Articles are inversely correlated with case in Indo-European

David Goldstein

20 June 2025

44th East Coast Indo-European Conference, Ludwig-Maximilians-Universität München

Introduction

“The typological revolution” in IE (Hewson and Bubeník 2006)

- “Free” word order → fixed configuration

“The typological revolution” in IE (Hewson and Bubeník 2006)

- “Free” word order → fixed configuration
- Rich morphology → analytic structure

“The typological revolution” in IE (Hewson and Bubeník 2006)

- “Free” word order → fixed configuration
- Rich morphology → analytic structure
- Case-marking → adpositions and syntax

The debate

- Is the size of case inventories inversely correlated with the presence of articles?

The debate

- Is the size of case inventories inversely correlated with the presence of articles?
- An association has repeatedly been claimed for certain languages and clades within Indo-European, in particular Germanic (Paul 1919, §149; Tschirch 1975, p. 168; Giusti 1995) and Romance (Giusti 1995; Posner 1996, p. 126; Ledgeway 2011, p. 718)

The debate

- Is the size of case inventories inversely correlated with the presence of articles?
- An association has repeatedly been claimed for certain languages and clades within Indo-European, in particular Germanic (Paul 1919, §149; Tschirch 1975, p. 168; Giusti 1995) and Romance (Giusti 1995; Posner 1996, p. 126; Ledgeway 2011, p. 718)
- Others have denied the relationship by pointing out counterexamples and questioning its motivation (Wood 2003, p. 74; Barðdal 2009, p. 131; Börjars et al. 2016, e31)

Two distinct questions

- The nature of the correlation

  Is the correlation between case and articles deterministic or probabilistic?

Two distinct questions

- The nature of the correlation

  Is the correlation between case and articles deterministic or probabilistic?

- Causal relationship

  Does the reduction in case inventory **cause** the grammaticalization of definite articles?

The correlation is not deterministic

- Counterexamples in both directions

The correlation is not deterministic

- Counterexamples in both directions
- Classical Armenian has both a rich case inventory and a definite article

The correlation is not deterministic

- Counterexamples in both directions
- Classical Armenian has both a rich case inventory and a definite article
- Nepali (Indic) has neither case nor articles

The correlation is not deterministic

- Counterexamples in both directions
- Classical Armenian has both a rich case inventory and a definite article
- Nepali (Indic) has neither case nor articles
- The relationship between case and articles is not deterministic

The correlation is not deterministic

- Counterexamples in both directions
- Classical Armenian has both a rich case inventory and a definite article
- Nepali (Indic) has neither case nor articles
- The relationship between case and articles is not deterministic
- Not clear from these observations alone if it's probabilistic

Previous claims of a probabilistic relationship are invalid

- **Insufficient data**

  Previous claims have been based on limited data (e.g., Romance, Germanic)

Previous claims of a probabilistic relationship are invalid

- **Insufficient data**

  Previous claims have been based on limited data (e.g., Romance, Germanic)

- **Galton's problem unaddressed**

  The previous literature does not take into account non-independence due to spatial proximity and shared ancestry

Previous claims of a probabilistic relationship are invalid

- **Insufficient data**

  Previous claims have been based on limited data (e.g., Romance, Germanic)

- **Galton's problem unaddressed**

  The previous literature does not take into account non-independence due to spatial proximity and shared ancestry

- **Methodological mismatch**

  Estimating probabilistic relationships requires statistical methods (Evans and Levinson 2009, p. 439; Ladd et al. 2015, p. 223)

Today's questions

1. Does the evidence support a statistical association between case-inventory size and the presence of a definite article in Indo-European?

Today's questions

1. Does the evidence support a statistical association between case-inventory size and the presence of a definite article in Indo-European?
2. Does the evidence support a statistical association between case inventory and the presence of an indefinite article in Indo-European?

Today's questions

1. Does the evidence support a statistical association between case-inventory size and the presence of a definite article in Indo-European?
2. Does the evidence support a statistical association between case inventory and the presence of an indefinite article in Indo-European?
3. Is there are also a statistical association between definite and indefinite articles?

Today's questions

1. Does the evidence support a statistical association between case-inventory size and the presence of a definite article in Indo-European?
2. Does the evidence support a statistical association between case inventory and the presence of an indefinite article in Indo-European?
3. Is there are also a statistical association between definite and indefinite articles?
4. If there is a statistical association between case inventory and articles, is it a causal relationship?

Claims

1. Articles are inversely associated with nominal case in Indo-European

Claims

1. Articles are inversely associated with nominal case in Indo-European
2. The association is **not** between articles and absence of case: as cases are lost, the probability of articles increases

Claims

1. Articles are inversely associated with nominal case in Indo-European
2. The association is **not** between articles and absence of case: as cases are lost, the probability of articles increases
3. The correlation between definite and indefinite articles is upheld even when controlling for case inventory

Claims

1. Articles are inversely associated with nominal case in Indo-European
2. The association is **not** between articles and absence of case: as cases are lost, the probability of articles increases
3. The correlation between definite and indefinite articles is upheld even when controlling for case inventory
4. The question of causation has not been properly understood and remains open

Roadmap

1. Case and articles in Indo-European

Roadmap

1. Case and articles in Indo-European
2. Data

Roadmap

1. Case and articles in Indo-European
2. Data
3. Methods

Roadmap

1. Case and articles in Indo-European
2. Data
3. Methods
4. Results

Roadmap

1. Case and articles in Indo-European
2. Data
3. Methods
4. Results
5. Discussion

Roadmap

1. Case and articles in Indo-European
2. Data
3. Methods
4. Results
5. Discussion
6. Final thoughts

Case and articles in Indo-European

Case inventories

- Eight cases (including vocative) standardly reconstructed to Proto-Indo-European

Case inventories

- Eight cases (including vocative) standardly reconstructed to Proto-Indo-European
- Case inventories generally decrease over time

Case inventories

- Eight cases (including vocative) standardly reconstructed to Proto-Indo-European
- Case inventories generally decrease over time
- There are exceptions to this general trend (e.g., Old Lithuanian; Kulikov 2012, pp. 295–296; Kim 2012, p. 125)

Article inventories in Indo-European

(3) Russian

krasivyj novyj derevjannyj **dom**
beautiful.NOM.SG new.NOM.SG wooden.NOM.SG house.NOM.SG

'**a**/**the** beautiful new wood **house**' (Bailyn 2012, p. 45)

(4) Old Irish

a. **in** macc
DEF boy
'**the** boy'

b. macc
boy
'**a** boy'

(5) Persian

a. ketab
book
'The book'

b. ketab-**i**
book-INDEF
'**A** book'

(6) English

a. **the** farmer

b. **a** farmer

The history of articles in Indo-European (Goldstein 2022)

- All articles emerge within the so-called major clades (Indic, Celtic, Germanic, etc.)

The history of articles in Indo-European (Goldstein 2022)

- All articles emerge within the so-called major clades (Indic, Celtic, Germanic, etc.)
- They are thus a relatively recent trait and not reconstructed to Proto-Indo-European

The history of articles in Indo-European (Goldstein 2022)

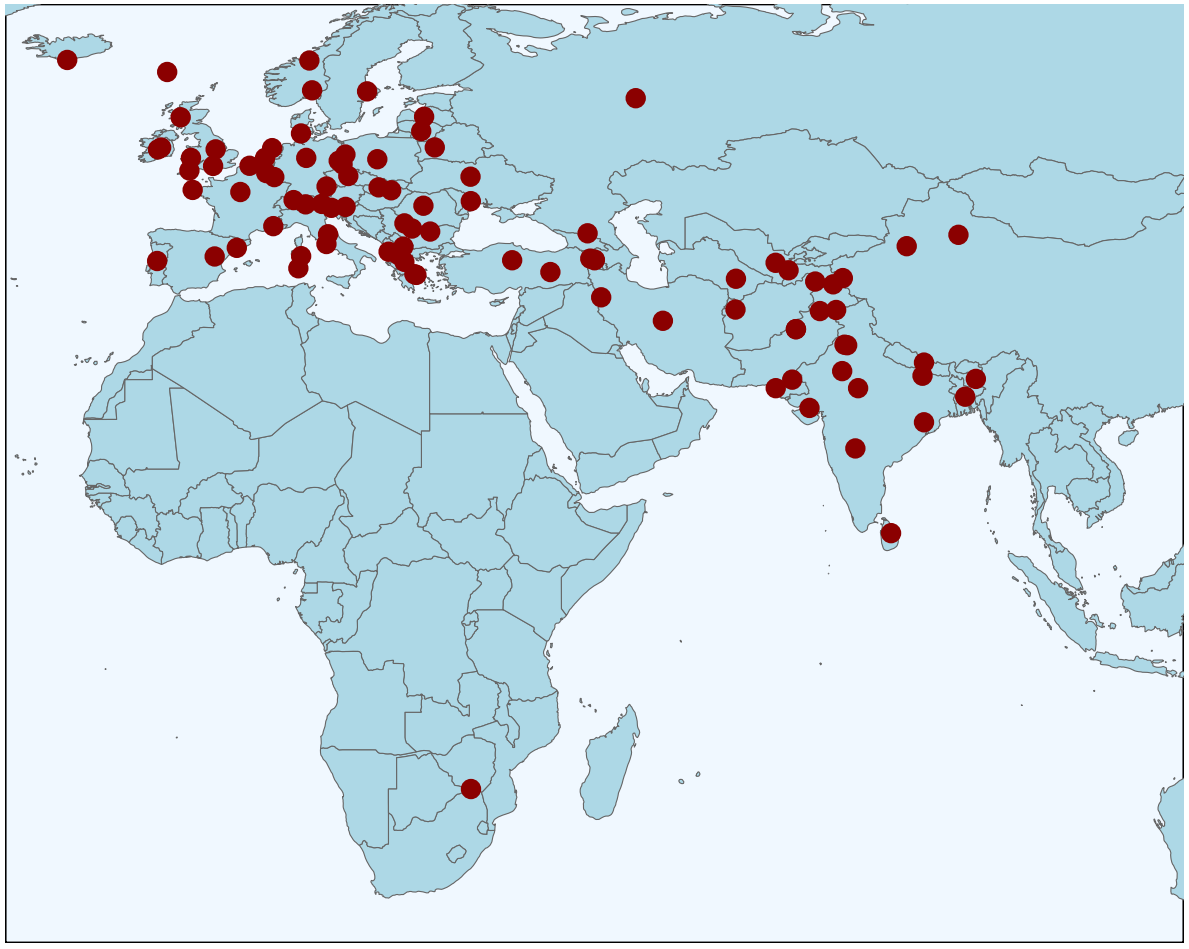
- All articles emerge within the so-called major clades (Indic, Celtic, Germanic, etc.)
- They are thus a relatively recent trait and not reconstructed to Proto-Indo-European
- In languages with both definite and indefinite articles, definite articles emerge first where we have evidence (e.g., Greek, Armenian, Germanic)

