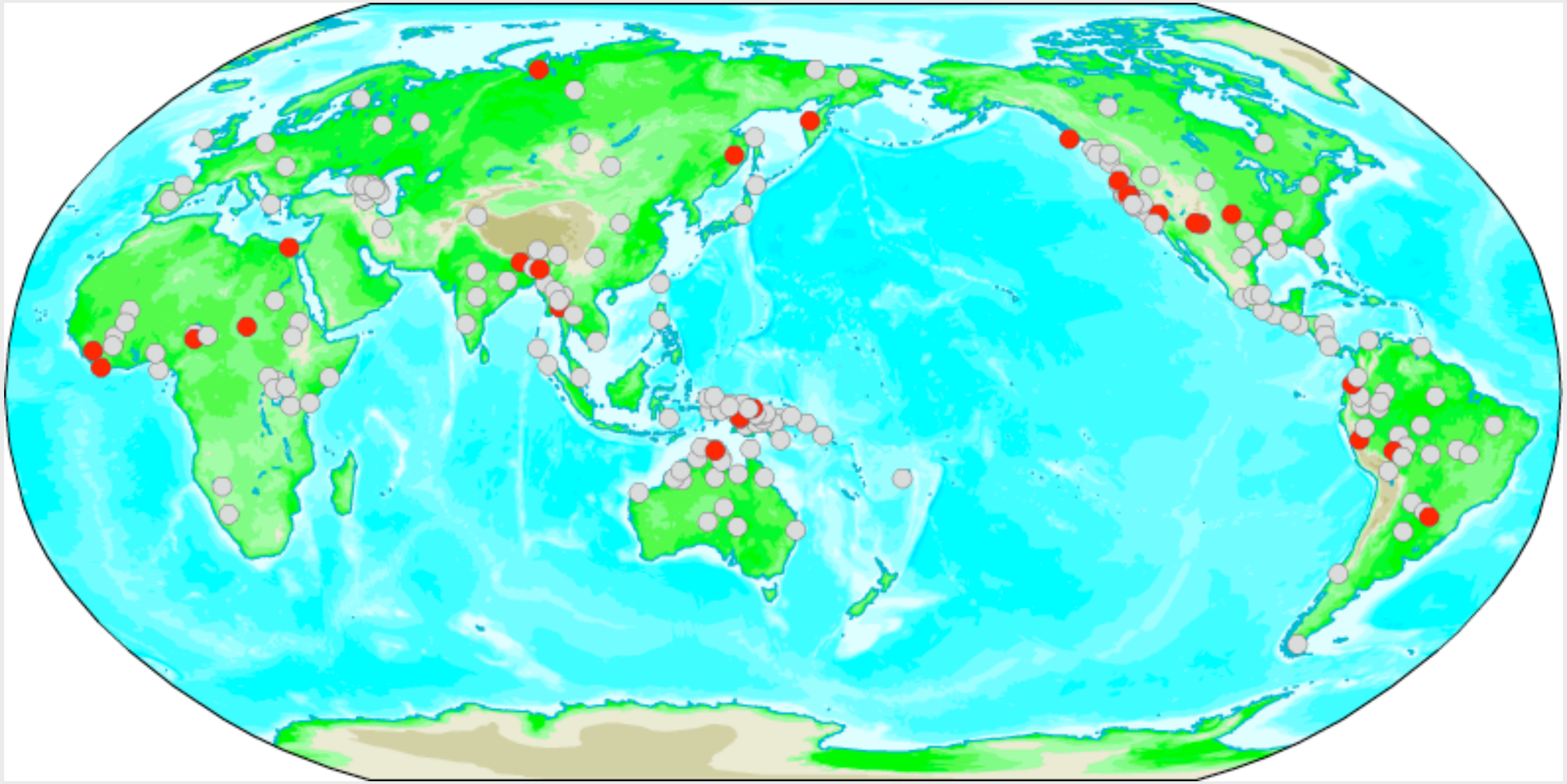
DRYNEG2 in WALSG sample (*N*=162)



for all hypotheses tested: *p* (FE) > .05 (ns)

Result: examples

Replication: AUTOTYP simul/circum vs. other NEG (GEN, *N*=203)



for all hypotheses tested: *p* (FE) > .05 (ns)

Conclusions

- WALS contains many areal signals supporting hypotheses on distributional effects of
  - Circumpacific
  - Eurasia
  - Enclaves in Eurasia: Caucasus and Himalayas
- For many variables, it does not make a difference whether SEA is counted with Eurasia or the Circumpacific macroarea.
- The data we looked at provides only very little evidence for distinctively *European* as apposed to *Eurasian* effects.
- Areal effects are often at odds with the visual impression gained from the non-sampled (printed) maps
- Distributions are numbers, not pictures.

Acknowledgments

- **The AUTOTYP research team (as of July 2005)**
  - Johanna Nichols (Co-Director, Berkeley)
  - Balthasar Bickel (Co-Director, Leipzig)
  - Kristine Hildebrandt (Post-Doc, Leipzig)
  - Tracy Alan Hall (Research Associate, Bloomington)
  - Fernando Zúñiga (Research Associate, Santiago)
  - *RAs in Berkeley:* Gabriela Caballero, Nicole Marcus, Suzanne Wilhite
  - *RAs in Leipzig:* **Anja Gampe**, Sebastian Hellmann, **Jenny Seeg**, Thomas Goldammer, Michael Riessler, **Sven Siegmund**, Sindy Poppitz, Franziska Crell, Kathi Stutz, Josh Wilbur
  - *Past team members:* Rebecca Voll, Alena Witzlack-Makarevich, Sandra Biewald, Aimee Lahaussois-Bartosik, Dave Peterson, Keith Sanders

- **Funding:** Swiss NSF Grant Nos. 08210-053455 (1998-2001, Bickel) and 610-0627 (2001-2002, Bickel), German DFG Grant No. BI 799/2-1 (2003-2005, Bickel & Hall), US NSF Grant No. 96-16448 (1998-2001, Nichols)