45th Annual Meeting of the Societas Linguistica Europaea,
Stockholm University, Sweden, 30 August 2012

An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches

Costanza Asnaghi, Jack Grieve

Prof. Maria Luisa Maggioni
Prof. Dirk Speelman

The Survey

Goals:

determine if Modern Standard Written Californian English contains regional variation;

map and describe these written dialects of English.

Basis:

a computational analysis of lexical variation in California online newspapers.

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

Evaluation

Conclusions

Why study Californian English?

1. Incomplete Previous Research

Major dialect surveys in the US
either excluded the West
or defined the West as one region:

1930s– Kurath's Linguistic Atlas of the United States and Canada (lexical survey);

1985 Cassidy's Dictionary of American Regional English (lexical survey);

2006 Labov's Atlas of North American English (phonological survey).

Overall, linguists assume there isn't much regional variation in the West.

Why study Californian English?

1. Incomplete Previous Research

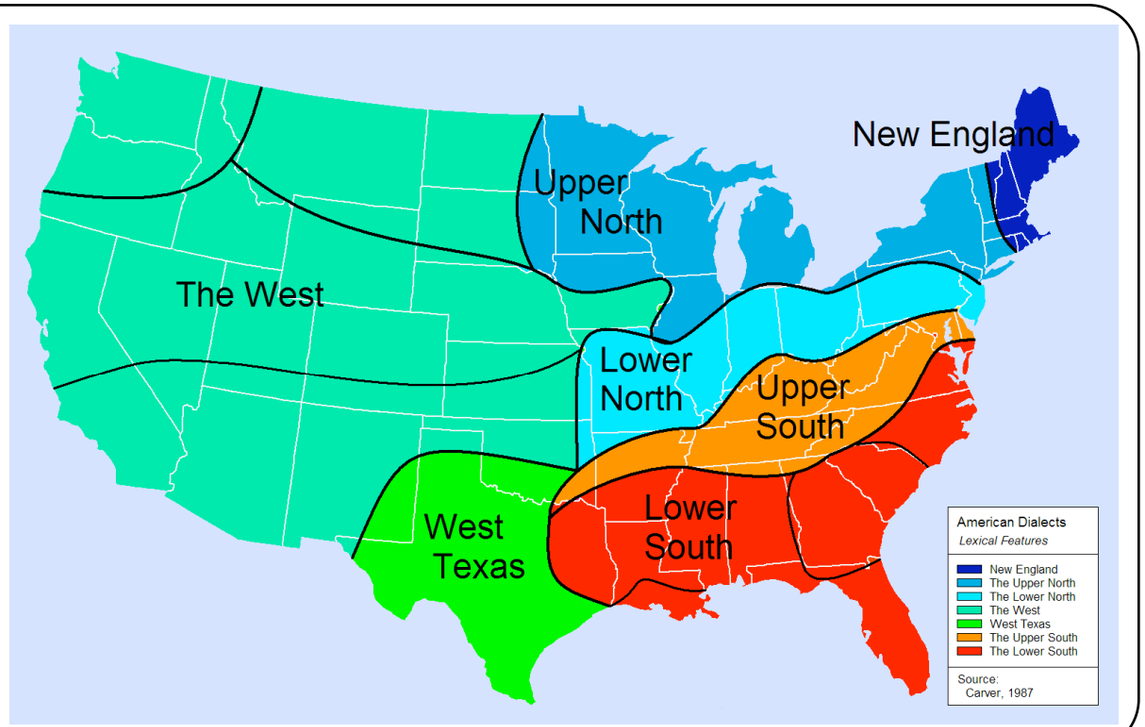
1930s– Kurath's Linguistic Atlas of the United States and Canada (lexical survey);

1985 Cassidy's Dictionary of American Regional English (lexical survey);

2006 Labov's Atlas of North American English (phonological survey).

1987
Carver's map
(based on Cassidy's survey)
divides the West
into North, North Central, and South.

He argues there's linguistic cohesion
in the West.



Why study Californian English?

1. Incomplete Previous Research

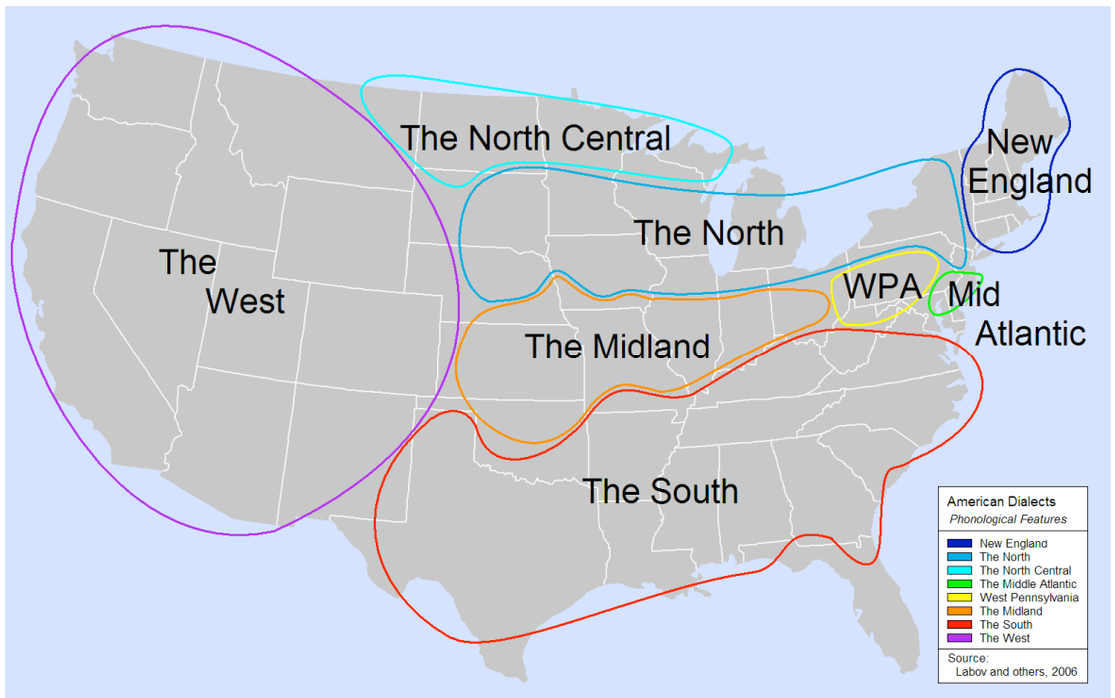
1930s– Kurath's Linguistic Atlas of the United States and Canada (lexical survey);

1985 Cassidy's Dictionary of American Regional English (lexical survey);

2006 Labov's Atlas of North American English (phonological survey).

2006
Labov's Atlas

Labov considers the West as one phonological area.



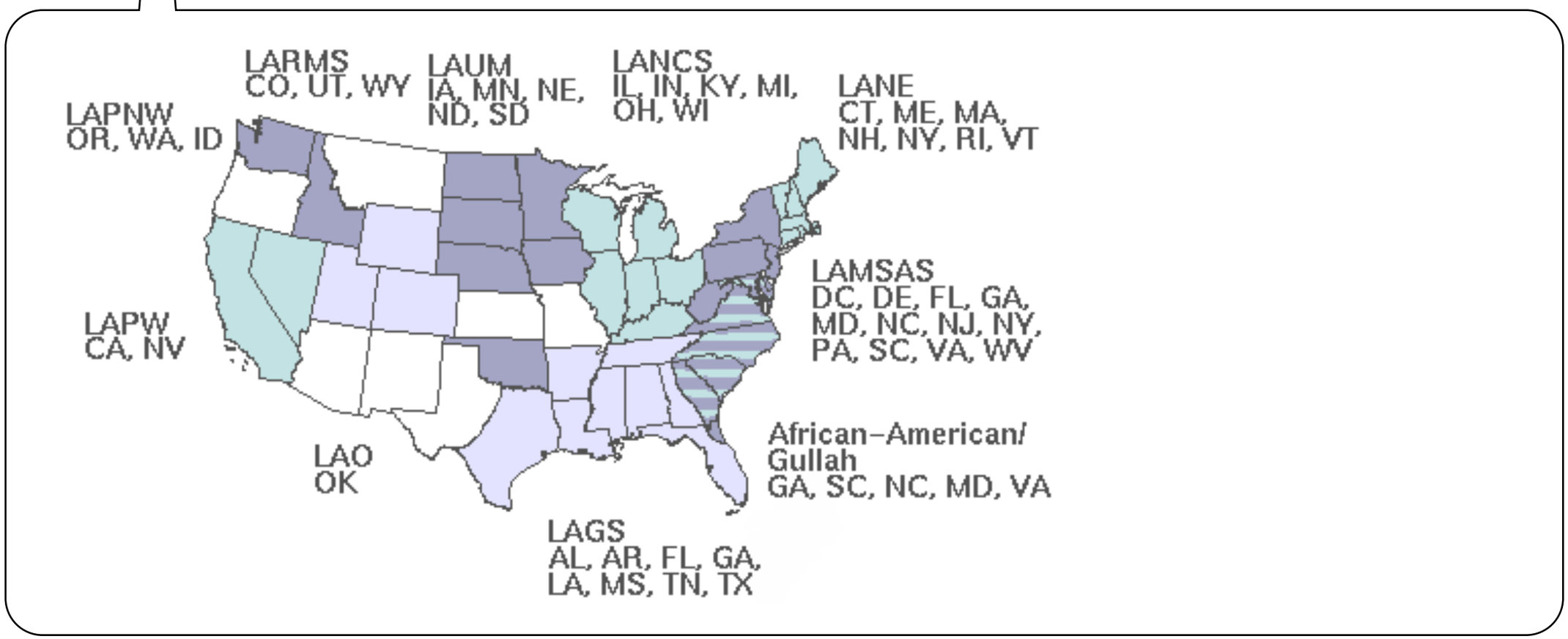
Why study Californian English?