The history of articles in Indo-European (Goldstein 2022)

- All articles emerge within the so-called major clades (Indic, Celtic, Germanic, etc.)
- They are thus a relatively recent trait and not reconstructed to Proto-Indo-European
- In languages with both definite and indefinite articles, definite articles emerge first where we have evidence (e.g., Greek, Armenian, Germanic)
- No evidence for loss of articles in my study group

Data

Geographic distribution of sampled languages ($N = 94$)



Measurement criteria for case

1. Only synthetic case markers were considered; analytic case markers such as adpositions were not (Blake 2001, pp. 9–12)

Measurement criteria for case

1. Only synthetic case markers were considered; analytic case markers such as adpositions were not (Blake 2001, pp. 9–12)
2. Only cases on nouns (as opposed to pronouns) or determiners were considered

Measurement criteria for case

1. Only synthetic case markers were considered; analytic case markers such as adpositions were not (Blake 2001, pp. 9–12)
2. Only cases on nouns (as opposed to pronouns) or determiners were considered
3. Vestigial case markers were excluded (e.g., Gothic instrumental)

Measurement criteria for case

1. Only synthetic case markers were considered; analytic case markers such as adpositions were not (Blake 2001, pp. 9–12)
2. Only cases on nouns (as opposed to pronouns) or determiners were considered
3. Vestigial case markers were excluded (e.g., Gothic instrumental)
4. Vocatives excluded

Measurement criteria for case

1. Only synthetic case markers were considered; analytic case markers such as adpositions were not (Blake 2001, pp. 9–12)
2. Only cases on nouns (as opposed to pronouns) or determiners were considered
3. Vestigial case markers were excluded (e.g., Gothic instrumental)
4. Vocatives excluded
5. Variation in case exponence is ignored

The lower bound on case inventories is 1

- This approach to measuring case inventories essentially counts the number of rows in nominal paradigms

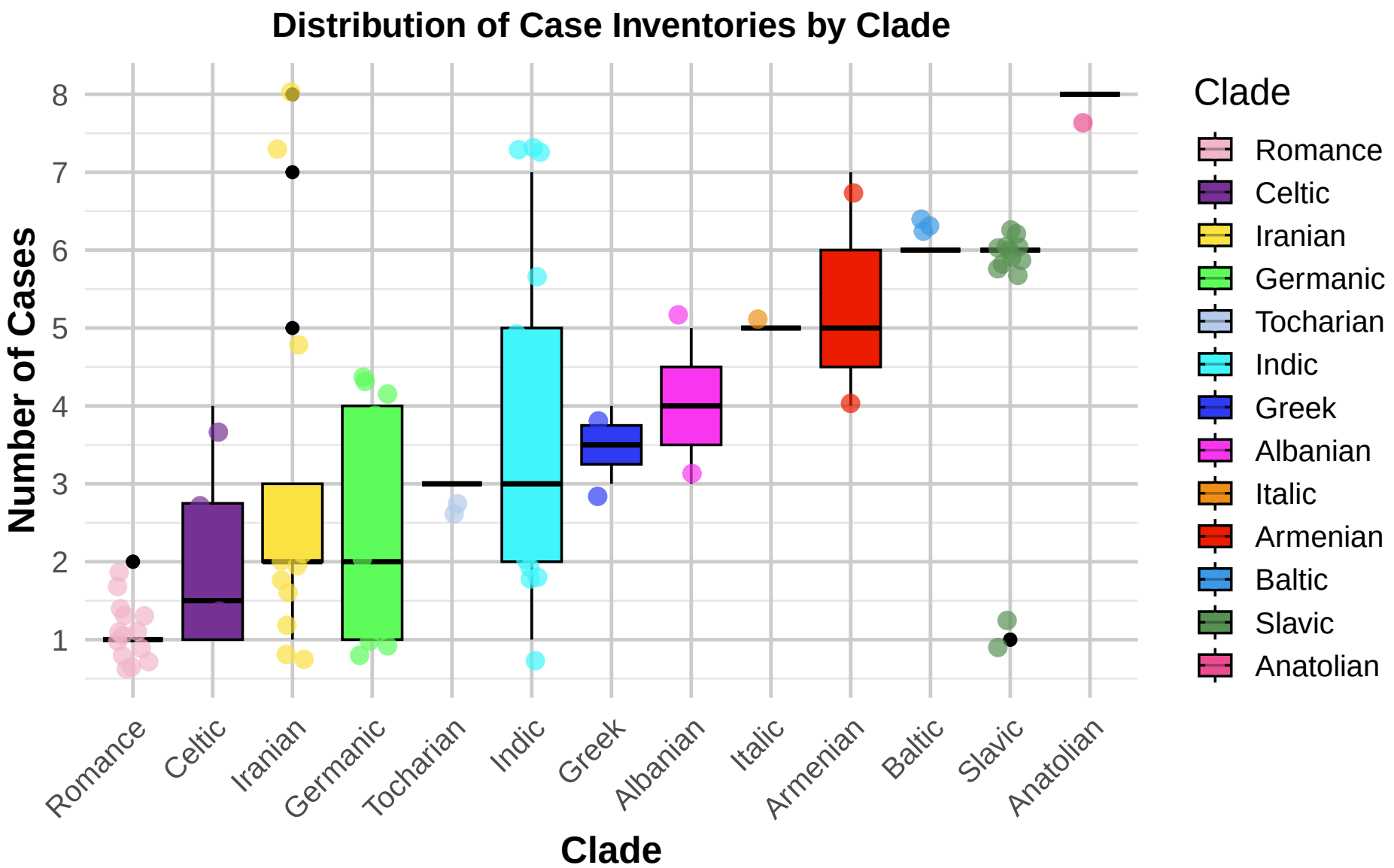
The lower bound on case inventories is 1

- This approach to measuring case inventories essentially counts the number of rows in nominal paradigms
- The lower bound for case inventories is therefore 1—not 0

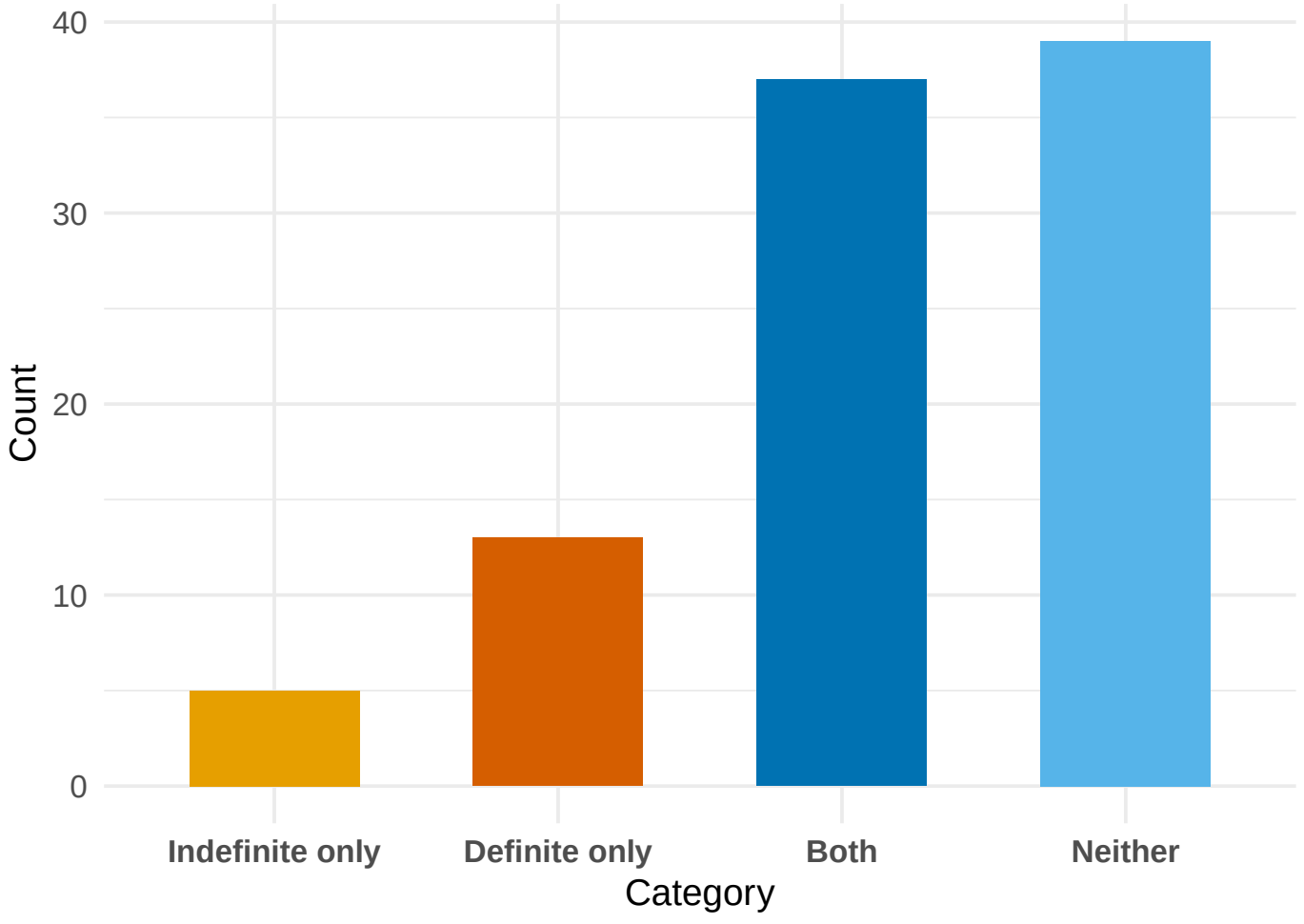
The lower bound on case inventories is 1

