What can a corpus of YouTube videos tell us about grammatical variation in North American English?

Joseph A. Stanley
Brigham Young University

Brett Hashimoto
Brigham Young University

Jack Grieve
University of Birmingham

American Dialect Society 2026 Annual Meeting
January 11, 2026
New Orleans, Louisianna

Background and Motivation

- In the 1950s, dialectologists relied on a mix of lexical, grammatical, phonological, and phonetic features.
  - Nowadays, dialect mapping primarily based on phonetic features (e.g. Labov, Ash & Boberg 2006)

- Some (but not much) corpus-based dialectology (Grieve 2016)
  - Some grammatical/lexical research, but the minority and mostly elicited rather than naturalistic (e.g. Kurath 1939, Carver 1987, Leemann et al 2018, Leemann et al 2020)

- Little that maps large areas or focuses on multiple features simultaneously (though see Kim et al 2019 and Stanley 2022)

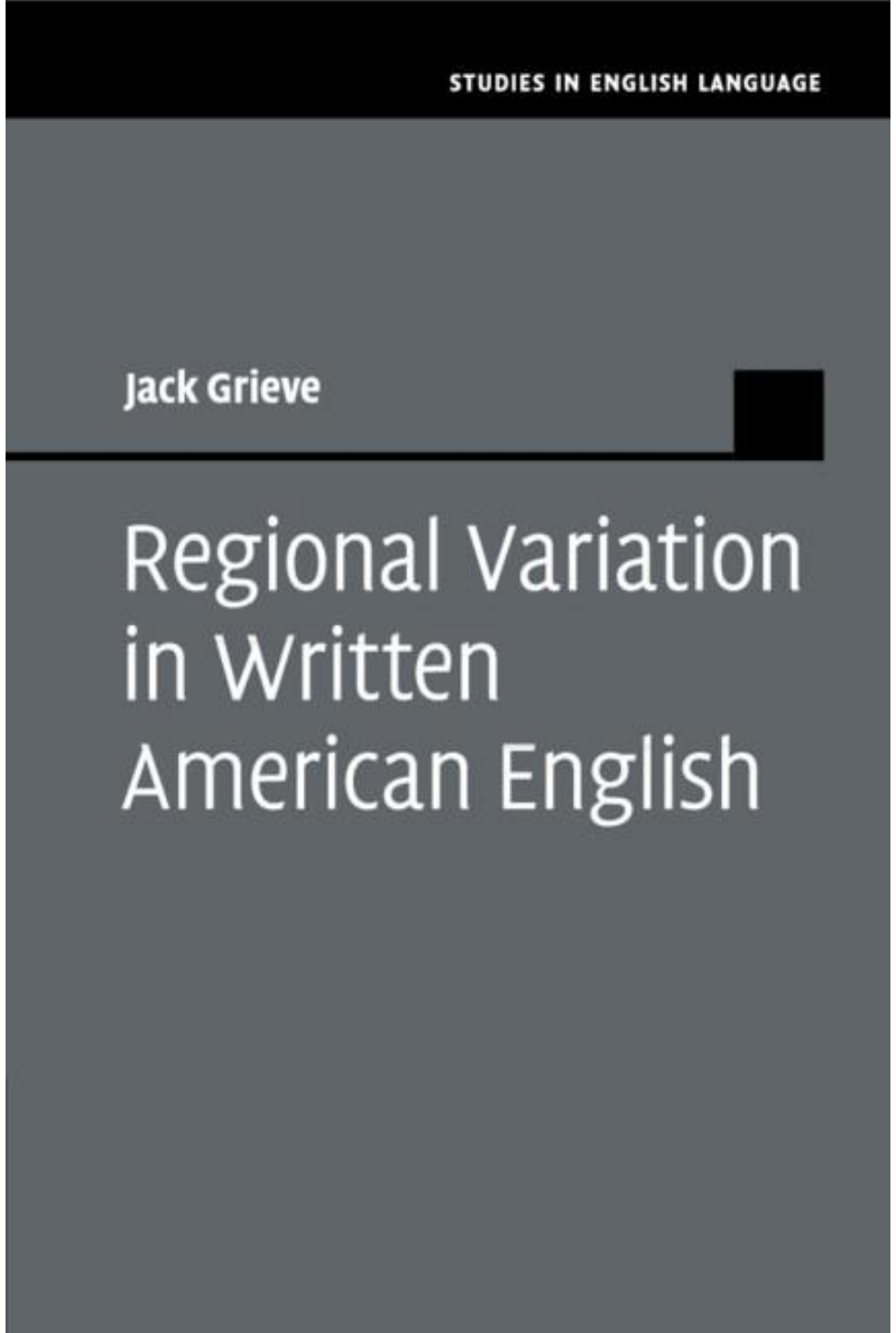

STUDIES IN ENGLISH LANGUAGE
Jack Grieve
Regional Variation in Written American English

Table 3.1 *The complete set of 135 grammatical alternations and their 295 variants*

Pronouns (7, 14)
Anyone/Anybody
Everyone/Everybody
Someone/Somebody
No one/Nobody
Each other/One another
Let's/Let us
'Em/Them

Relative Pronouns (4, 11)
Who/That Human Subject
Zero/*That/Whom/Who* Human Object
That/Which Non-human Subject
Zero/*That/Which* Non-human Object

Determiners (8, 21)
Fewer/Less
Most/Almost all
With no/Without
All/All of
Half/Half of
One/A Number
Many/A lot of/Lots of/Plenty of
Much/A lot of/Lots of/Plenty of/A great deal of

Adjectives (8, 16)
Previous/Prior
Past/Last Numeral
Only/Sole
Un-/Not Negation
In-/Not Negation
Adjective *-est/Most* Adjective
At all/Whatsoever
Attributive/Predicative Adjective

Nouns (5, 11)
Kind of/Sort of/Type of
Common Noun Genitive *Of's*
Proper Noun Genitive *Of's*
Pre-/Post-nominal Modification
Nouns/Pronoun

Prepositions (22, 48)
Among/Amongst
Toward/Towards
Outside/Outside of
Off/Off of
In spite of/Despite
Because of/Due to
Instead of/Rather than
As a result of/Resulting from
Regardless of/No matter
With the exception of/Except for/Except
Get In/Into
Depend+etc. *On/Upon*
Compare+etc. *To/With*
Concerned+etc. *About/With/By*
Hear+etc. *Of/About*
Complaint+etc. *Of/About*
Concern+*Concerns About/With*
Article+etc. *About/On/Regarding/Concerning*
Different from/than
About/Around Number
More than/Over Number
Preposition Stranding/Fronting

Particles (6, 12)
Verb *Away* NP/Verb NP *Away*
Verb *Back* NP/Verb NP *Back*
Verb *Down* NP/Verb NP *Down*
Verb *Off* NP/Verb NP *Off*
Verb *Out* NP/Verb NP *Out*
Verb *Up* NP/Verb NP *Up*

Subordinators (7, 15)
Until/Till
Because/Since
As long as/So long as
As if/As though
Although/Though/Even Though
If/Whether
*If . . . then/*Zero

Coordinators (5, 10)
But/Yet
As well as/In addition to
Neither . . . nor/or
Not only . . . but/but also
Sentence Initial *And/*No *And*

Verbs (7, 14)
Be Contraction/Full
Pronoun *Have* Contraction/Full
Modal *Have* Contraction/Full
Strong/Weak Past Tense
Strong/Weak Perfect Aspect
Conditional *Were/Was*
By-Passive/Non-*By*-Passive

Table 3.1 *(cont.)*

Modals (7, 16)
Should/Ought
Will/Shall/Be going to
May/Might
Can/Could Question
Must/Need to/Have to
Will Contraction/Full
Would Contraction/Full

Infinitives (4, 8)
To Contraction/Full
In order to/So as to
Infinitive/Present Participle
Seem/Seem to be

Not (6, 12)
Be Not Contraction/Full
Do Not Contraction/Full
Have Not Contraction/Full
Modal Not Contraction/Full
Ain't/Standard Negation
Not/No Negation

Adverbs (13, 27)
-ward/-wards
-where/-place
Very/Really
Especially/Particularly Adj.+Adv.
Especially/Particularly Prep.+Subord.
Especially/Particularly NP
Totally/Completely/Entirely
Almost/Nearly Adjective
Almost/Nearly NP
Almost/Nearly Numeral
Only/Just Numeral
Much/Far Comparative Adjective
Much more/Far more Adjective

Adverbials (26, 60)
Previously/Formerly
Frequently/Often
Occasionally/Sometimes
Rarely/Seldom
Repeatedly/Again and again
Immediately/Right away
Suddenly/All of a sudden
Meanwhile/In the meantime
Simultaneously/At the same time
Currently/Presently/Right now
Usually/Normally/Most of the time
Clearly/Obviously
Maybe/Perhaps
Probably/Likely
Not Only/Not Just
However/Nevertheless/Nonetheless
Therefore/Thus
For example/For instance
Actually/In fact
Additionally/In addition
Furthermore/Further/Moreover
Ordinal-o/*-ly*
Finally/Lastly/Last
Stance Advl.: Initial/Internal/Final
Temporal Advl.: Initial/Internal/Final
Linking Advl.: Initial/Internal/Final