1. Incomplete Previous Research

1930s– Kurath's Linguistic Atlas of the United States and Canada (lexical survey);

1985 Cassidy's Dictionary of American Regional English (lexical survey);

2006 Labov's Atlas of North American English (phonological survey).



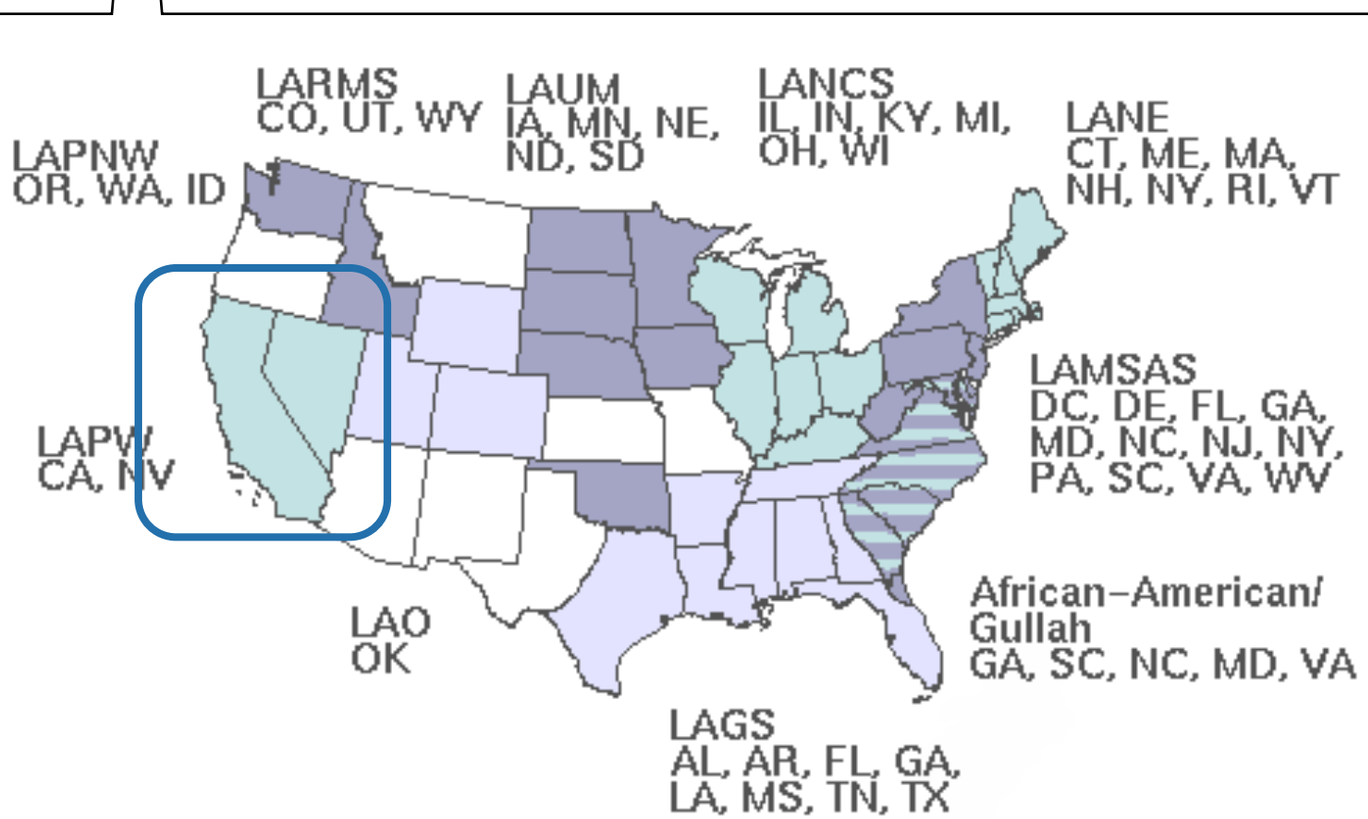
Why study Californian English?

1. Incomplete Previous Research

1930s– Kurath's Linguistic Atlas of the United States and Canada (lexical survey);

1985 Cassidy's Dictionary of American Regional English (lexical survey);

2006 Labov's Atlas of North American English (phonological survey).



The Linguistic Atlas Projects never really completed the analysis for the West Coast.

Although some data was collected, it was never analyzed with the rest of the US data.

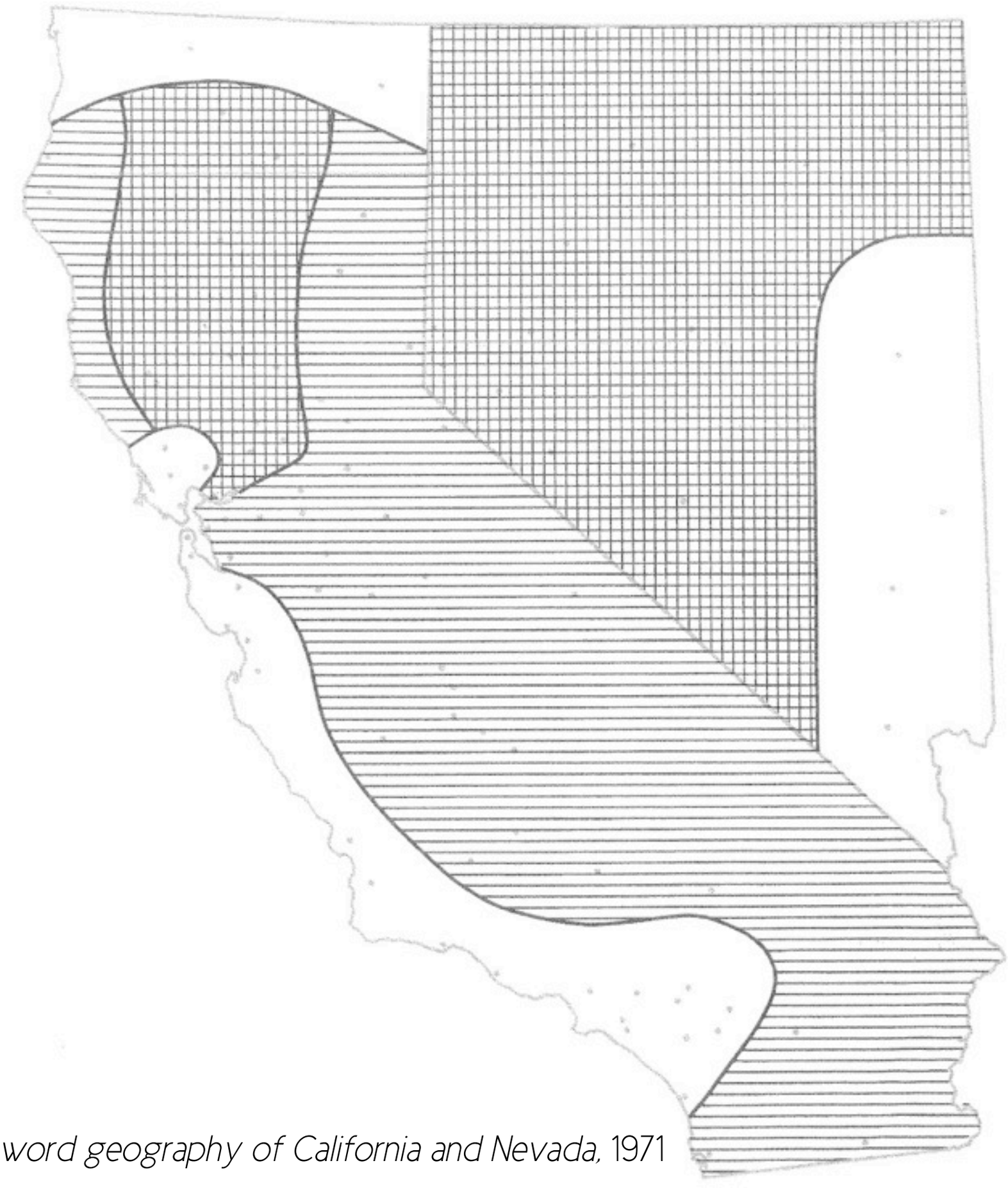
Why study Californian English?

1. Incomplete Previous Research

1971
Elizabeth Bright's
A Word Geography
of California and Nevada

based on 270 field records
eliciting lexical items
made by David Reed (1952–59)

Bright printed only one map
without really explaining it.



E. Bright, *A word geography of California and Nevada,* 1971

Why study Californian English?

2. Significant Population Size

Most populous state in the US
12% of the population

Surprising that previous dialectology studies
have not paid much attention to it,

Major gap in our knowledge of American dialect regions.

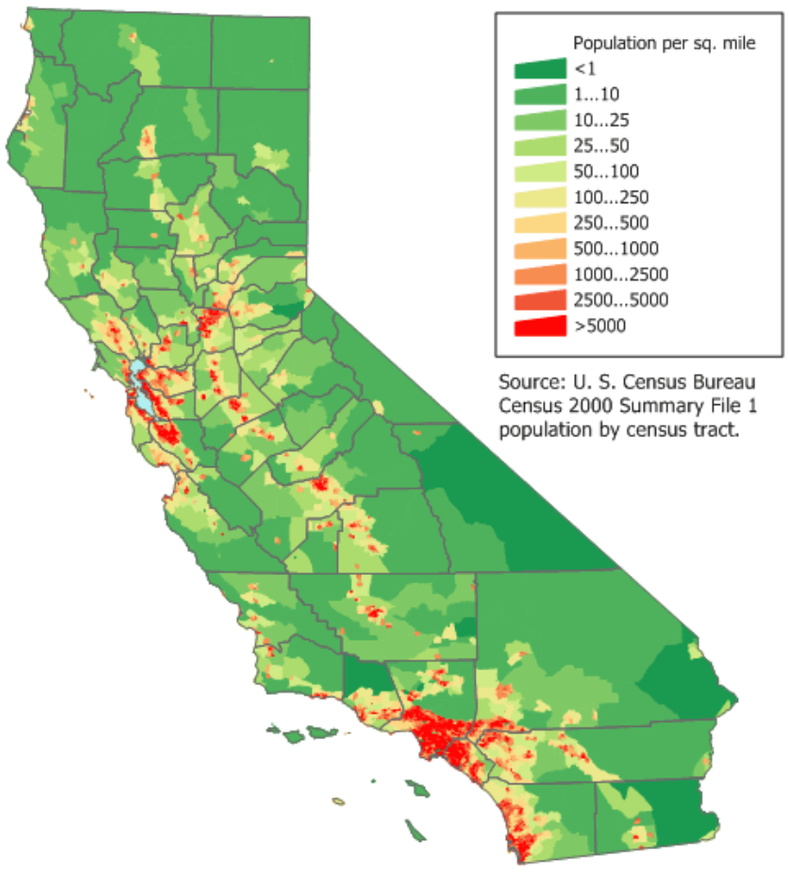
Why study Californian English?