- This approach to measuring case inventories essentially counts the number of rows in nominal paradigms
- The lower bound for case inventories is therefore 1—not 0
- Counting this away avoids the discontinuity 0, 2, 3, ...

Phylogenetic distribution of case inventories



Frequency distribution of article inventories



Methods

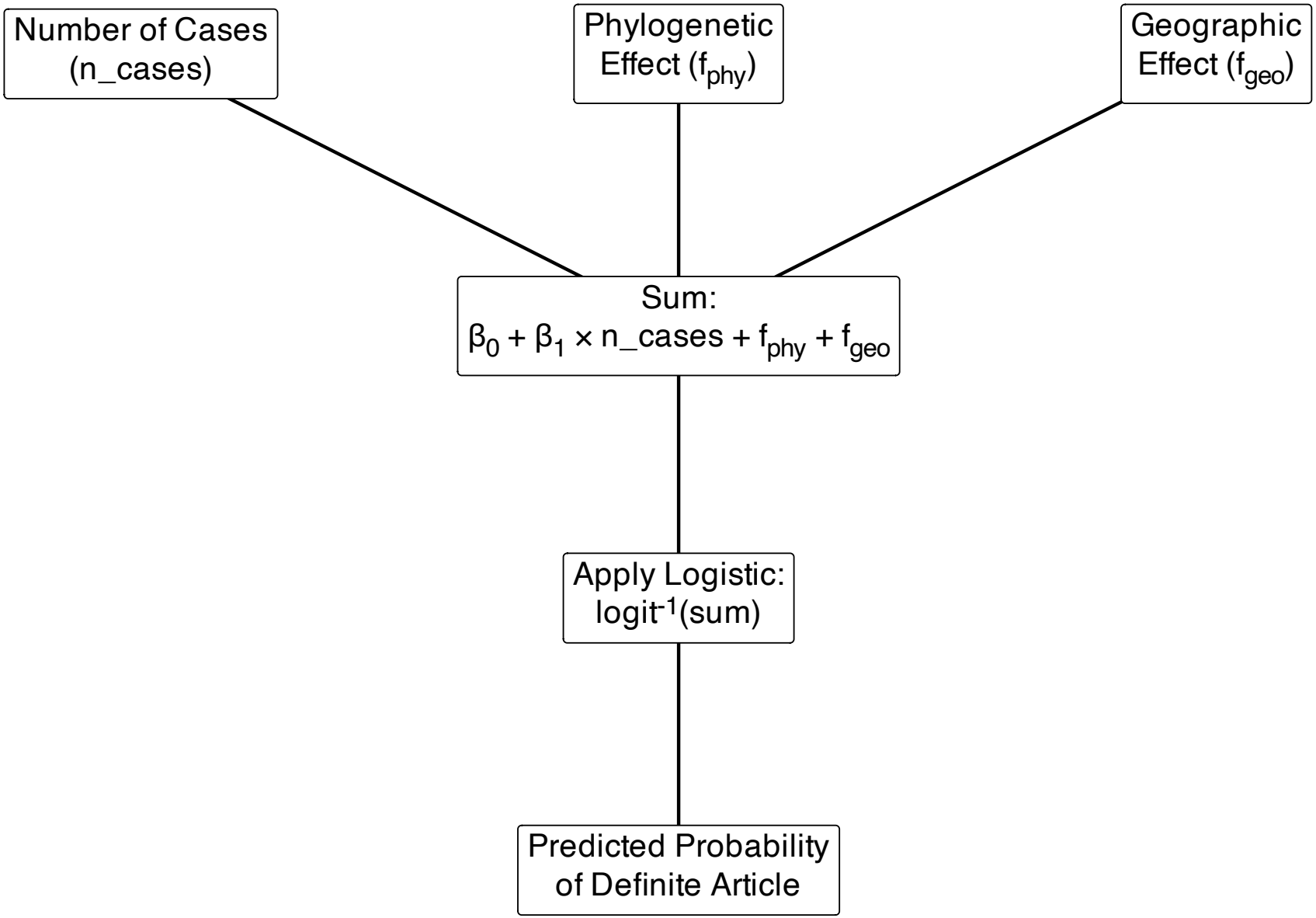
Description of model

- Two Bayesian logistic regression submodels with Gaussian processes for spatial and phylogenetic autocorrelation (Guzmán Naranjo and Becker 2021; Guzmán Naranjo and Mertner 2022)

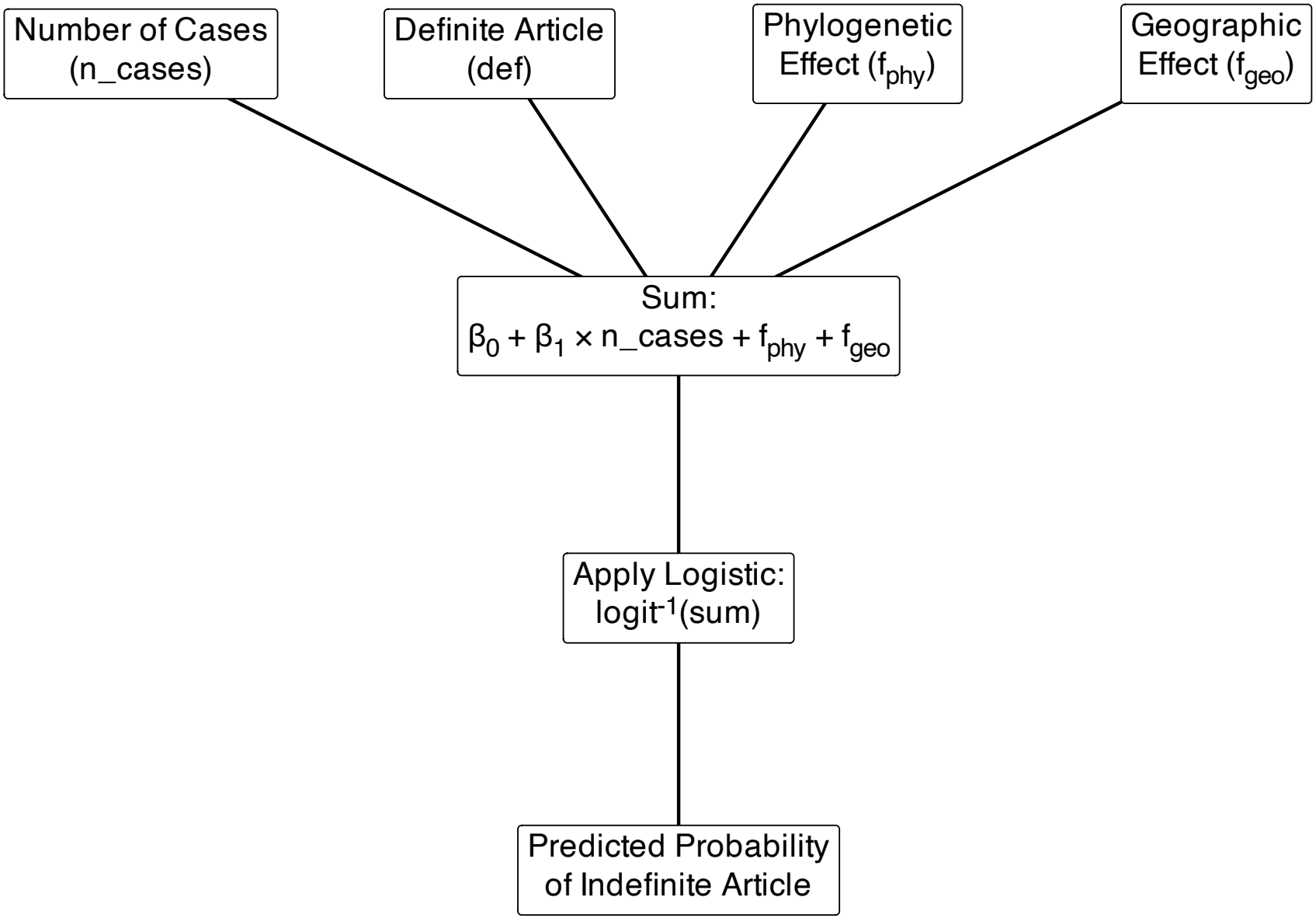
Description of model

- Two Bayesian logistic regression submodels with Gaussian processes for spatial and phylogenetic autocorrelation (Guzmán Naranjo and Becker 2021; Guzmán Naranjo and Mertner 2022)
- Stan and `cmdstanr`