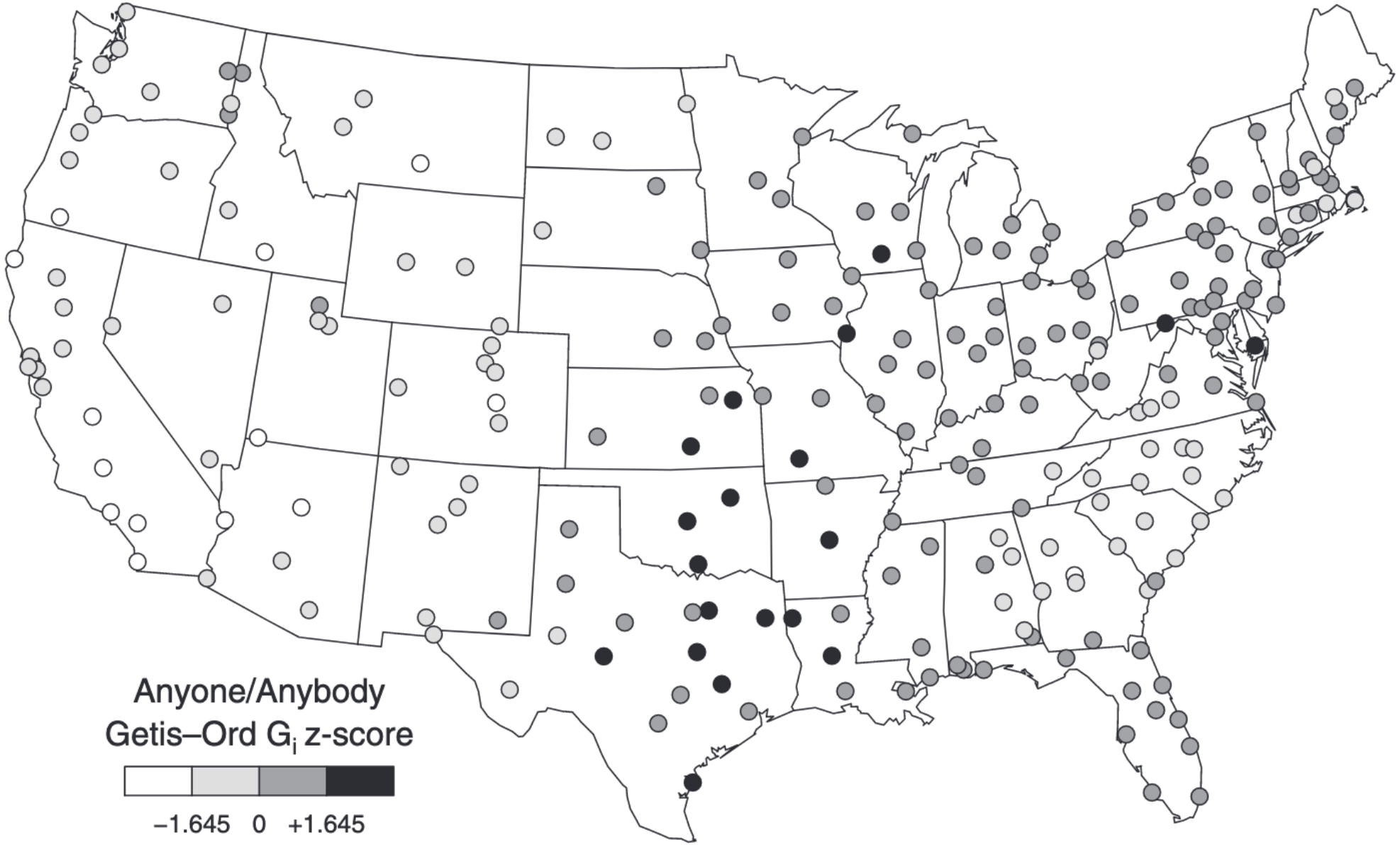


Figure 4.3 *Anyone/Anybody* local spatial autocorrelation map

Figure 4.4 *Anyone* and *Anybody* simplified local spatial autocorrelation map

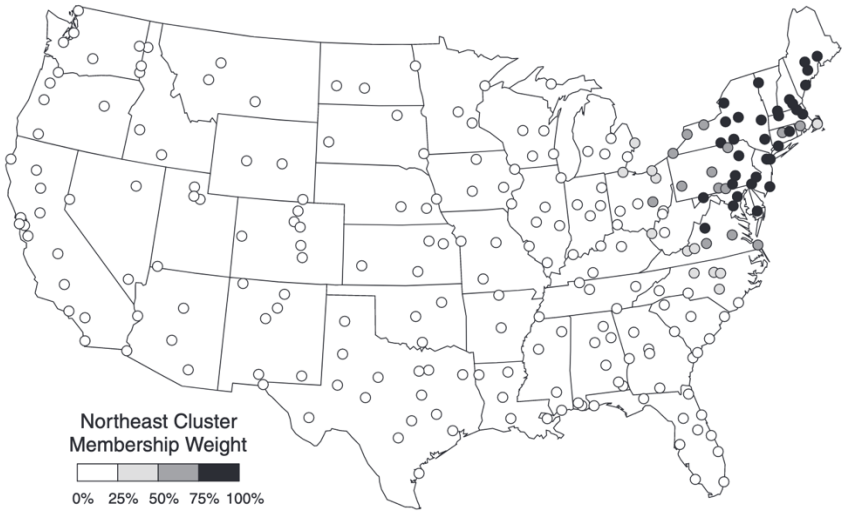


Figure 5.12 Northeastern dialect region cluster

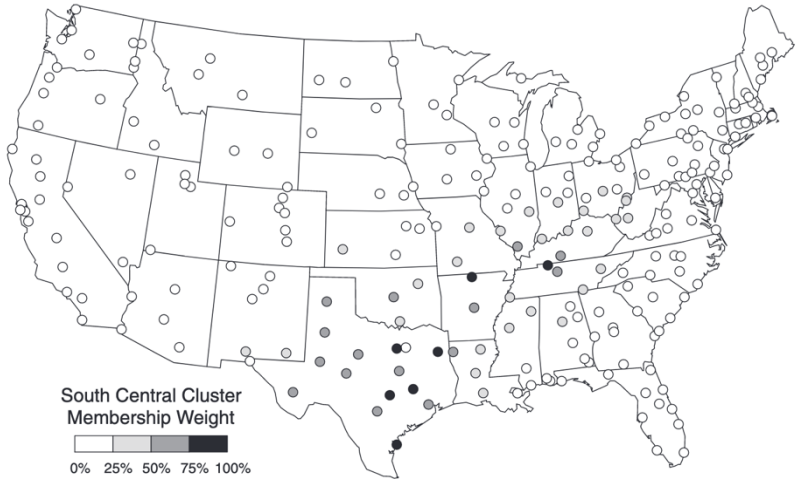


Figure 5.15 South Central dialect region cluster

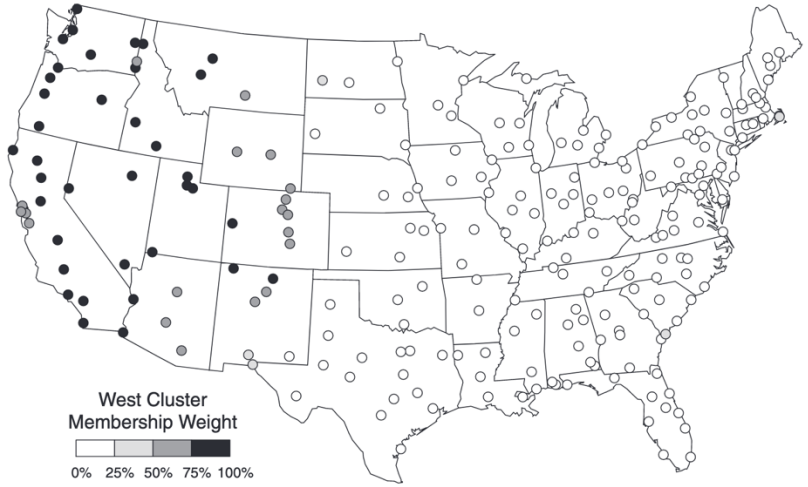


Figure 5.16 Western dialect region fuzzy cluster

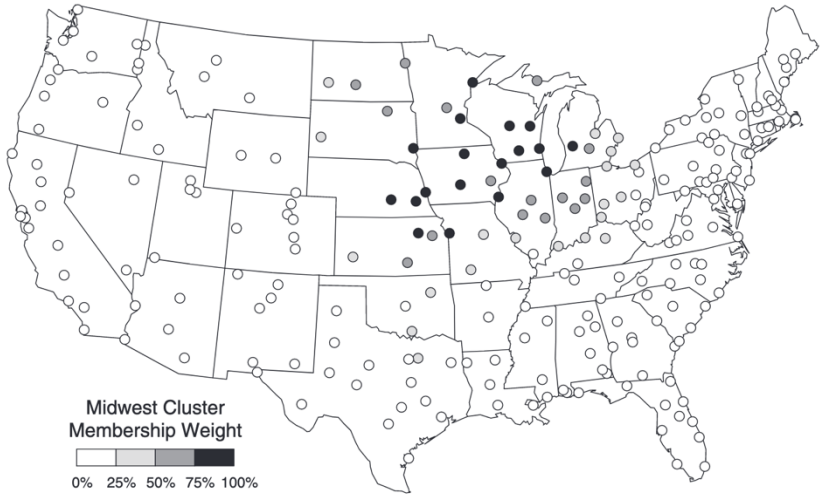


Figure 5.13 Midwestern dialect region cluster

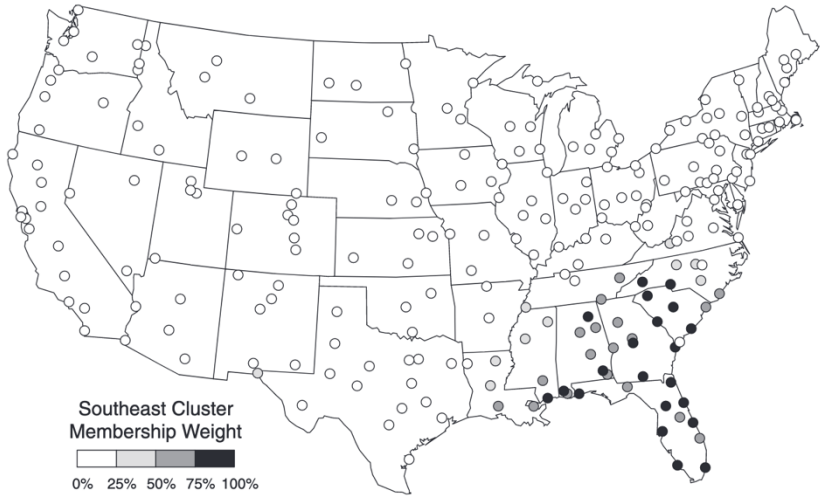


Figure 5.14 Southeastern dialect region cluster

Our Goal

Generate maps of the distributions of those
100+ lexico-grammatical feature alternations
in spoken North American English.

Corpus

- Corpus of North American Spoken English (CoNASE: Coats, 2019; 2023)
  - YouTube channels of mainly regional and local government entities or other governmental/civic organizations.
  - Stratified sampling from counties across the US and Canada
  - 301,847 texts; 154,041 hours of spoken language; 1,252,066,371 words
  - Autotranscribed and geotagged.
  - Stanza lemmatized; Part-of-speech tagged
- Same 135 grammatical alternation variables as Grieve (2016)
  - But, algorithms for feature identification were altered from Grieve (2016) to be more suitable.
  - Accuracy checking of features