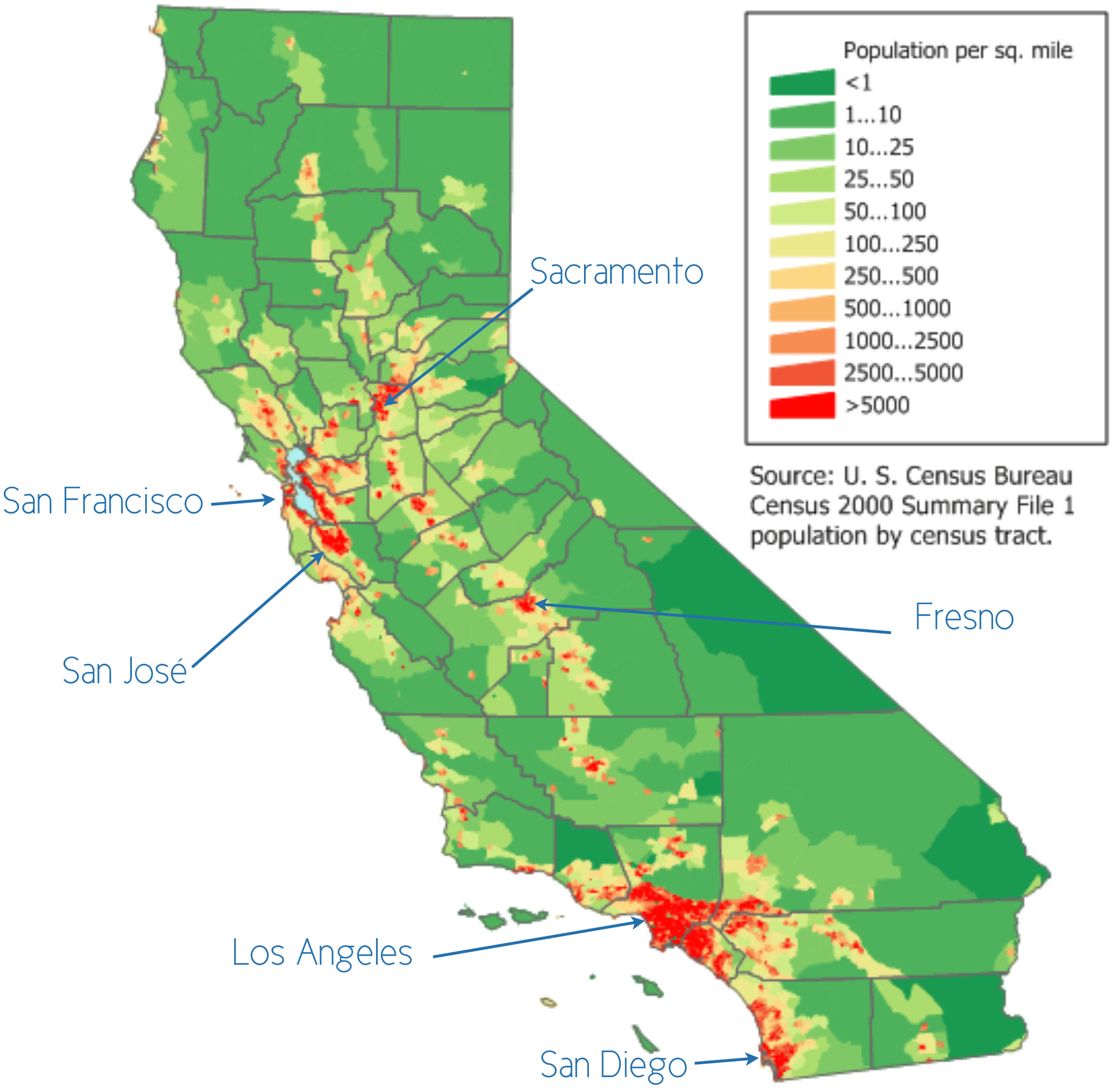
3. Multiple Population Centers

Multiple population centers

may lead to a more complicated picture of regional diversity

allows to compare the language in the different regions.



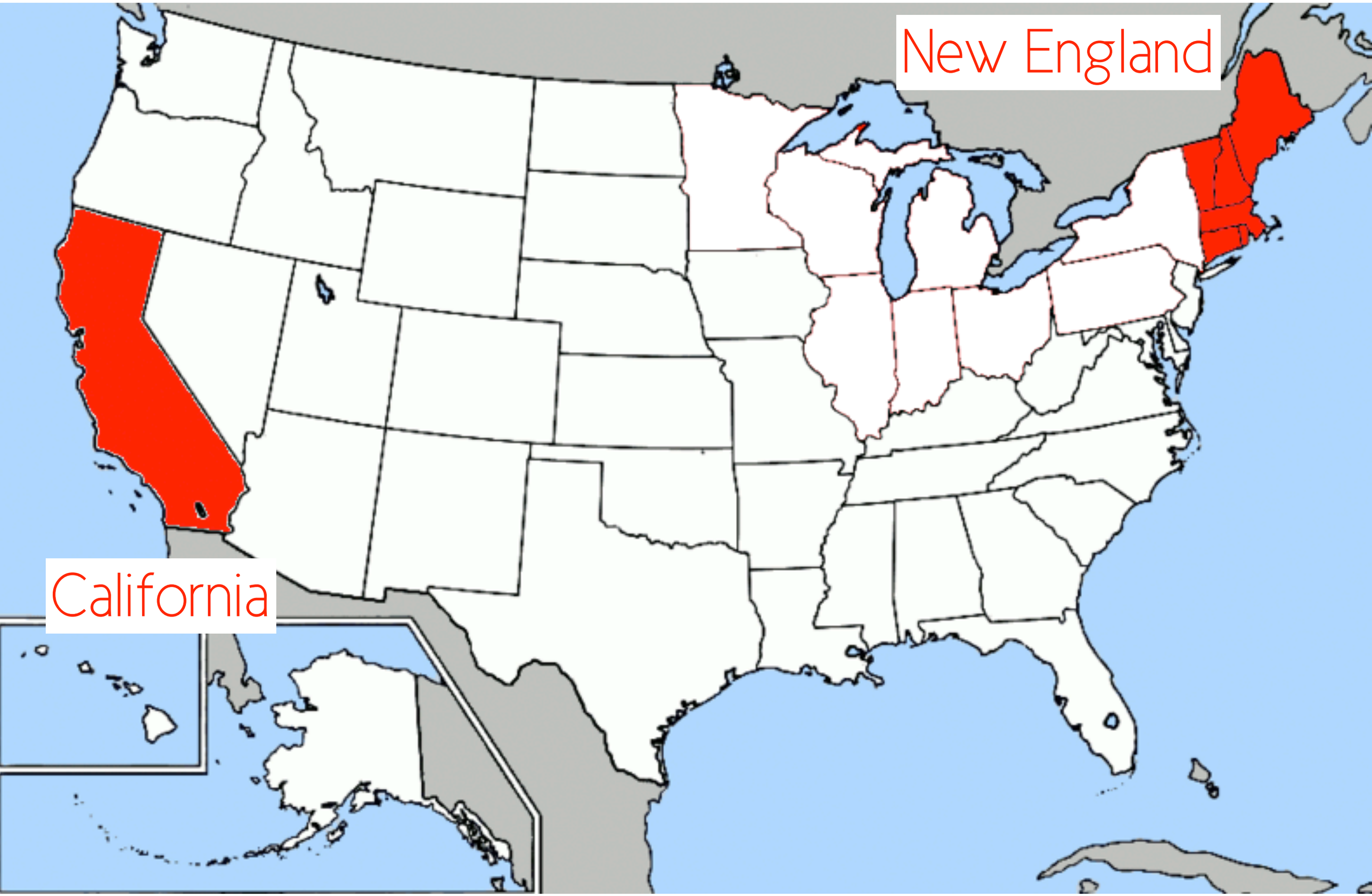

Population per sq. mile
<1
1...10
10...25
25...50
50...100
100...250
250...500
500...1000
1000...2500
2500...5000
>5000
Source: U. S. Census Bureau
Census 2000 Summary File 1
population by census tract.
Sacramento
San Francisco
San José
Fresno
Los Angeles
San Diego

Why study Californian English?

4. Significant Land Size

While previous studies concentrated
on the whole US
or on a big sub-region
(e.g. New England),

we can study regional variation
in one state
because of the significant size
of Californian land and population.


New England
California

Why study Californian English?

5. Relatively New American Dialect Region

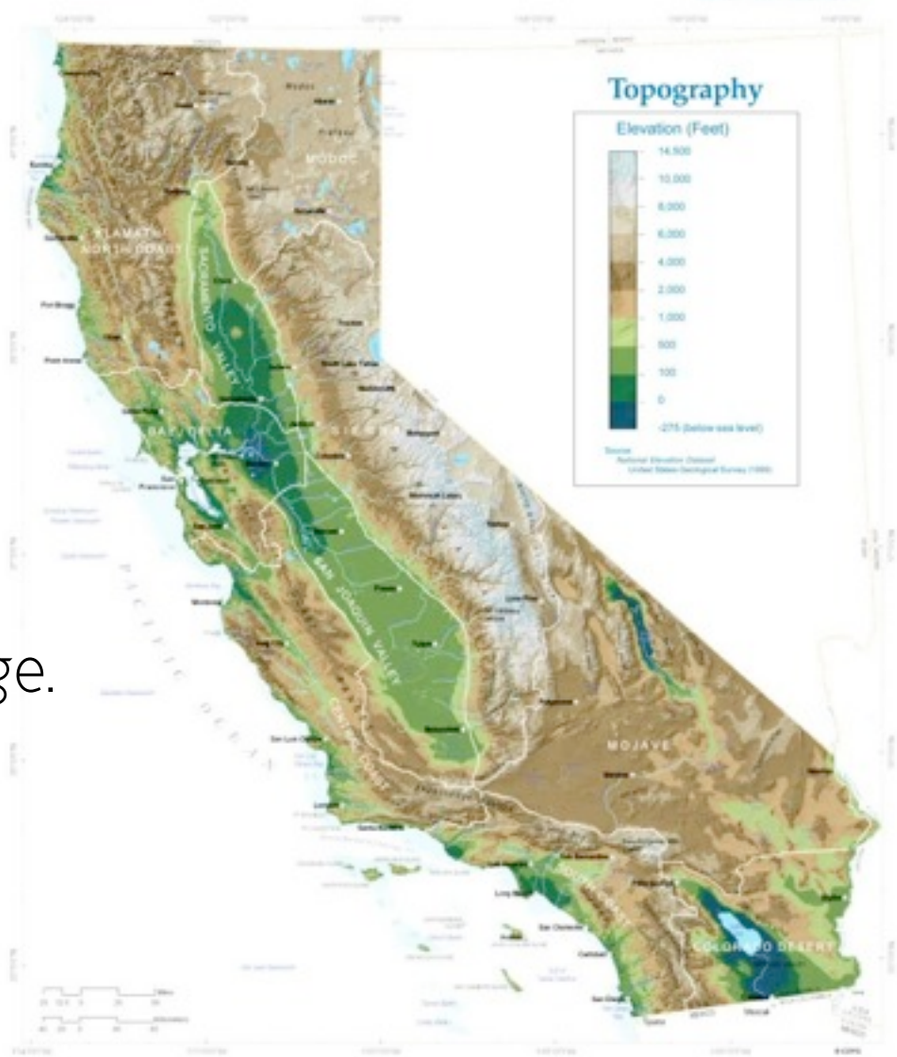
Relatively new American dialect region:

English was first spoken on a wide scale in California starting only from 1848 (the Gold Rush).