Definite submodel

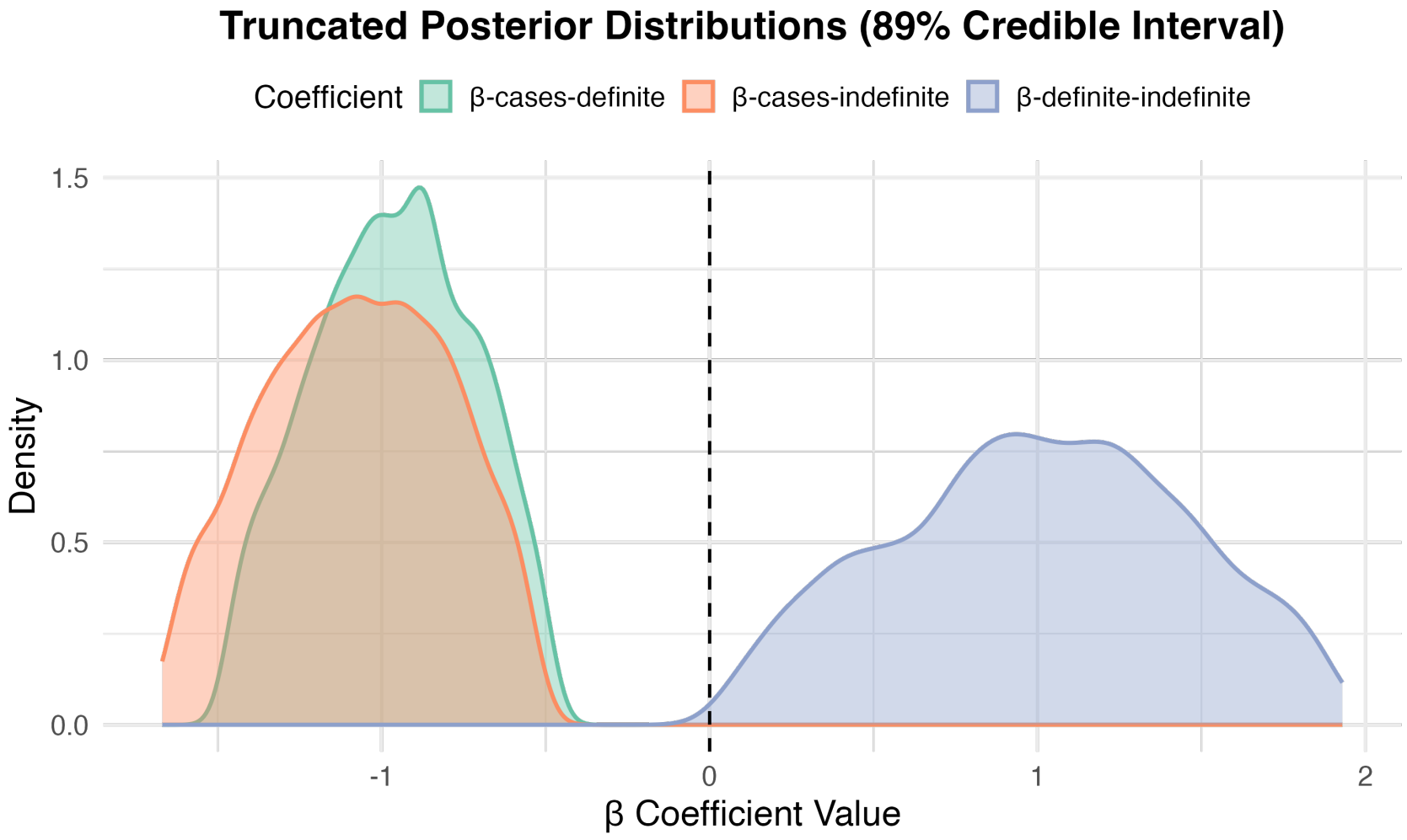


Indefinite submodel

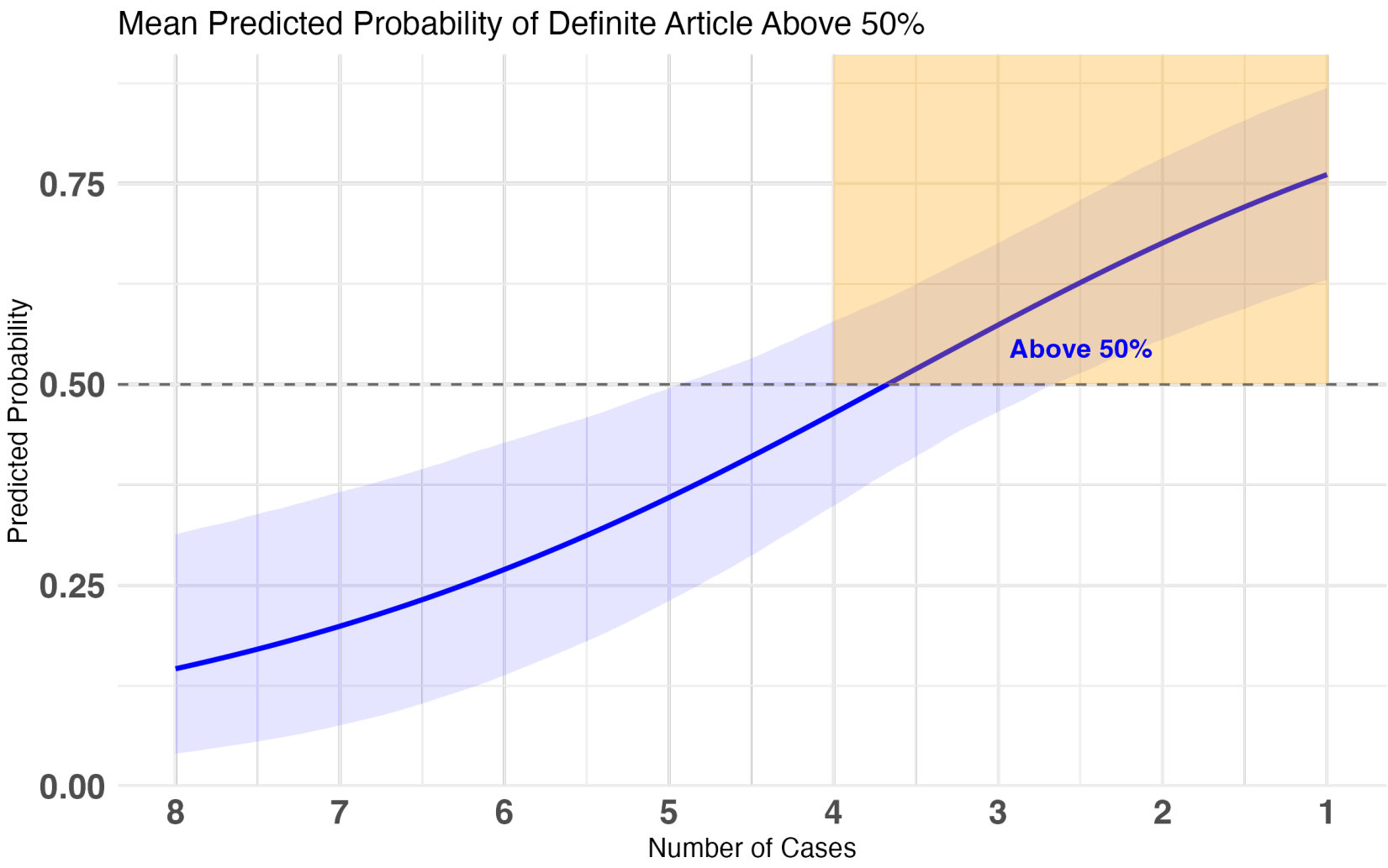


Results

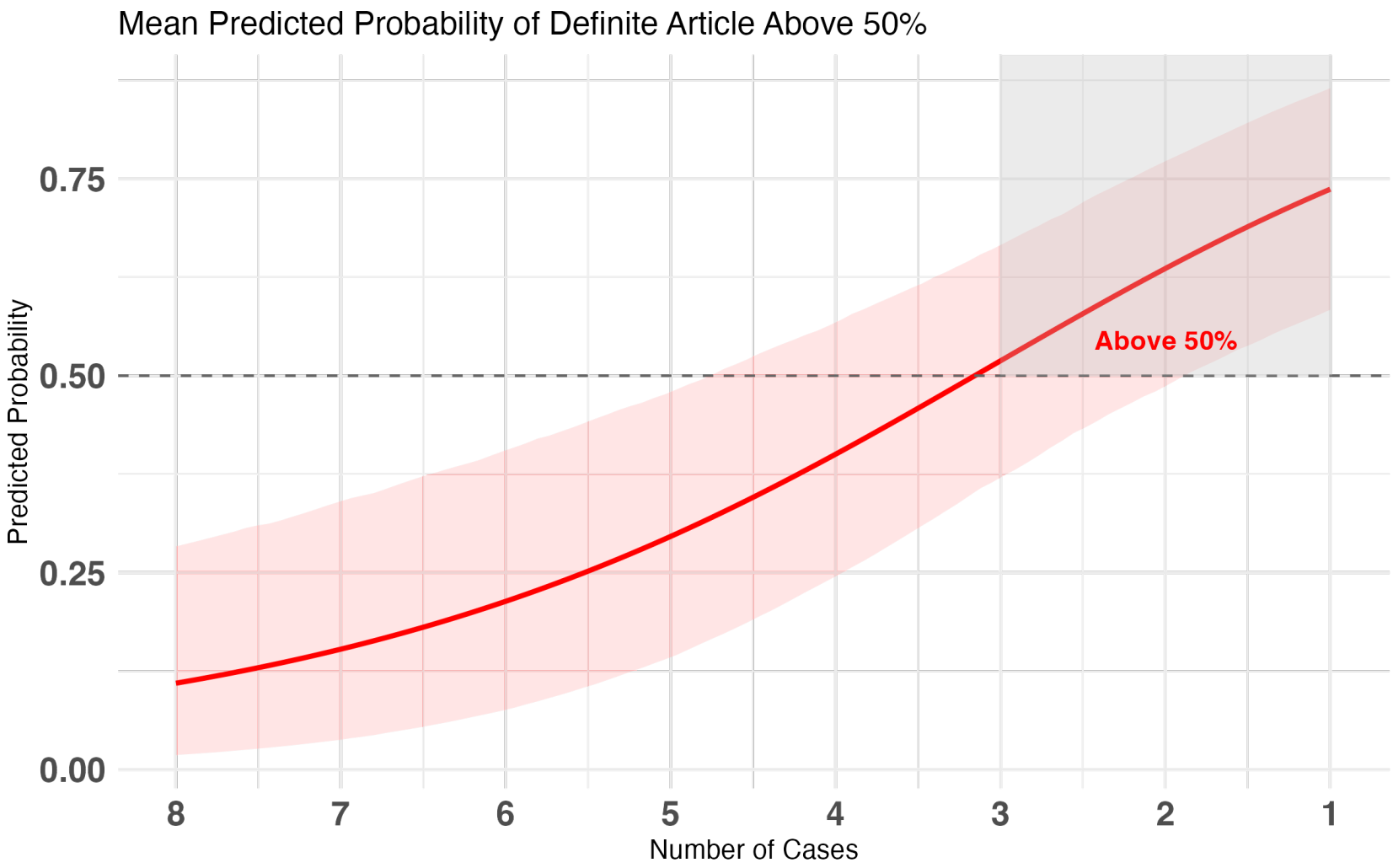
Posterior distributions of predictor variables



Posterior predictive distribution: Definite article



Posterior predictive distribution: Indefinite article



Discussion

Blake's case hierarchy (Blake 2001, p. 89)

- Nominative » Accusative/Ergative » Genitive » Dative » Locative » Ablative/Instrumental » Others

Blake's case hierarchy (Blake 2001, p. 89)

- Core: Nominative » Accusative/Ergative » Genitive » Dative

Blake's case hierarchy (Blake 2001, p. 89)

- Peripheral: Locative » Ablative/Instrumental » Others

Reframing the generalization

- Articles start to emerge among case inventories dedicated to only core relationships

Reframing the generalization

- Articles start to emerge among case inventories dedicated to only core relationships
- As the number of cases within the core class diminishes, the probability of articles increases

Is the relationship between case and articles causal?

A compensatory account

With the repurposing of the demonstrative as the article, Old High German developed a very simple means of addressing the collapse of case distinctions: the task that the case ending could no longer reliably fulfill was taken over by the demonstrative, which was thereby transformed into the definite article. The task assigned to it by the grammatical system has proven to be one that it has successfully managed to this day! (Tschirch 1975, p. 175, my trans.)

Theoretical accounts

- Various theoretical proposals try to make articles and case realization of the same underlying category or grammatical function (e.g., Giusti 1995; Martin et al. 2021)

Theoretical accounts

- Various theoretical proposals try to make articles and case realization of the same underlying category or grammatical function (e.g., Giusti 1995; Martin et al. 2021)
- These proposals reflect a compensatory account between articles and the *absence* of case

Theoretical accounts

- Various theoretical proposals try to make articles and case realization of the same underlying category or grammatical function (e.g., Giusti 1995; Martin et al. 2021)
- These proposals reflect a compensatory account between articles and the *absence* of case
- The relationship is **cumulative**, however: the fewer cases a language has, the more likely it is to have articles

Motivation for a causal account

- Accusative case can be used to signal definiteness/specificity (Enç 1991; Paul 1998, pp. 21–26)

Motivation for a causal account

- Accusative case can be used to signal definiteness/specificity (Enç 1991; Paul 1998, pp. 21–26)
- Definiteness marker > nominative > ergative (König 2011, p. 514)

Motivation for a causal account

- Accusative case can be used to signal definiteness/specificity (Enç 1991; Paul 1998, pp. 21–26)
- Definiteness marker > nominative > ergative (König 2011, p. 514)