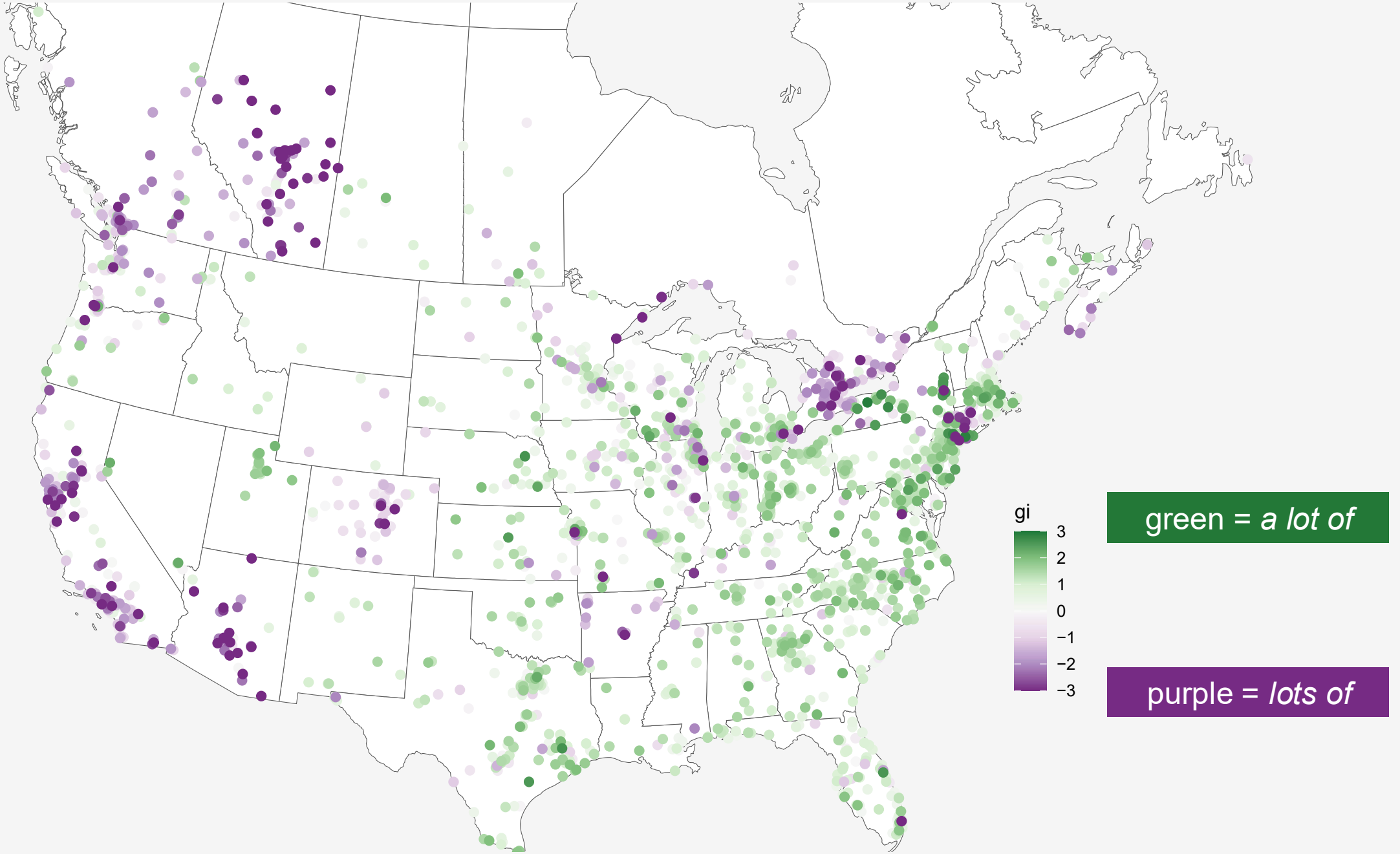
Quantitative analysis

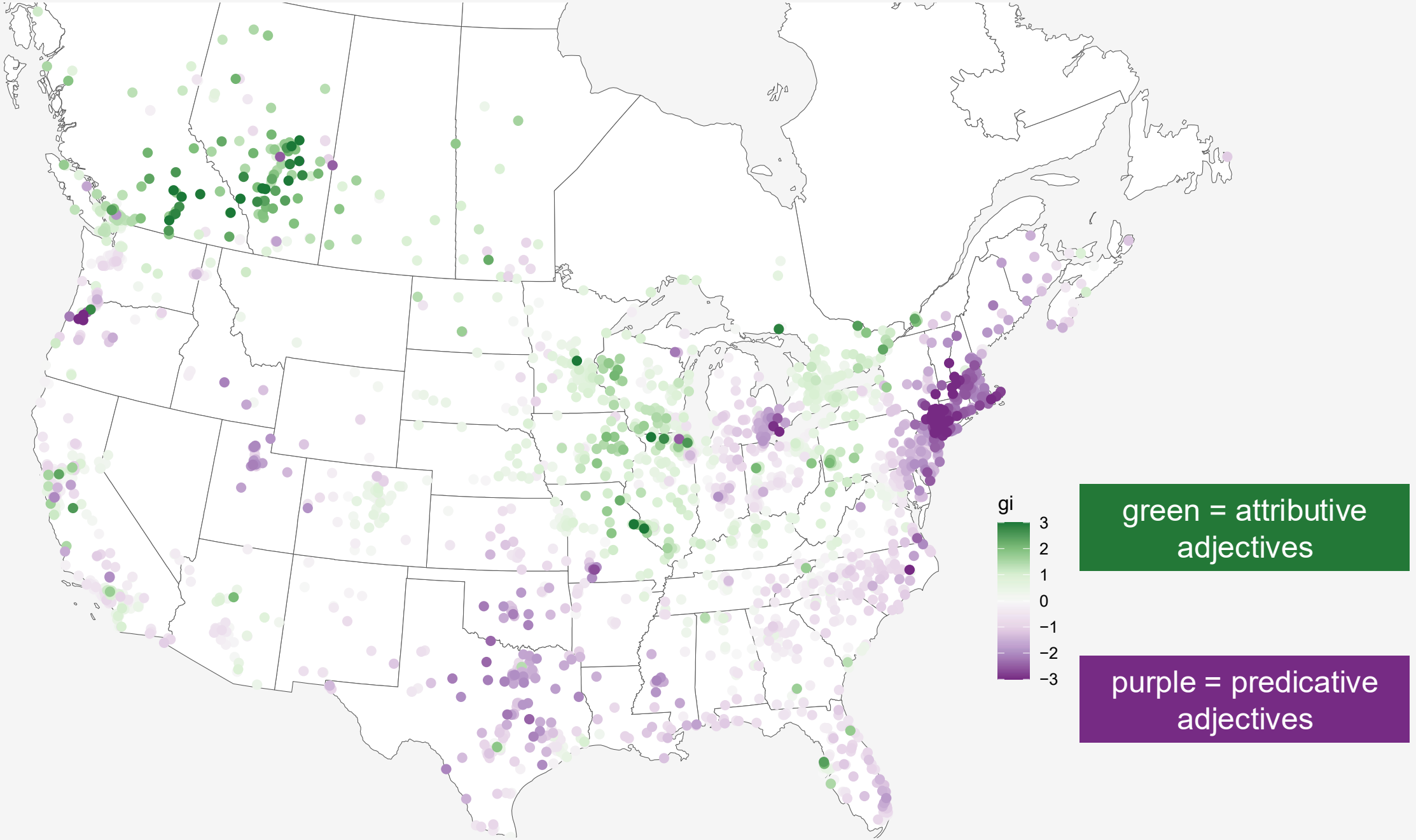
- Proportions by location
  - We calculated the proportion of each variant for each feature, i.e. A/(A+B).
  - Weighted average per location (301K texts → 2,537 locations)

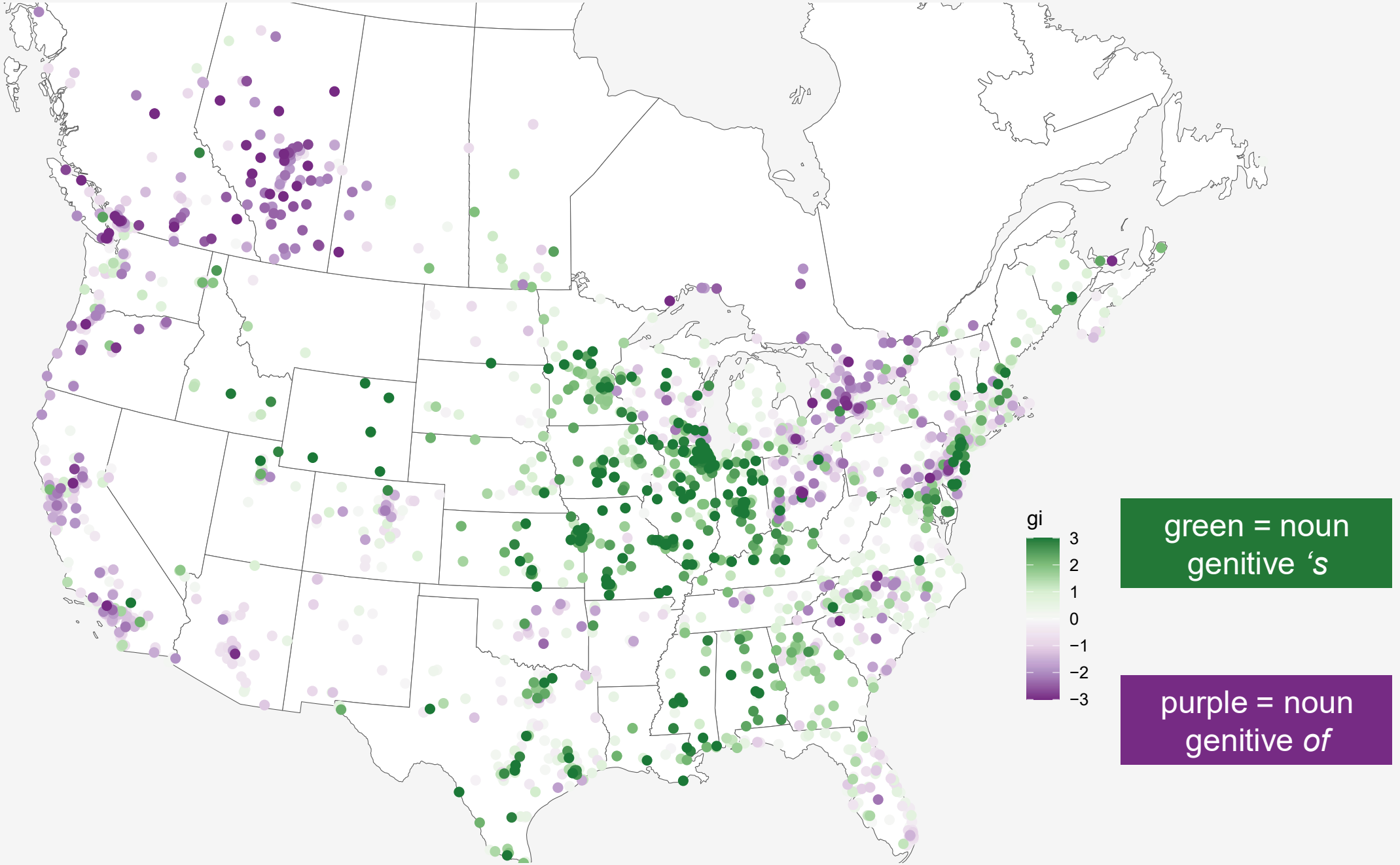
- Spatial stats following Grieve (2016)