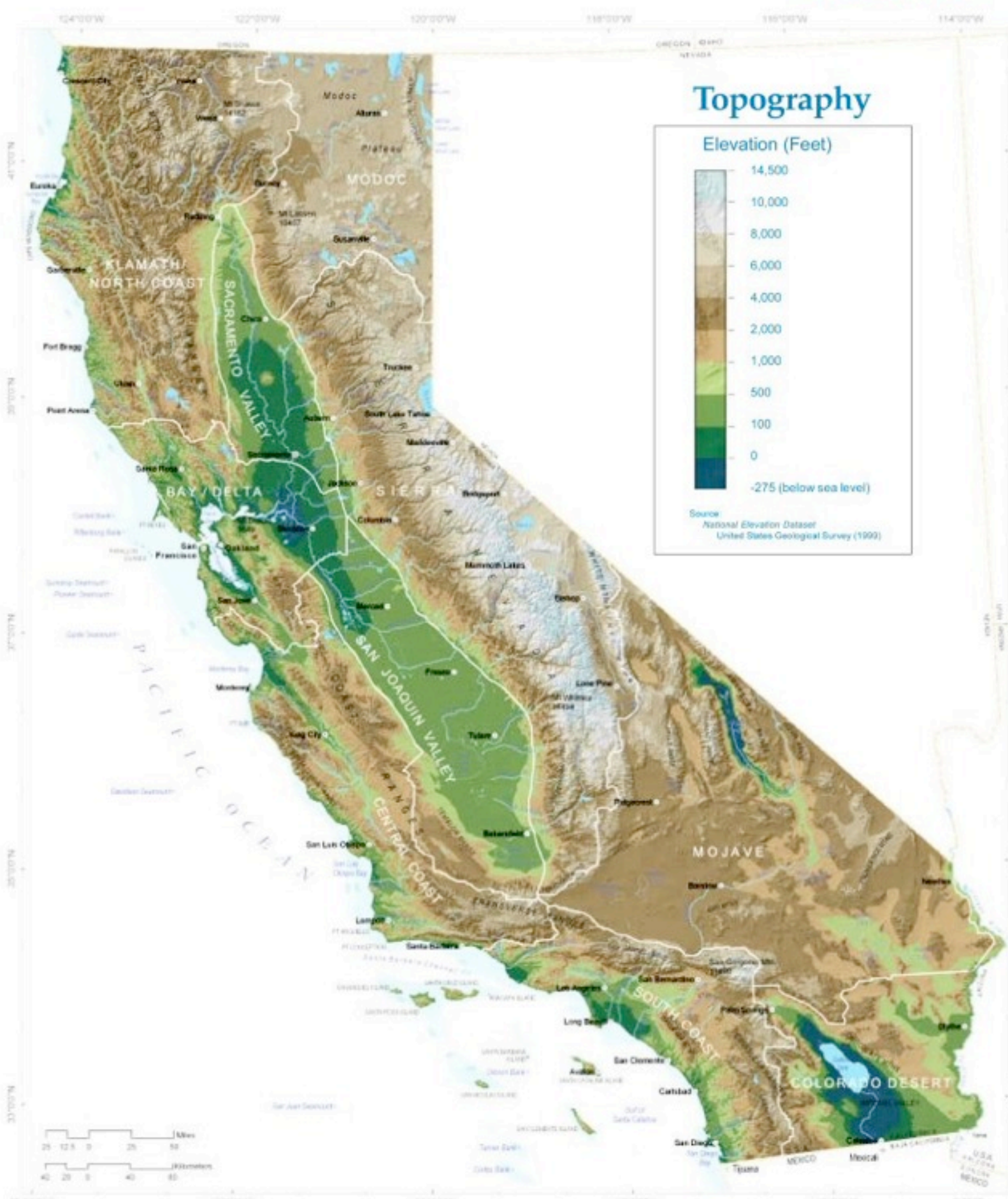
6. Physical Geography and Language

Why study California English?

Because of its topography and size, California is a good region to analyze the effect of physical geography (mountains, rivers, deserts) on language.



Topography


Elevation (Feet)
14,500
10,000
8,000
6,000
4,000
2,000
1,000
500
100
0
-275 (below sea level)
Source
National Elevation Dataset
United States Geological Survey (1999)
MODOC
Plateau
KLAMATH/
NORTH COAST
SACRAMENTO VALLEY
BAY / DELTA
SIERRA
SAN JOAQUIN VALLEY
COAST
RANGES
CENTRAL COAST
MOJAVE
SOUTH COAST
COLORADO DESERT
PACIFIC OCEAN
Miles
Kilometers
© CDFG

Why study California English?

7. Language Contact

We can look at the effect of language contact (e.g. Spanish) and dialect variation in a multi ethnical society (White, Hispanic, African American, Asian).

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

Evaluation

Conclusions

Specific Research Questions
Keywords: *Lexical Variation; Online Newspapers*

Do my new maps align with Bright's map (1971)?

Is there a north/south distinction?

Is there an inland/coastal distinction?

Is there a urban/rural distinction?

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

Evaluation

Conclusions

Traditional Methods of Data Collection

postal questionnaires
e.g. Davis 1948;

fieldworker interviews
e.g. Kurath 1939–43;

telephone interviews
e.g. Labov et al 2006;

traditional corpus-based techniques
e.g. Szmrecsanyi 2008, Grieve et al 2011.

Word Alternations

We collected a list of 41 word alternations, choosing variables mostly from previous dialectology studies[1].

1. Vaux 2003, Kurath 1949; Cassidy 1985–2002; Grieve 2009).

aim_purpose

assessment_evaluation

basement_cellar

car_automobile

cemetery_graveyard

dinner_supper

characteristic_feature

corridor_hallway

pail_bucket

pants_trousers

soda_coke

analysis_study

bag_sack

bro_brother

buddy_pal

client_customer

cloth_fabric

coat_jacket

concept_notion

context_framework

dad_father

earnings_revenue

expensive_costly

expert_specialist

grandma_grandmother

grandpa_grandfather

happiness_joy

holiday_vacation

ill_sick

law_legislation

mesa_butte

mom_mother

obstinate_stubborn

outcome_result

personnel_staff

porch_veranda

procedure_technique

regulation_rule

sundown_sunset

sunrise_dawn

trash_rubbish

New Method of Data Collection: Why?

A lexical dialect survey
requires
an extremely large corpus
or a large number of interviews:
lexical words do not appear often.

Newspaper Selection

Sample for this survey:
modern newspaper register of English, as published in mainstream newspapers from across California.
245 Californian newspapers from 176 Californian cities.

Why newspapers?
newspapers are plentiful,
freely available in machine readable form,
written in Standard English,
and annotated for their place of publication.

Register?
Distinguishing between registers can be important only at a different level of resolution;
not necessary for determining basic regional variation when dealing with a large set of data.

First Wave Sociolinguistics: "developing the big picture". (Eckert 2005)

New Method: Data Extraction Procedure

For each variant of a lexical alternation,
we counted the number of pages
containing that variant
in a series of city newspaper websites.

We used a Perl (LWP) script
to automatically query online search engines
and extract the number of hits
from the html source code for the results page.

Proportioning

After data collection,
we measured the alternation quantitatively as a proportion.

Example

newspaper:
    Los Angeles Times

website:
    latimes.com

alternation variable tested:
    pail/bucket

Google

site:latimes.com pail

Ricerca

Circa 3.840 risultati (0,38 secondi)

Web
Immagini
Maps
Video
Notizie
Shopping
Più contenuti

Milano
Cambia località

Nel Web
Pagine in italiano
Pagine da: Italia
Pagine straniere tradotte

Più strumenti

[Gift guide: Compost **pail**, cocktail glasses, blankets and more ...]
latimesblogs.latimes.com/.../best-gift... - Traduci questa pagina
26 Nov 2011 – Best gifts 2011: Our home and garden gift guide is loaded with picks -- some practical, some pure fun -- at every price. Clever piggy banks ...

[Detroit Red Wings Bucket: Galvanized 5 Quart **Pail**]
fanshop.latimes.com/Detroit-Red-W... - Traduci questa pagina
Detroit Red Wings Bucket: Galvanized 5 Quart **Pail** at LA Times. Buy your Detroit Red Wings Bucket: Galvanized 5 Quart **Pail** and from LA Times and proudly ...

[Edmonton Oilers Bucket: Galvanized 5 Quart **Pail**]
fanshop.latimes.com/Edmonton-Oil... - Traduci questa pagina
Edmonton Oilers Bucket: Galvanized 5 Quart **Pail** at LA Times. Buy your Edmonton Oilers Bucket: Galvanized 5 Quart **Pail** and from LA Times and proudly show ...

[UFC **Pail**: 5-Quart]
fanshop.latimes.com/UFC-**Pail**-5-Qu... - Traduci questa pagina
Buy your UFC **Pail**: 5-Quart and from LA Times and proudly show off your fandom! LA Times carries a full selection of Cooler from the MMA and all your favorites.

[Airline Ticket Consolidators Look to Improve **Pail** Image - Los ...]
articles.latimes.com/.../tr-15539_1_c... - Traduci questa pagina
16 Jun 1996 – Two questions arrive in this department with great frequency.One: What is a bucket shop?Two: What does an airline consolidator do?There is ...

[Team USA Bucket: 17 Quart Olympics **Pail**]
fanshop.latimes.com/Team-USA-Bu... - Traduci questa pagina
Buy your Team USA Bucket: 17 Quart Olympics **Pail** and from LA Times and proudly show off your fandom! LA Times carries a full selection of Cooler from the ...

Google site:latimes.com pail

Ricerca Circa 3.840 risultati (0,38 secondi)

Google site:latimes.com bucket

Ricerca Circa 12.900 risultati (0,21 secondi)

Web
Immagini
Maps
Video
Notizie
Shopping
Più contenuti

Milano
Cambia località

Nel Web
Pagine in italiano
Pagine da: Italia
Pagine straniere tradotte
Più strumenti

[San Diego student complains of being forced to urinate in **bucket ...**](#)
latimesblogs.latimes.com/.../teacher-... - Traduci questa pagina
13 Mar 2012 – A tenured teacher at a San Diego high school has been put on paid leave while the district investigates an allegation that she refused to allow a ...

[Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat](#)
fanshop.latimes.com/Los-Angeles-D... - Traduci questa pagina
Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat at LA Times. Buy your Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat and from LA ...

[Roseanne for pres: A chicken in every **bucket**, a pie in every face ...](#)
opinion.latimes.com/.../roseanne-pre... - Traduci questa pagina
7 Feb 2012 – In a review last year of Roseanne Barr's new reality TV series "Roseanne's Nuts," Times TV critic Mary McNamara noted that the show ...

[Articles about **Bucket** - Los Angeles Times](#)
articles.latimes.com/keyword/**bucket** - Traduci questa pagina
Bucket News. Find breaking news, commentary, and archival information about **Bucket** From The Los Angeles Times.

['**Bucket** & Skinner's Epic Adventures': TV review - Los Angeles Times](#)
articles.latimes.com/.../la-et-**bucket**-s... - Traduci questa pagina
1 Jul 2011 – The most, and almost the only, surprising thing about "**Bucket** & Skinner's Epic Adventures," a new tweencom debuting Friday on Nickelodeon ...