- Genitive/partitive case can be used to signal indefiniteness (Philippi 1997; Kiparsky 1998)

Motivation for a causal account

- Accusative case can be used to signal definiteness/specificity (Enç 1991; Paul 1998, pp. 21–26)
- Definiteness marker > nominative > ergative (König 2011, p. 514)
- Genitive/partitive case can be used to signal indefiniteness (Philippi 1997; Kiparsky 1998)
- Case loss thought to impact other aspects of syntax (Dragomirescu and Nicolae 2016, p. 911)

The challenges of causal attribution

- Causation in language change not well understood (Winter-Froemel 2013)

The challenges of causal attribution

- Causation in language change not well understood (Winter-Froemel 2013)
- What do we mean by causal effect? (Precondition? Proxy?)

The challenges of causal attribution

- Causation in language change not well understood (Winter-Froemel 2013)
- What do we mean by causal effect? (Precondition? Proxy?)
- Inferring causation from observational data requires devoted statistical methods (e.g., Pearl 2009)

The challenges of causal attribution

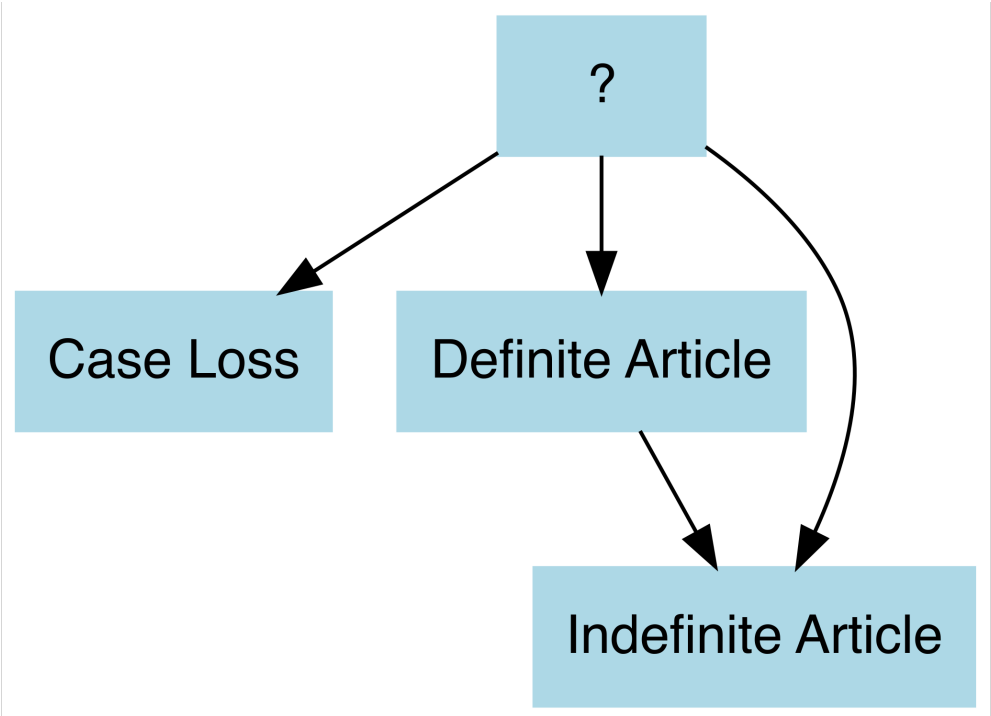
- Causation in language change not well understood (Winter-Froemel 2013)
- What do we mean by causal effect? (Precondition? Proxy?)
- Inferring causation from observational data requires devoted statistical methods (e.g., Pearl 2009)
- Since no study evaluates all of the causal factors that have been proposed (e.g., aspect, word order), they are all premature

The challenges of causal attribution

- Causation in language change not well understood (Winter-Froemel 2013)
- What do we mean by causal effect? (Precondition? Proxy?)
- Inferring causation from observational data requires devoted statistical methods (e.g., Pearl 2009)
- Since no study evaluates all of the causal factors that have been proposed (e.g., aspect, word order), they are all premature
- How to rule out other scenarios?

Indirect causal effect?

Unobserved confounder?



Final thoughts

Summary

- Evidence from Indo-European provides robust support for the statistical association between nominal case and articles

Summary

- Evidence from Indo-European provides robust support for the statistical association between nominal case and articles
- Articles become more likely as languages lose case markers, in particular after they've lost peripheral cases

Summary

- Evidence from Indo-European provides robust support for the statistical association between nominal case and articles
- Articles become more likely as languages lose case markers, in particular after they've lost peripheral cases
- The question of causation remains open as does the import of the association between case and articles for linguistic theory

Future work

- What would a causal model look like?

Future work

- What would a causal model look like?
- Confounders: What happens if we include other predictor variables (such as word order and aspect)?