  - Getis Ord-Gi statistic: For each location, indicates whether there is high/low clustering at that location (without regard to political boundaries)
  - Interpret this like a z-score, so high absolute values = statistically significant.

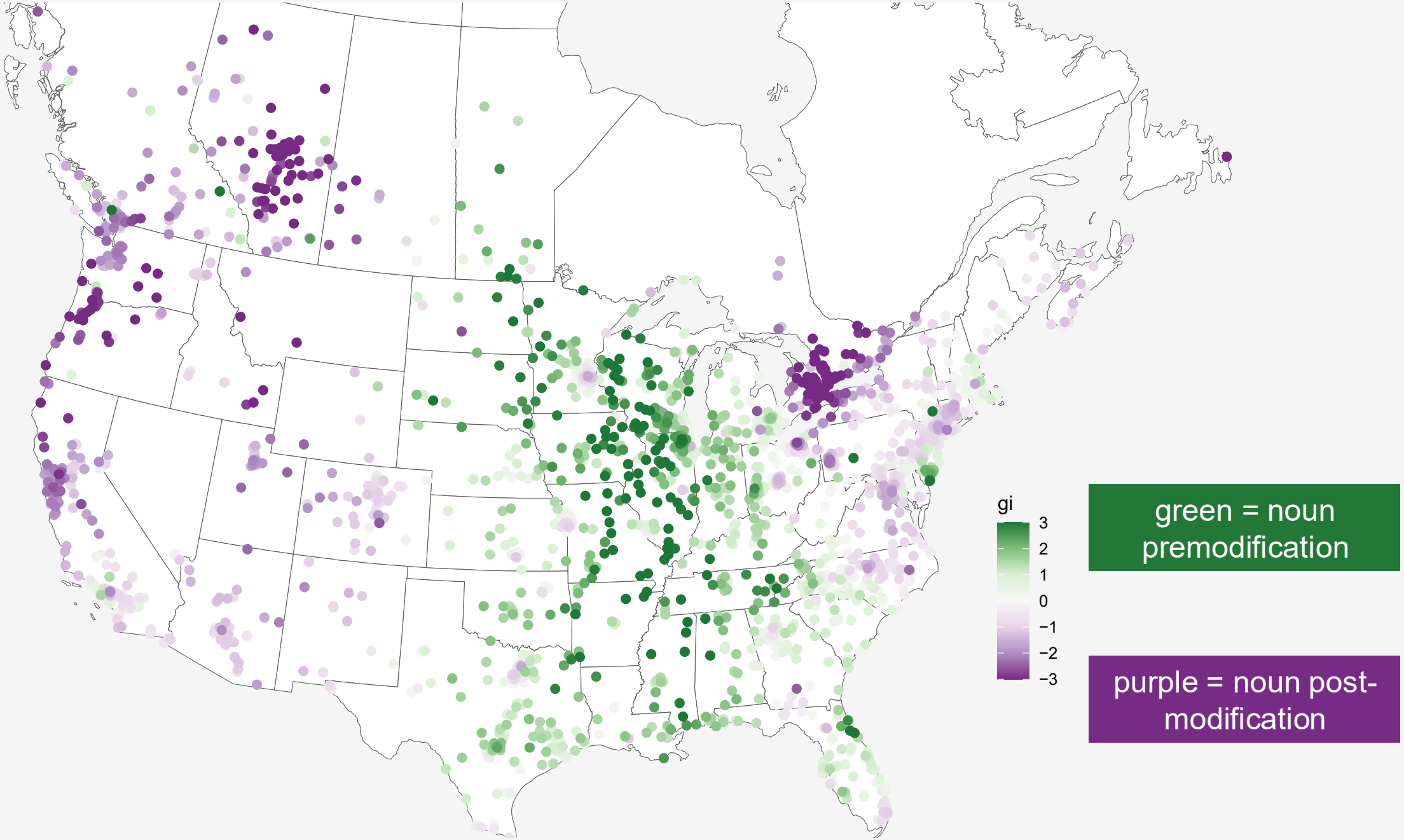
- Maps
  - Plot points (if there's enough data).
  - One variant is green; the other is purple.