[Detroit Pistons Party Cooler **Bucket**](#)
fanshop.latimes.com/Detroit-Pistons... - Traduci questa pagina

Google site:latimes.com pail

Ricerca Circa 3.840 risultati (0,38 secondi)

Google site:latimes.com bucket

Ricerca Circa 12.900 risultati (0,21 secondi)

$$\frac{3.840}{3.840 + 12.900} = 0.2293$$

Web
Immagini
Maps
Video
Notizie
Shopping
Più contenuti

Milano
Cambia località

Nel Web
Pagine in italiano
Pagine da: Italia
Pagine straniere tradotte
Più strumenti

[San Diego student complains of being forced to urinate in **bucket ...**]
latimesblogs.latimes.com/.../teacher-... - Traduci questa pagina
13 Mar 2012 – A tenured teacher at a San Diego high school has been put on paid leave while the district investigates an allegation that she refused to allow a ...

[Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat]
fanshop.latimes.com/Los-Angeles-D... - Traduci questa pagina
Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat at LA Times. Buy your Los Angeles Dodgers Newborn and Infant Mascot **Bucket** Hat and from LA ...

[Roseanne for pres: A chicken in every **bucket**, a pie in every face ...]
opinion.latimes.com/.../roseanne-pre... - Traduci questa pagina
7 Feb 2012 – In a review last year of Roseanne Barr's new reality TV series "Roseanne's Nuts," Times TV critic Mary McNamara noted that the show ...

[Articles about **Bucket** - Los Angeles Times]
articles.latimes.com/keyword/**bucket** - Traduci questa pagina
Bucket News. Find breaking news, commentary, and archival information about **Bucket** From The Los Angeles Times.

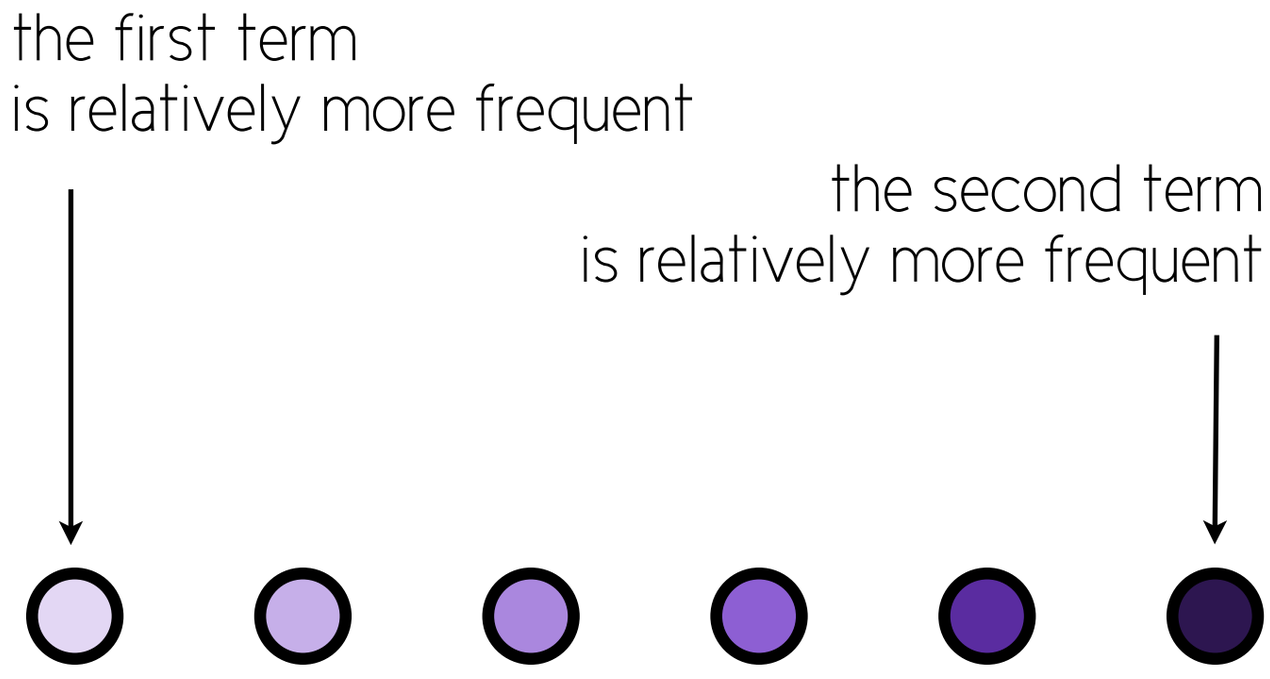
['**Bucket** & Skinner's Epic Adventures': TV review - Los Angeles Times]
articles.latimes.com/.../la-et-**bucket**-s... - Traduci questa pagina
1 Jul 2011 – The most, and almost the only, surprising thing about "**Bucket** & Skinner's Epic Adventures," a new tweencom debuting Friday on Nickelodeon ...

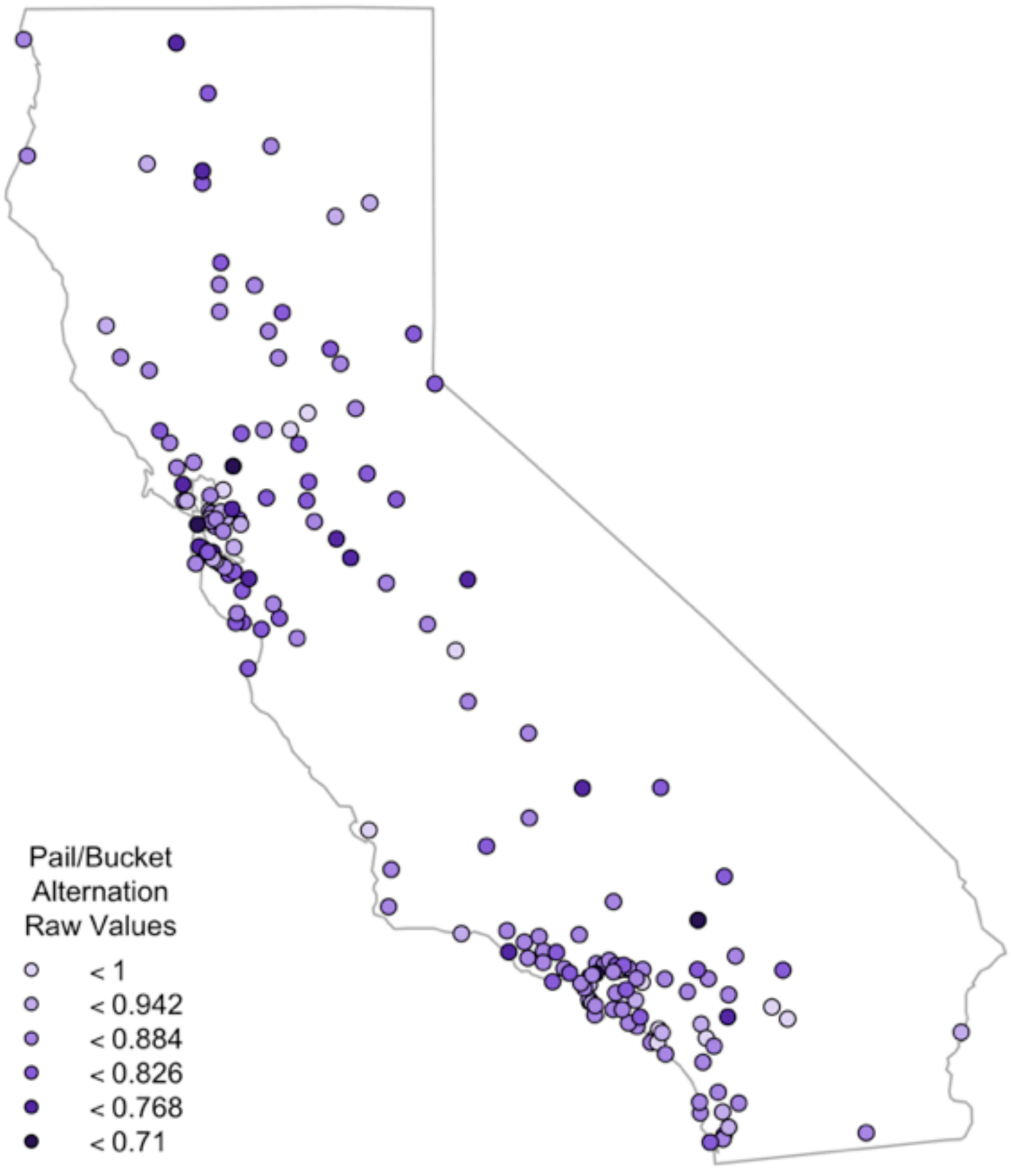
[Detroit Pistons Party Cooler **Bucket**]
fanshop.latimes.com/Detroit-Pistons... - Traduci questa pagina