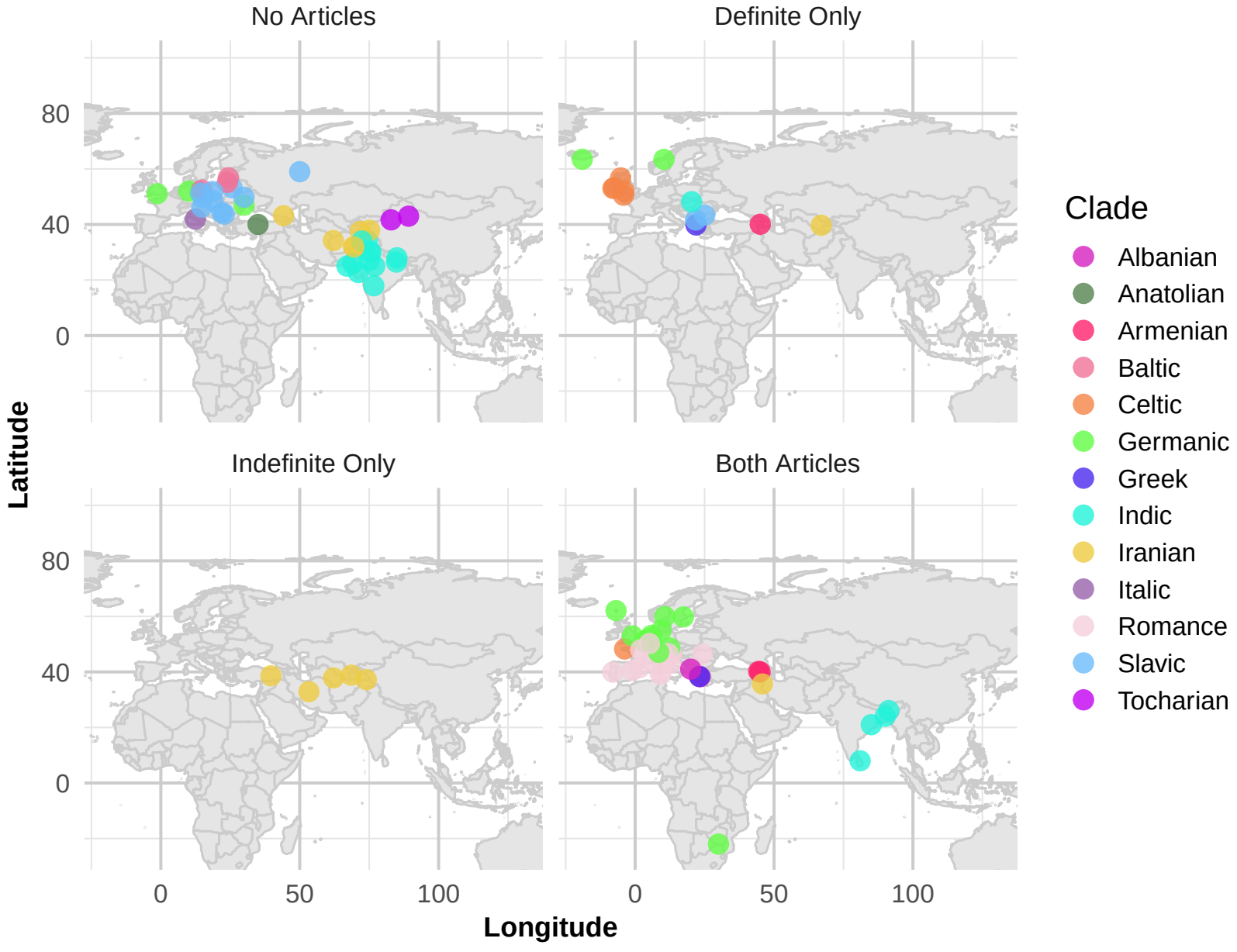
Future work

- What would a causal model look like?
- Confounders: What happens if we include other predictor variables (such as word order and aspect)?
- Is there evidence for the IE correlation elsewhere in the world's languages?

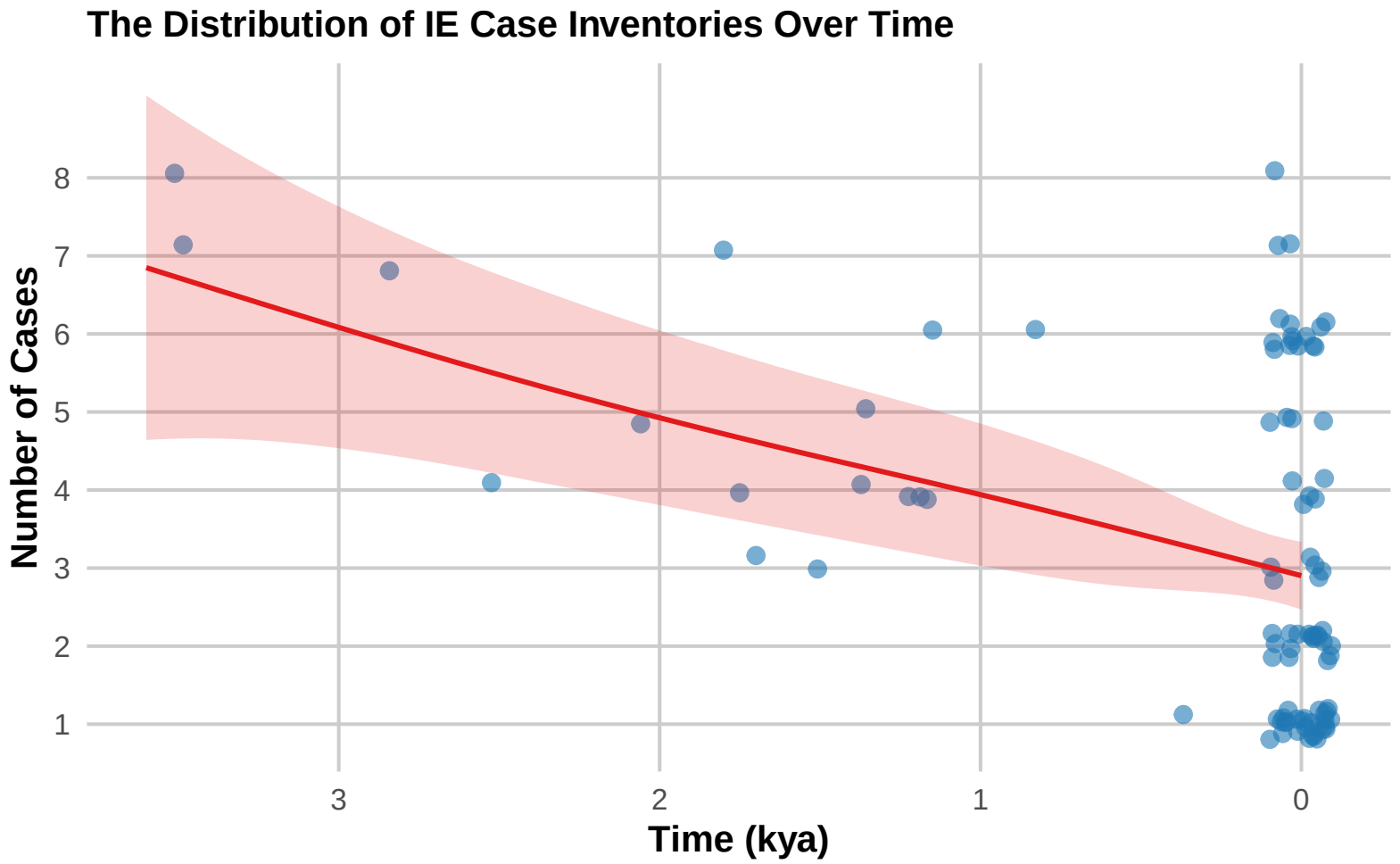
Vielen Dank!