Results


gi
3
2
1
0
−1
−2
−3
green = *a lot of*
purple = *lots of*

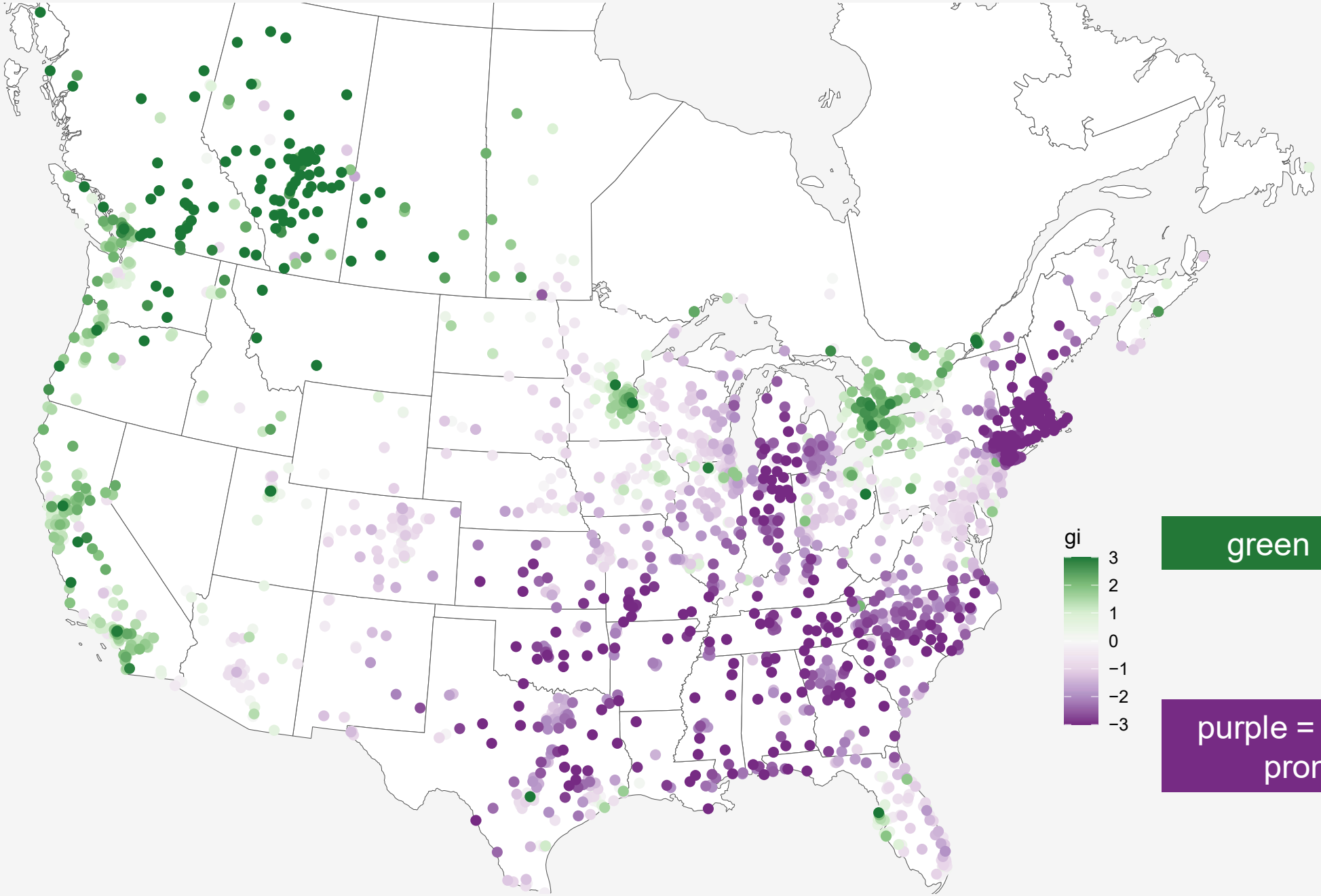

gi
3
2
1
0
−1
−2
−3
green = attributive adjectives
purple = predicative adjectives