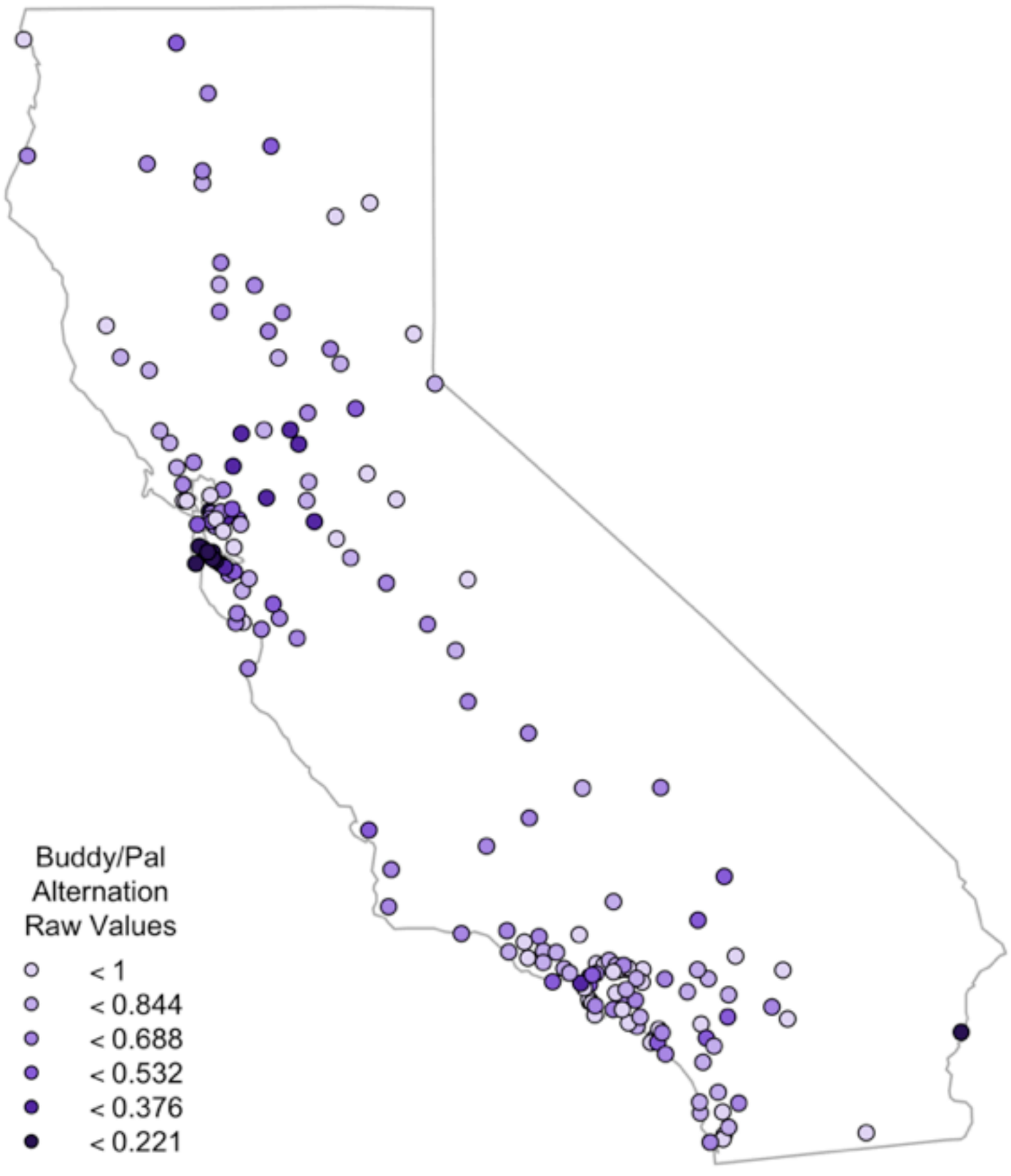
Raw Maps Plotting

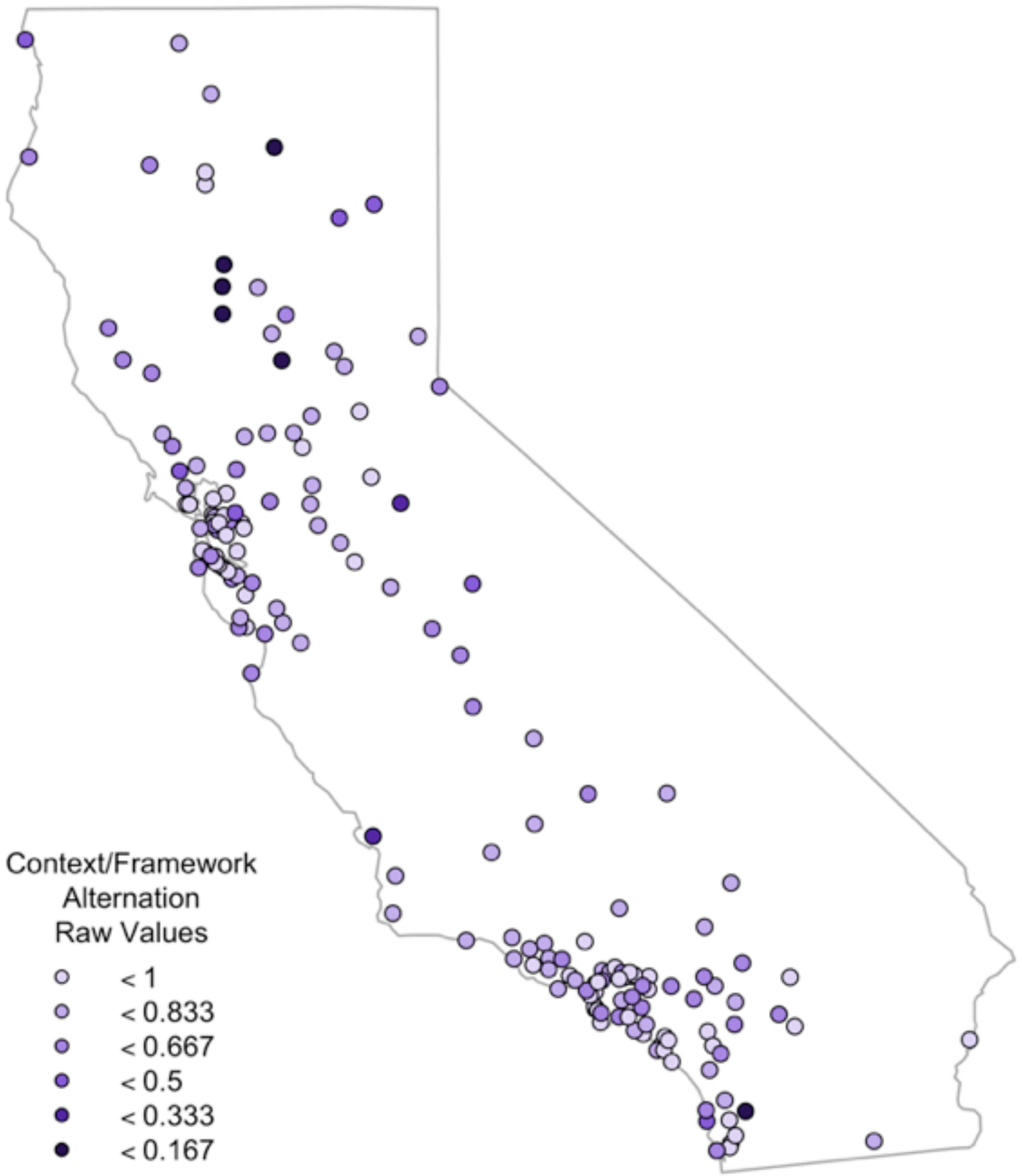
Maps out of the calculated proportions.

Frequency of the terms:




Pail/Bucket
Alternation
Raw Values
< 1
< 0.942
< 0.884
< 0.826
< 0.768
< 0.71


Buddy/Pal
Alternation
Raw Values
< 1
< 0.844
< 0.688
< 0.532
< 0.376
< 0.221


Context/Framework
Alternation
Raw Values
< 1
< 0.833
< 0.667
< 0.5
< 0.333
< 0.167

Local Getis-Ord Gi Autocorrelated Maps

Local Getis-Ord Gi was calculated for each location
to test if that location is part of a high- or low- value cluster.

Local Getis-Ord Gi returns a z-score
indicating the degree to which a location
is surrounded by locations with similar values.

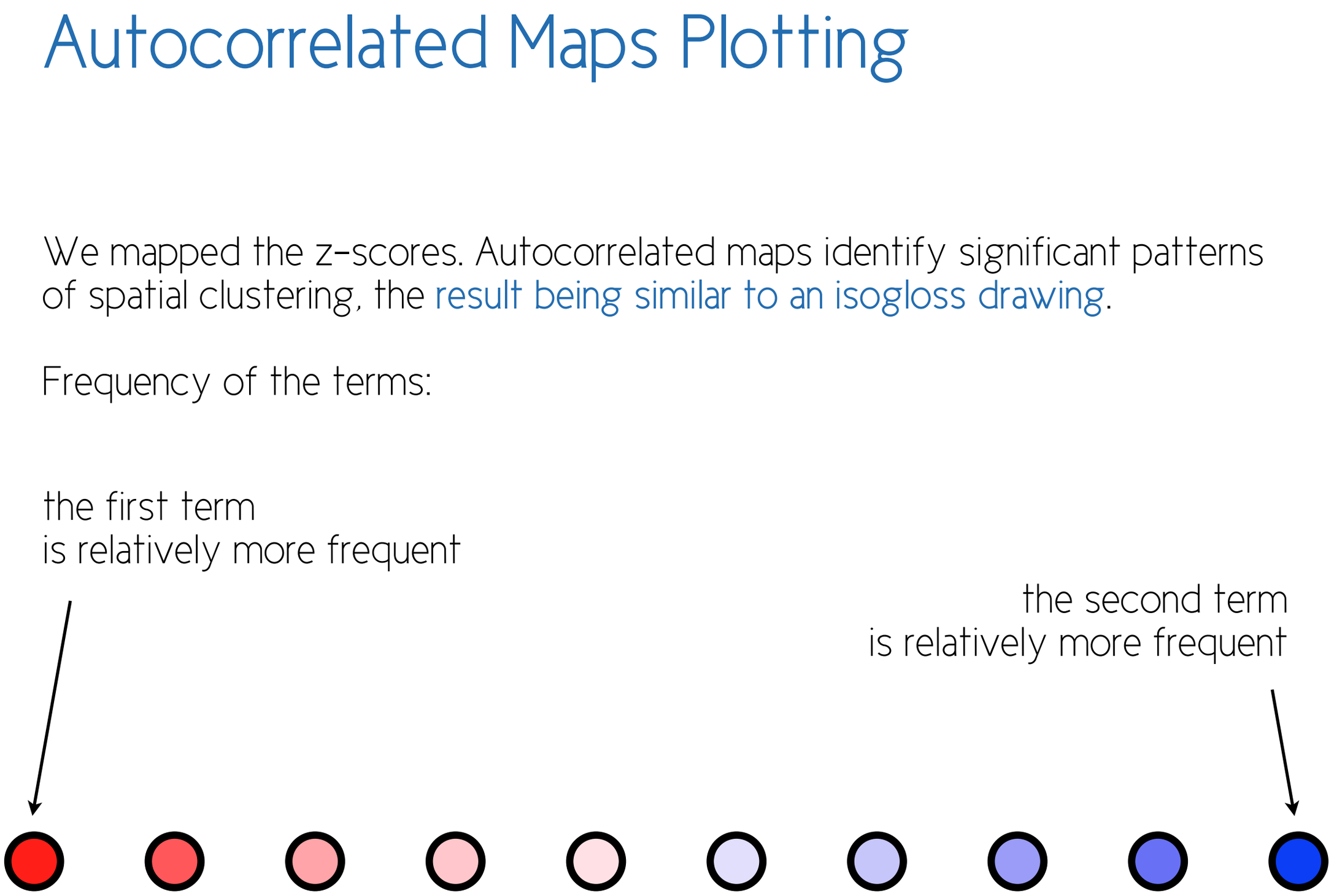
This statistical method smooths raw data,
cutting through the noise[1].
It is especially used in hot/cold spot testing, e.g. to detect crime hot spots.