Appendix

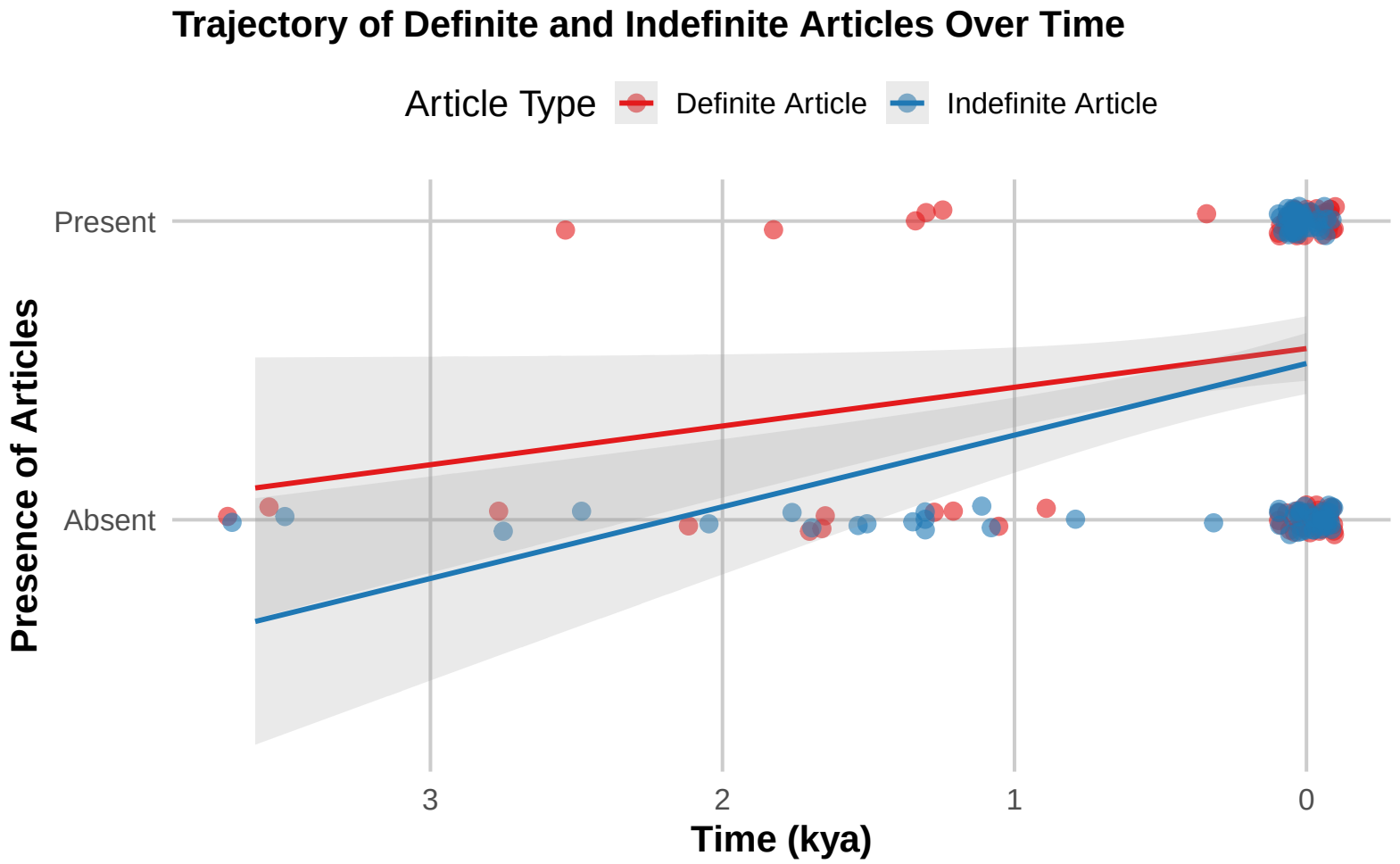
Distribution of article inventories by area and clade



Diachronic trajectory of case



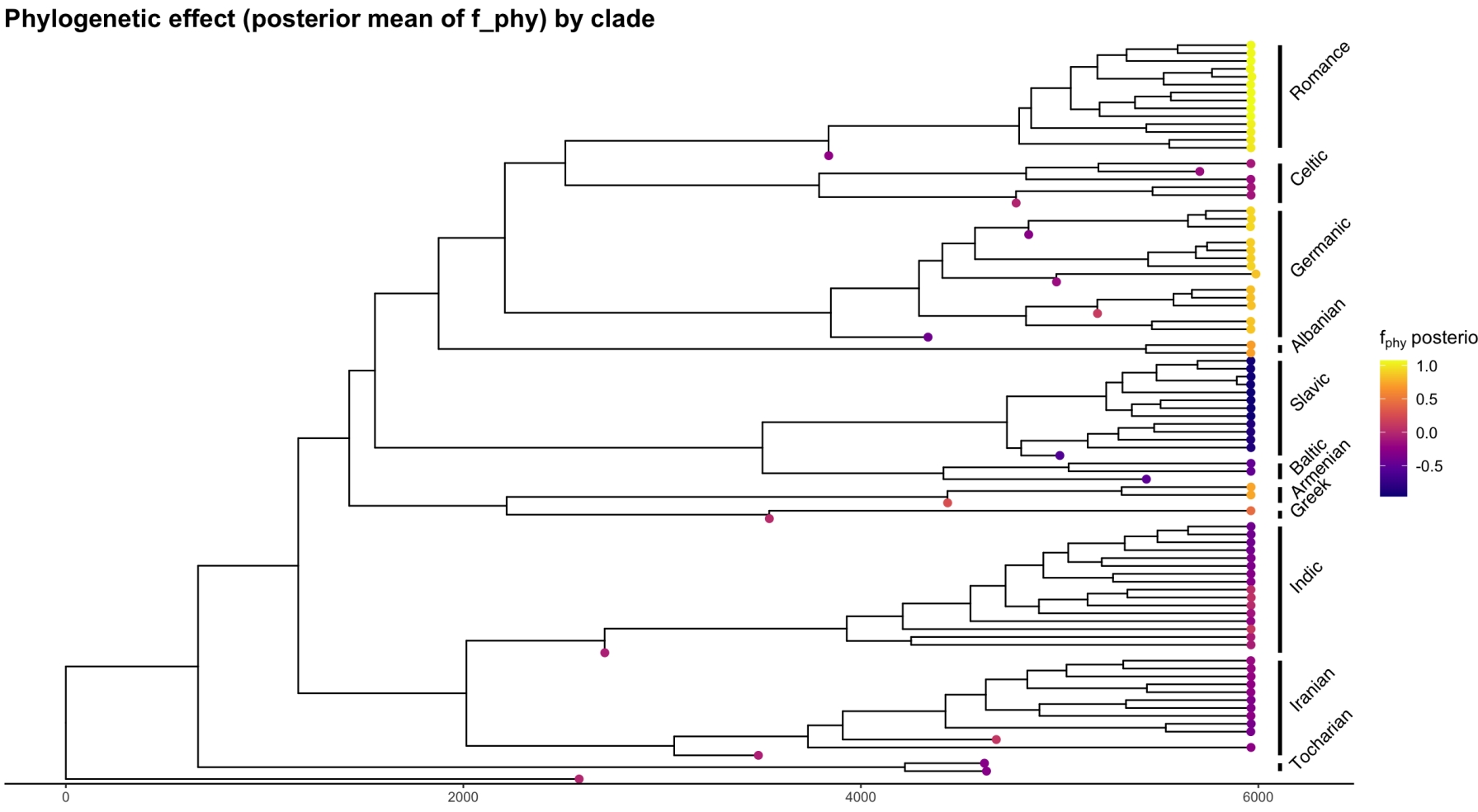
Diachronic trajectory of articles



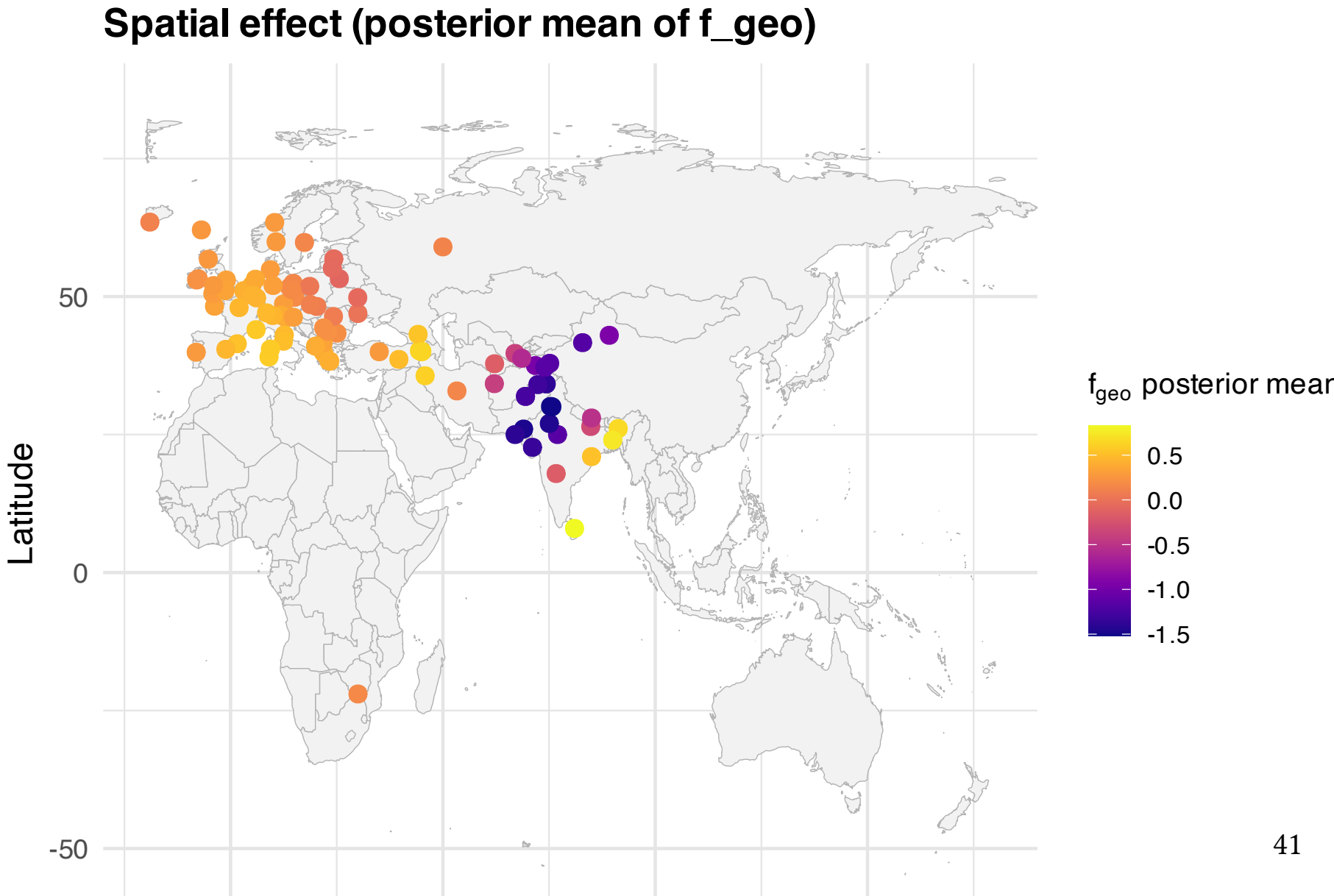
Predictive accuracy

Metric	Definite Article	Indefinite Article	Combined
Classification Accuracy	0.915	0.83	0.872
Area Under the Curve (AUC)	0.965	0.945	–
F1 Score	0.92	0.851	–
Brier Score	0.094	0.099	–