gi
3
2
1
0
−1
−2
−3
green = noun genitive ‘s
purple = noun genitive *of*



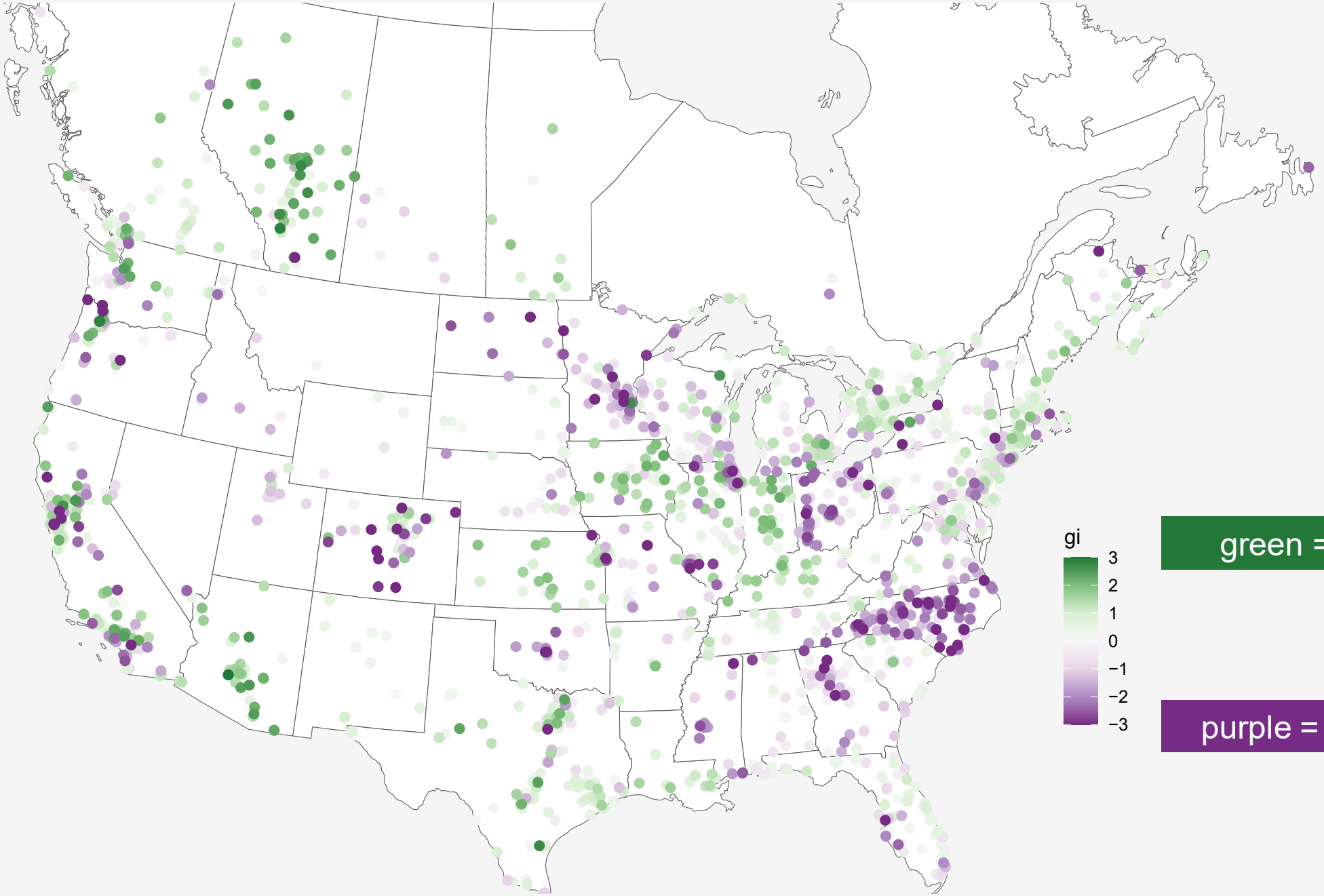
green = noun premodification

purple = noun post-modification



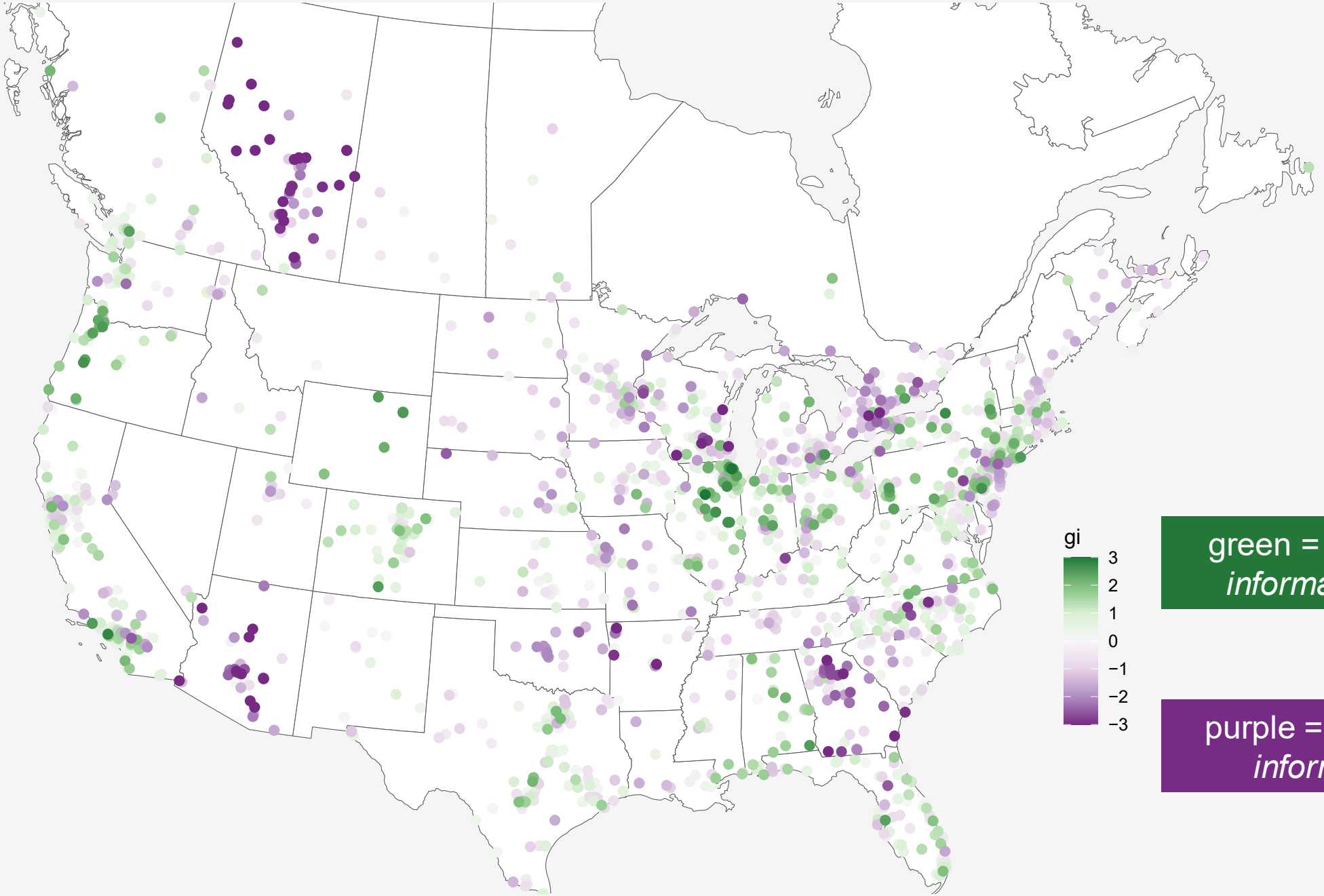
green = noun

purple = personal pronoun



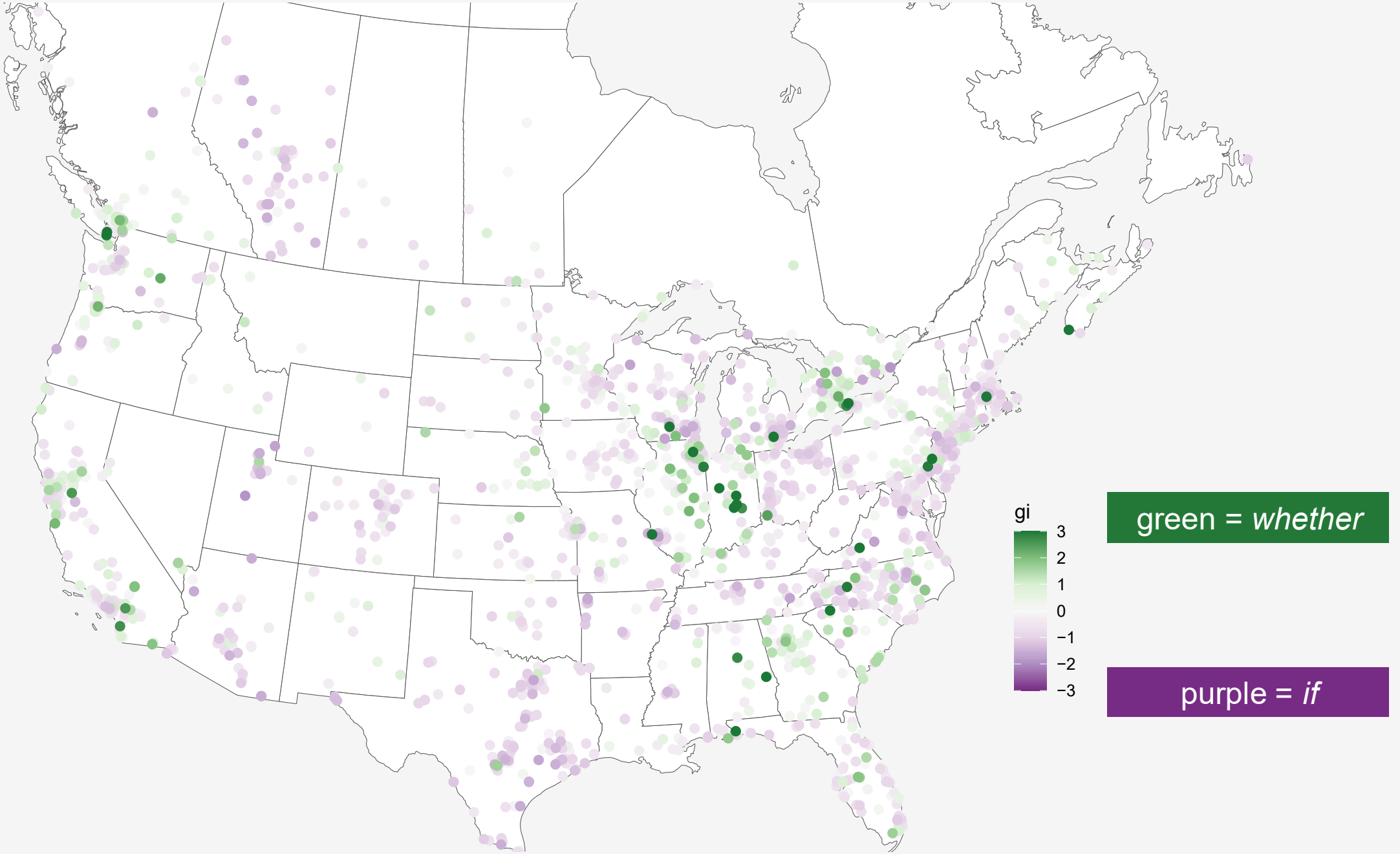
green = *compare to*

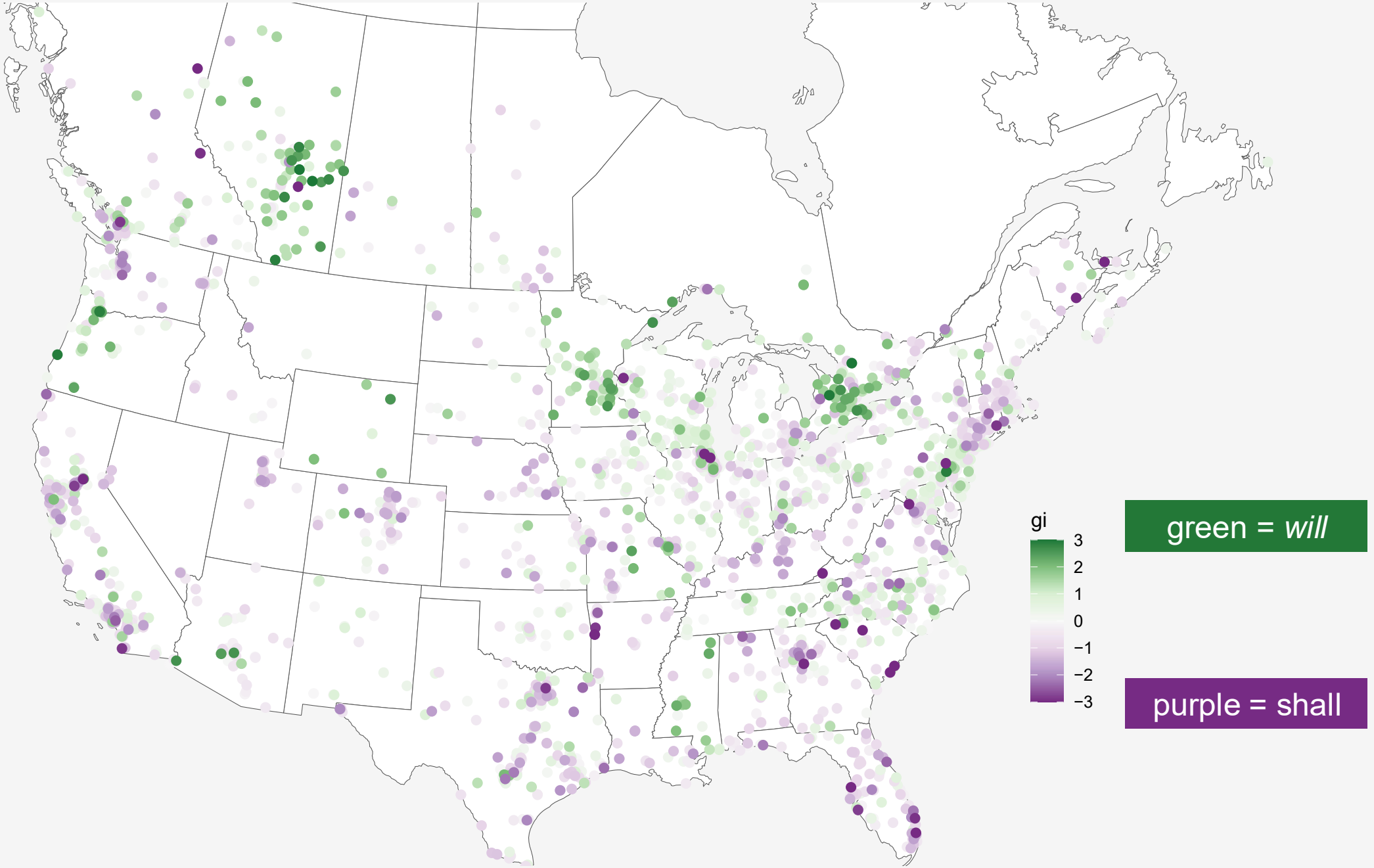
purple = *compare with*

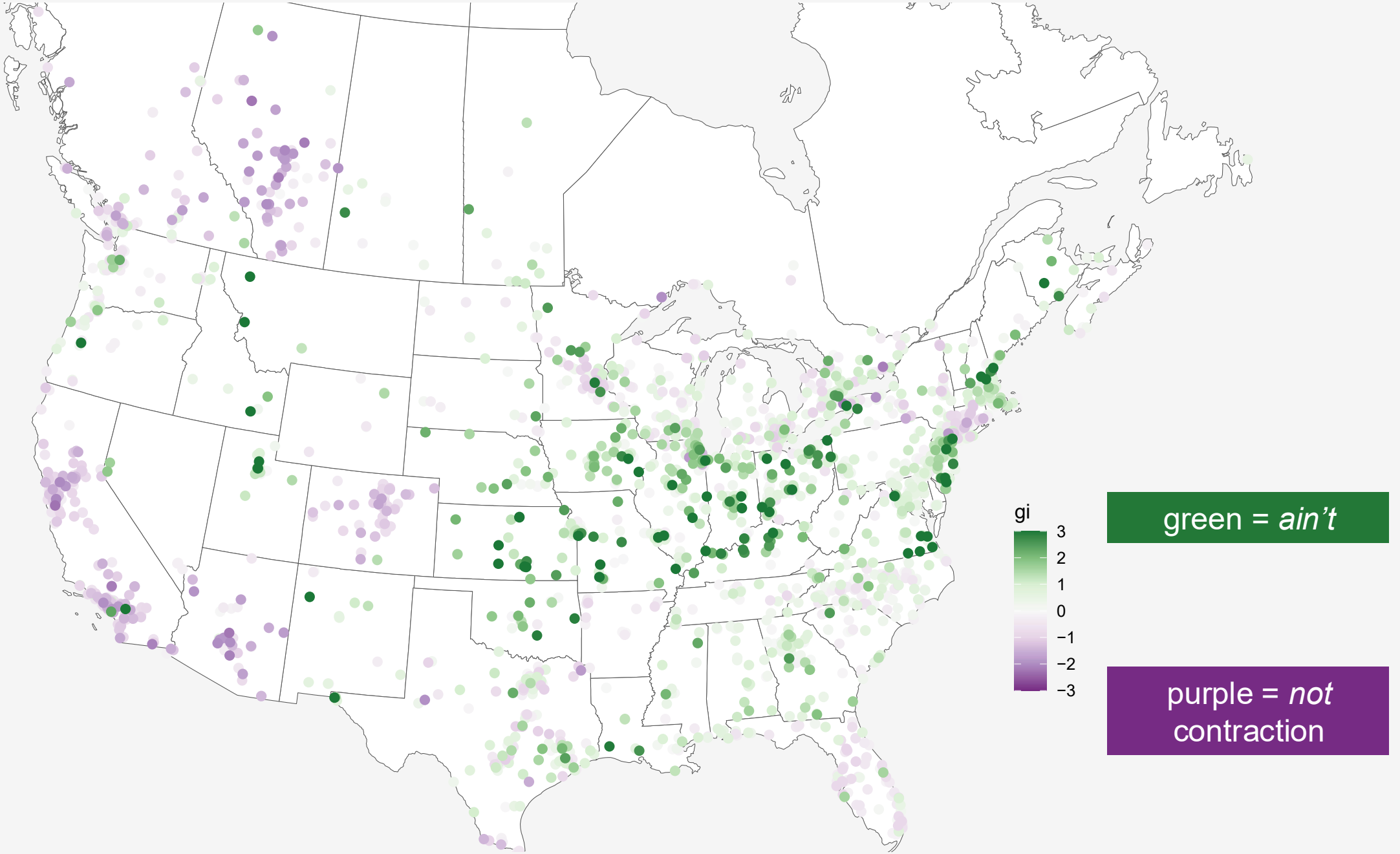


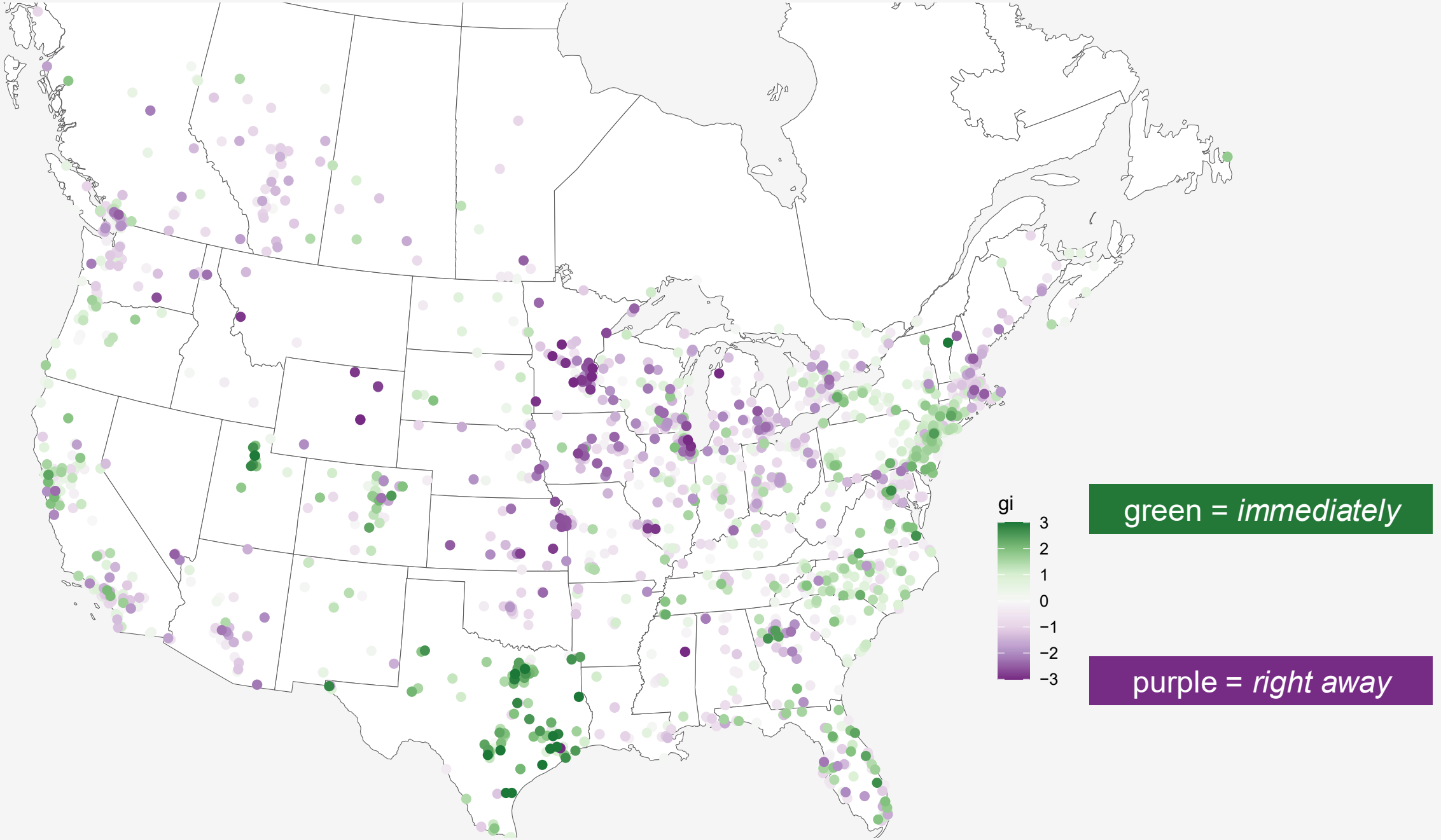
green = *article/book/ information about*

purple = *article/book/ information on*


gi
3
2
1
0
−1
−2
−3
green = *whether*
purple = *if*


gi
3
2
1
0
−1
−2
−3
green = *will*
purple = shall


gi
3
2
1
0
−1
−2
−3
green = *ain't*
purple = *not* contraction

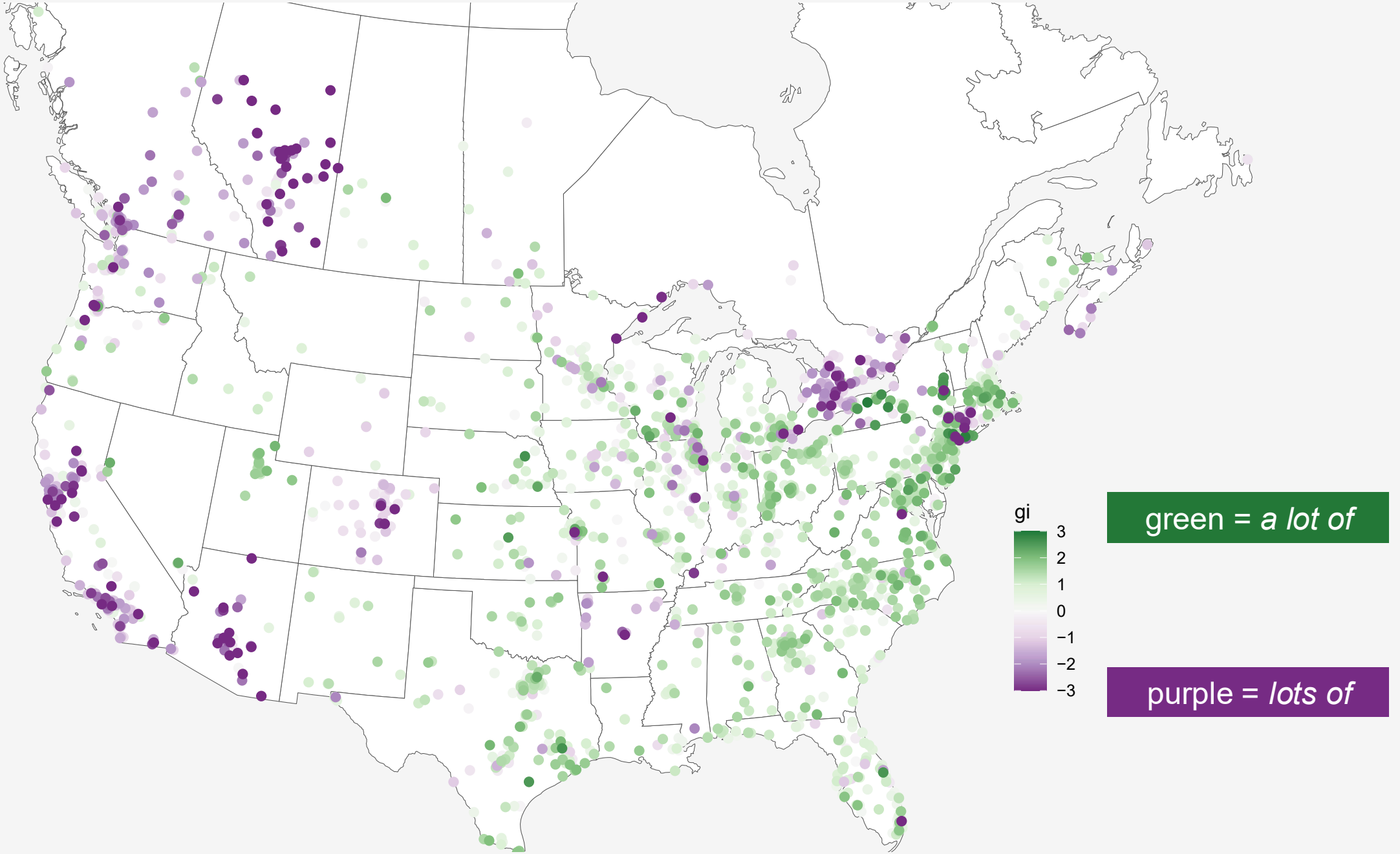

gi
3
2
1
0
−1
−2
−3
green = *immediately*
purple = *right away*

Discussion

Discussion

- Summary
  - Many features align nicely with known dialectal isoglosses
  - Many reveal interesting new geographic patterns.

- Limitations
  - Stylistic effects?
  - Speaker metadata?