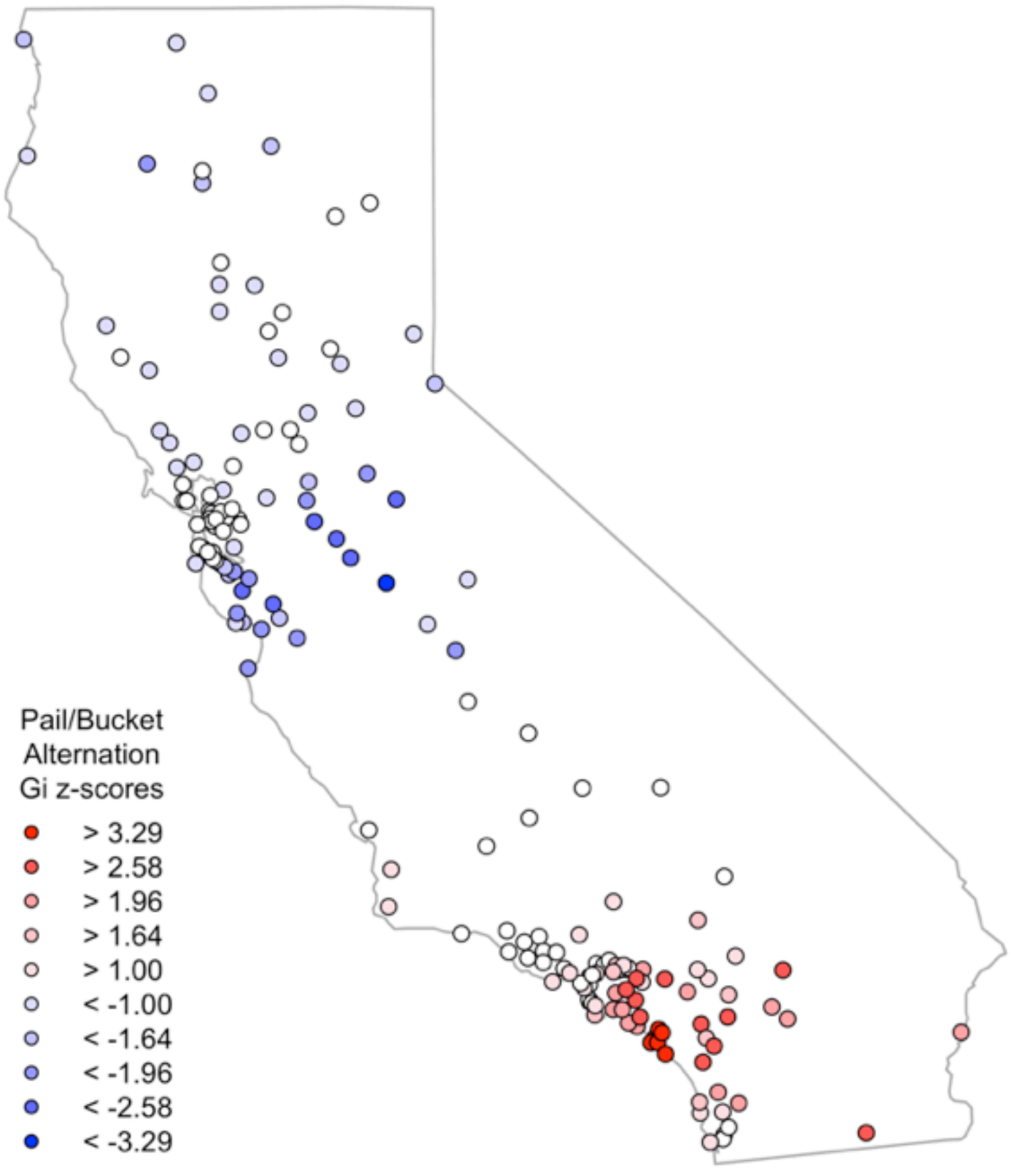
1. Ord and Getis 1995; Grieve 2011

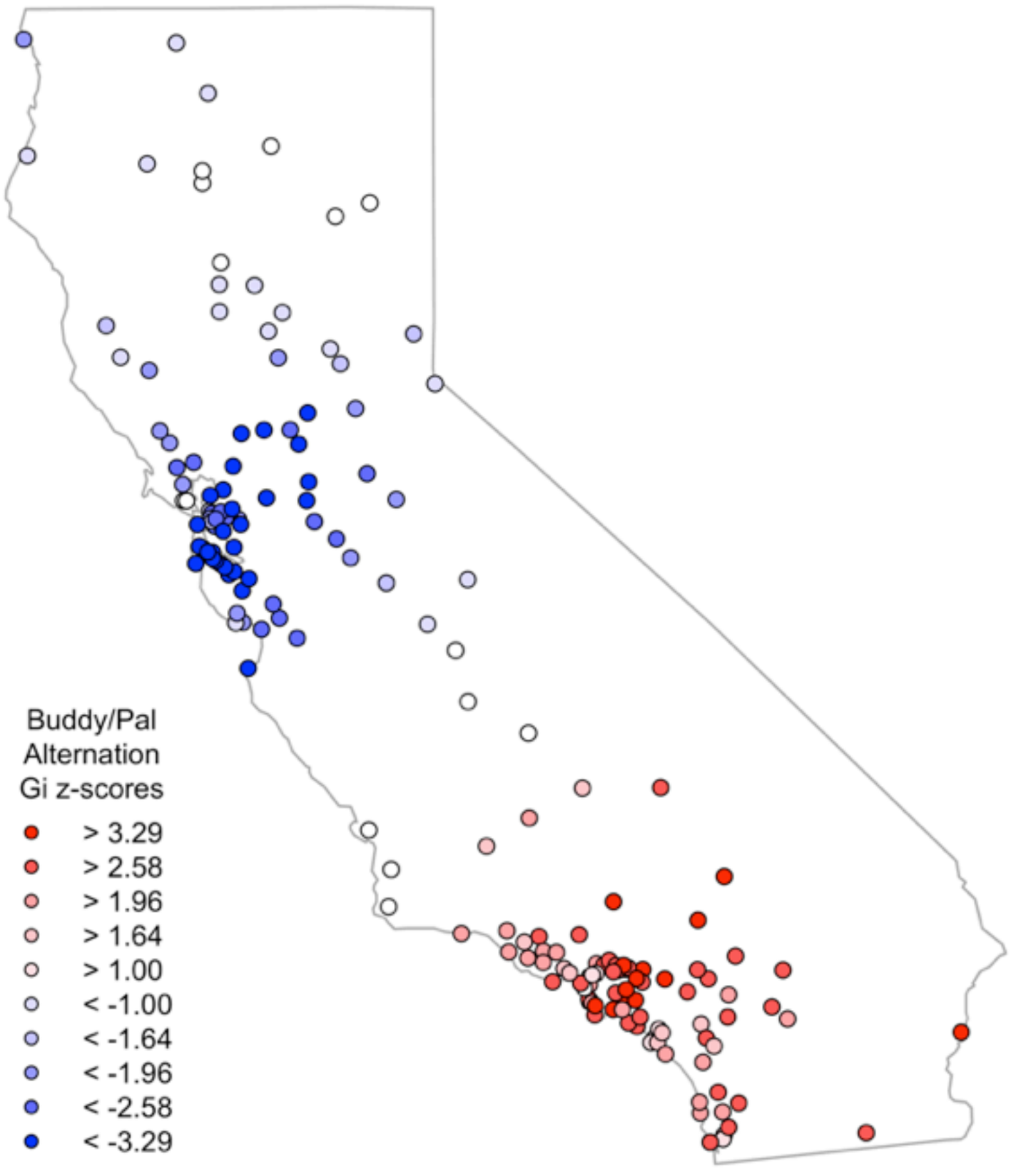
Autocorrelated Maps Plotting

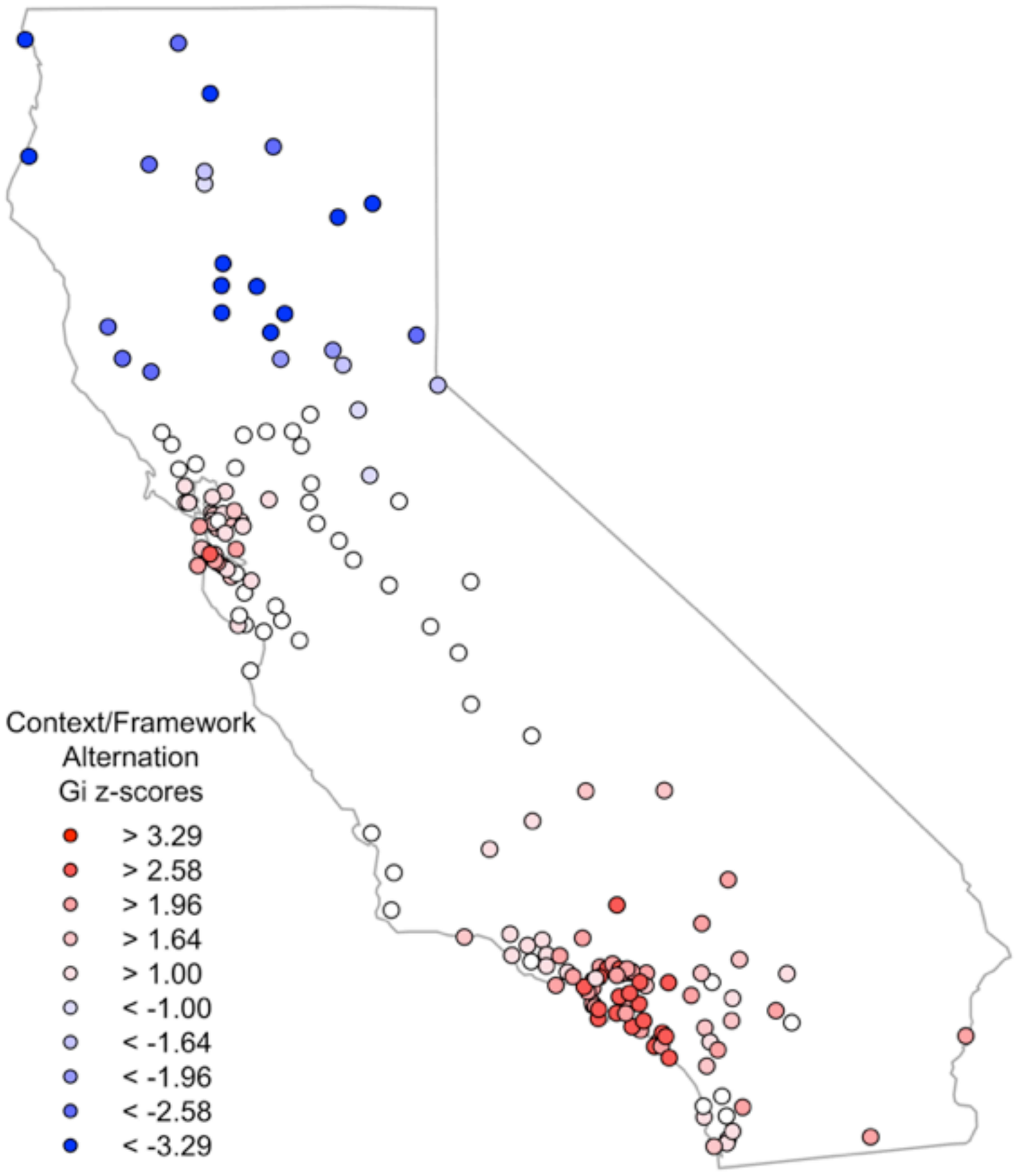
We mapped the z-scores. Autocorrelated maps identify significant patterns of spatial clustering, the result being similar to an isogloss drawing.