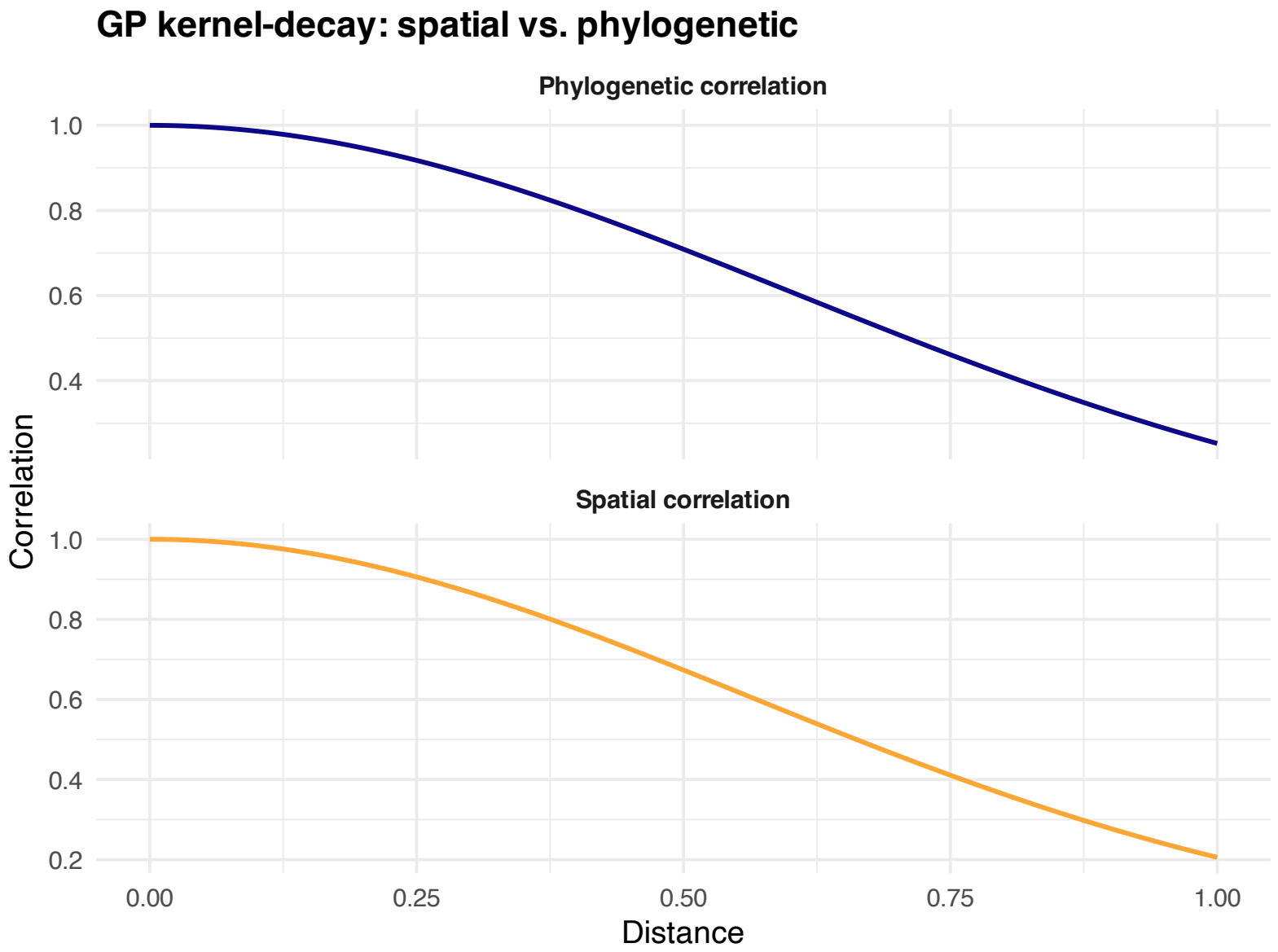
Phylogenetic autocorrelation



Spatial autocorrelation



Kernel decay



Paul, Hermann (1919). ***Deutsche Grammatik.*** German. Vol. 3. Halle an der Salle: Niemeyer.

Tschirch, Fritz (1975). ***Geschichte der deutschen Sprache.*** German. 2nd ed. Berlin: Schmidt.

Enç, Mürvet (1991). **"The semantics of specificity".** In: *Linguistic Inquiry* 22.1, pp. 1–25.

Giusti, Giuliana (1995). **"A unified structural representation of (abstract) case and article".** In: *Studies in comparative Germanic syntax*. Ed. by Hubert Haider, Susan Olsen, and Sten Vikner. Dordrecht: Kluwer, pp. 77–93.

Posner, Rebecca R. (1996). ***The Romance languages.*** Cambridge: Cambridge University Press.

Philippi, Julia (1997). **“The rise of the article in the Germanic languages”.** In: *Parameters of morphosyntactic change*. Ed. by Ans van Kemenade and Nigel Vincent. Cambridge: Cambridge University Press, pp. 62–93.

Kiparsky, Paul (1998). **“Partitive case and aspect. Lexical and compositional factors”.** In: *The projection of arguments*. Ed. by Miriam Butt and Wilhelm Geuder. Stanford: Center for the Study of Language and Information, pp. 265–307.

Paul, Ludwig (1998). ***Zazaki. Grammatik und Versuch einer Dialektologie.*** German. Reichert.

Blake, Barry J. (2001). ***Case.*** 2nd ed. Cambridge: Cambridge University Press.

Wood, Johanna L. (2003). **“Definiteness and number. Determiner phrase and number phrase in the history of English”.** PhD thesis. Arizona State University.

Hewson, John and Vít Bubeník (2006). ***From case to adposition. The development of configurational syntax in Indo-European languages.*** Amsterdam: John Benjamins.

Barðdal, Jóhanna (2009). **"The development of case in Germanic".** In: *The role of semantic, pragmatic, and discourse factors in the development of case*. Ed. by Jóhanna Barðdal and Shobhana L. Chelliah. Amsterdam: John Benjamins, pp. 123–159.

Evans, Nicholas D. and Stephen C. Levinson (Oct. 2009). **"The myth of language universals. Language diversity and its importance for cognitive science".** In: *Behavioral and Brain Sciences* 32.5, pp. 429–492.

Pearl, Judea (2009). ***Causality. Models, reasoning and inference.*** 2nd ed. Cambridge: Cambridge University Press.

König, Christa (2011). **“The grammaticalization of adpositions and case marking”.** In: *The Oxford handbook of grammaticalization.* Ed. by Bernd Heine and Heiko Narrog. Oxford: Oxford University Press, pp. 511–521.

Ledgeway, Adam (2011). **“Grammaticalization from Latin to Romance”.** In: *The Oxford handbook of grammaticalization.* Ed. by Bernd Heine and Heiko Narrog. Oxford: Oxford University Press, pp. 719–728.

Bailyn, John F. (2012). ***The syntax of Russian.*** Cambridge: Cambridge University Press.

Kim, Ronald I. (2012). **“The Indo-European, Anatolian, and Tocharian “secondary” cases in typological perspective”.** In: *Multi nominis grammaticus. Festschrift for Alan J. Nussbaum.* Ed. by Adam I. Cooper, Jeremy Rau, and Michael Weiss. Ann Arbor: Beech Stave Press, pp. 121–142.

Kulikov, Leonid I. (2012). **"The Proto-Indo-European case system and its reflexes in a diachronic typological perspective. Evidence for the linguistic prehistory of Eurasia".** In: *Rivista degli Studi Orientali* 84, pp. 289–309.

Winter-Froemel, Esme (2013). **"What does it mean to *explain* language change? Usage-based perspectives on causal and intentional approaches to linguistic diachrony, or. On *S*-curves, invisible hands, and speaker creativity".** In: *Energeia. Online Journal for Linguistics, Language Philosophy and History of Linguistics* 5, pp. 123–142.

Ladd, D. Robert, Seán G. Roberts, and Dan Dediu (2015). **"Correlational studies in typological and historical linguistics".** In: *Annual Review of Linguistics* 1, pp. 221–241.

Börjars, Kersti, Pauline Harries, and Nigel Vincent (Mar. 2016). **"Growing syntax. The development of a DP in North Germanic".** In: *Language* 92.1, e1–e37.

Dragomirescu, Adina and Alexandru Nicolae (2016). **"Case".** In: *The Oxford guide to the Romance languages*. Ed. by Adam Ledgeway and Martin Maiden. Oxford: Oxford University Press, pp. 911–923.

Guzmán Naranjo, Matías and Laura Becker (2021). **"Statistical bias control in typology".** In: *Linguistic Typology* 26.3, pp. 605–670.

Martin, Txuss, Ioanna Sitaridou, and Wolfram Hinzen (Dec. 2021). **"Correlations between Case and the D-system and the interpretability of Case".** In: *Borealis. An International Journal of Hispanic Linguistics* 10.2, pp. 238–263.

Goldstein, David M. (2022). **"Correlated grammaticalization. The emergence of articles in Indo-European".** In: *Diachronica* 39.5, pp. 658–706.

Guzmán Naranjo, Matías and Miri Mertner (Dec. 2022). **“Estimating areal effects in typology. A case study of African phoneme inventories”.** In: *Linguistic Typology* 27.2, pp. 455–480.