  - Exploring the effect of spatial density of sampling locations for our spatial methods, especially when it diverges from underlying variation in population density


gi
3
2
1
0
−1
−2
−3
green = *a lot of*
purple = *lots of*

Future Work

- Bigger project:
  - Compare our work with previous dialect mapping
  - Multivariate analyses (factor analysis, cluster analysis etc.)

- Future research
  - Improve accuracy of features
  - Sample underrepresented regions
  - More features!

References

Carver, Craig M. 1987. *American Regional Dialects: A Word Geography*. Ann Arbor: The University of Michigan Press.

Coats, Steven. 2019. A corpus of regional American language from YouTube. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen*, 79–91. Denmark. http://ceur-ws.org/Vol-2364/7_paper.pdf.

Coats, Steven. 2024. Naturalistic Double Modals in North America. *American Speech: A Quarterly of Linguistic Usage* 99(1). 47–77. https://doi.org/10.1215/00031283-9766889.

Grieve, Jack. 2016. *Regional Variation in Written American English*. 1st edn. Cambridge University Press. https://doi.org/10.1017/CBO9781139506137.

Kim, Chaeyoon, Sravana Reddy, James Stanford, Ezra Wyschogrod & Jack Grieve. 2019. Bring on the crowd! Using online audio crowdsourcing for large-scale New England dialectology and acoustic sociophonetics. *American Speech* 94(2). 151--194. https://doi.org/10.1215/00031283-7251252.

Kurath, Hans. 1939. *Linguistic Atlas of New England. 6 vols. bound as 3*. Providence: Brown University for the American Council of Learned Societies.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The Atlas of North American English: Phonetics, phonology and sound change*. Berlin: Walter de Gruyter.

Leemann, Adrian, Péter Jeszenszky, Carina Steiner, Melanie Studerus & Jan Messerli. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard* 6(s3). https://doi.org/10.1515/lingvan-2020-0061.

Stanley, Joseph A. 2022. Regional Patterns in Prevelar Raising. *American Speech* 97(3). 374–411. https://doi.org/10.1215/00031283-9308384.

What can a corpus of YouTube videos tell us about grammatical variation in North American English?

Joseph A. Stanley
Brigham Young University

Brett Hashimoto
Brigham Young University

Jack Grieve
University of Birmingham

Download these slides at
joeystanley.com/ads2026