Frequency of the terms:




Pail/Bucket
Alternation
Gi z-scores
> 3.29
> 2.58
> 1.96
> 1.64
> 1.00
< -1.00
< -1.64
< -1.96
< -2.58
< -3.29


Buddy/Pal
Alternation
Gi z-scores
> 3.29
> 2.58
> 1.96
> 1.64
> 1.00
< -1.00
< -1.64
< -1.96
< -2.58
< -3.29

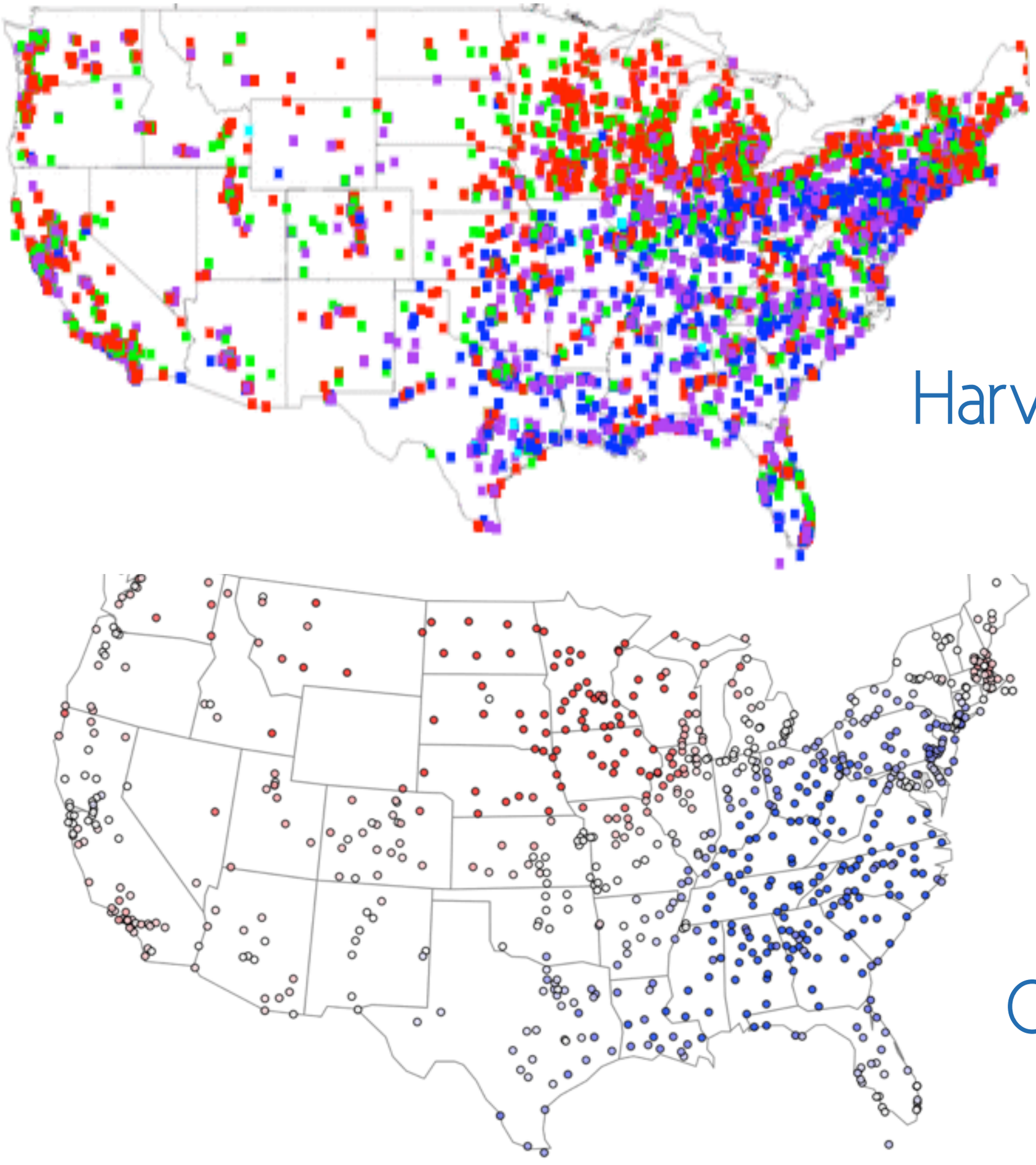

Context/Framework
Alternation
Gi z-scores
> 3.29
> 2.58
> 1.96
> 1.64
> 1.00
< -1.00
< -1.64
< -1.96
< -2.58
< -3.29

Method Evaluation

Method already validated in the US as a whole[1].

1. Grieve and Asnaghi 2011

Frosting / Icing



Harvard Dialect Survey

Our Survey

Method Evaluation: Cons

Polysemous words, idioms, proper names, unique localisms (Sunset Boulevard) and others.

Variants should be relatively unambiguous (e.g. creek/stream).

Alternations with too many of these problems (creek/stream) might mean they can't be analyzed using this method.

Method Evaluation: Pros

We can find regional patterns despite the noise
thanks to the quantity of data and the advanced statistics.

We compared our results
to the results of previous American dialect surveys.
In almost every case
the regional pattern identified by the web-based analysis
agreed with the results of previous dialect surveys.

Based on these results,
we believe that this approach is both a valid and efficient method for
gathering data on regional lexical variations.

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

Evaluation

Conclusions

Factor Analysis

Factor Analysis extracts a reduced set of factors
from a set of variables
that accounts for most of the variance in the variables.

Regression Analysis saves the scores so that it is possible to map the resulting factors.

Loadings:

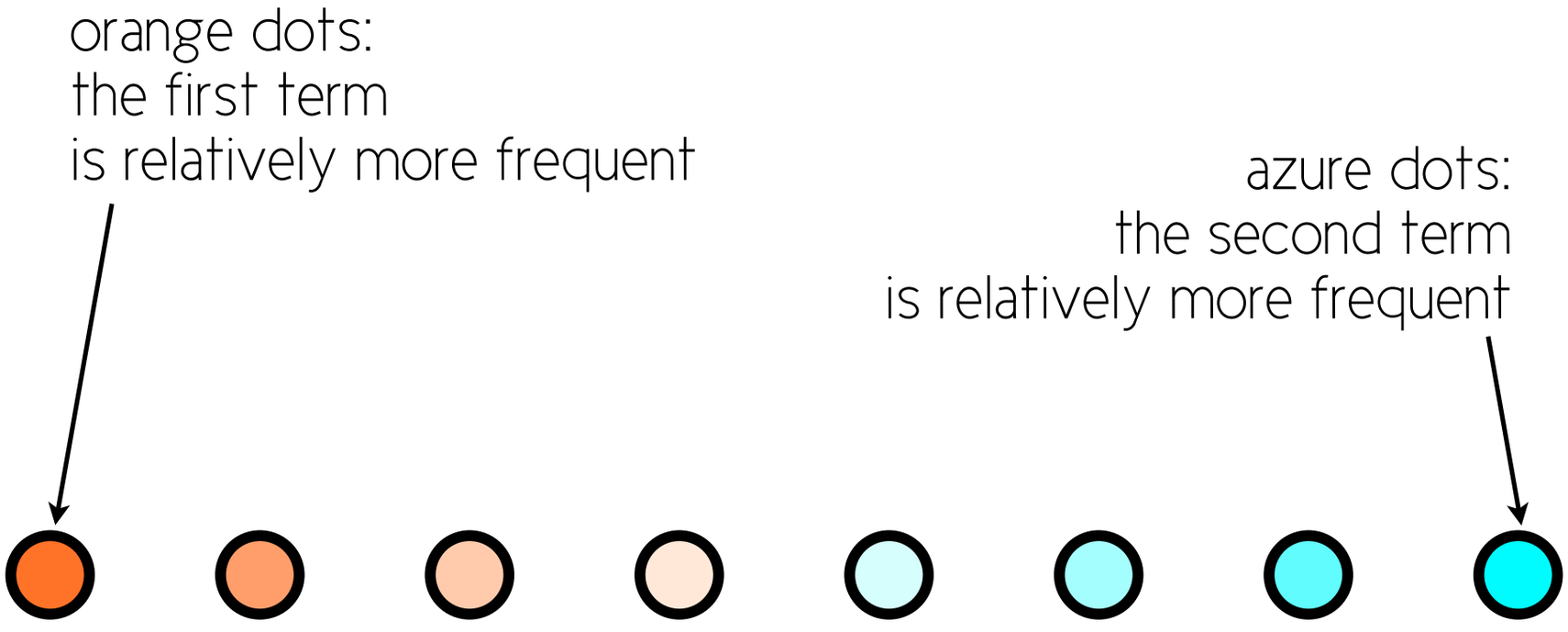
	Factor1	Factor2	Factor3
basement_cellar	-0.508	-0.216	0.168
cemetery_graveyard		-0.481	0.546
mesa_butte	0.467	-0.814	0.181
obstinate_stubborn	0.366	0.673	-0.354
porch_veranda	-0.308	0.571	-0.427
aim_purpose		0.316	
characteristic_feature	-0.654		
assessment_evaluation		0.108	
car_automobile			0.174
dinner_supper		0.461	
corridor_hallway	-0.365	0.473	-0.799
pail_bucket	0.366	-0.457	0.807
pants_trousers	-0.113		0.324
soda_coke			0.323
analysis_study		0.620	
bag_sack	0.709	0.168	-0.158
bro_brother	-0.187	0.792	
buddy_pal	-0.121	-0.779	0.515
client_customer		-0.241	0.360
cloth_fabric	-0.398	0.673	-0.436
coat_jacket	0.488	0.548	-0.427
concept_notion	0.126	-0.648	0.514
context_framework	0.819	-0.540	
dad_father	0.885	0.245	0.120
earnings_revenue	-0.592	0.309	-0.150
expensive_costly	0.627	-0.711	
expert_specialist	0.691	-0.232	
grandma_grandmother	0.632	-0.507	0.258
grandpa_grandfather		0.590	-0.302
happiness_joy	0.110	-0.353	0.167
holiday_vacation	0.854	-0.217	0.157
ill_sick	-0.863		
law_legislation	0.162	-0.565	0.124
mom_mother	0.896	0.114	0.144
outcome_result	0.488		0.122
personnel_staff	-0.643	-0.210	0.348
procedure_technique	-0.579	0.228	0.380
regulation_rule	-0.285	0.757	-0.333
sundown_sunset	0.275	0.581	-0.457
sunrise_dawn	-0.405		
trash_rubbish	0.459	-0.184	0.497

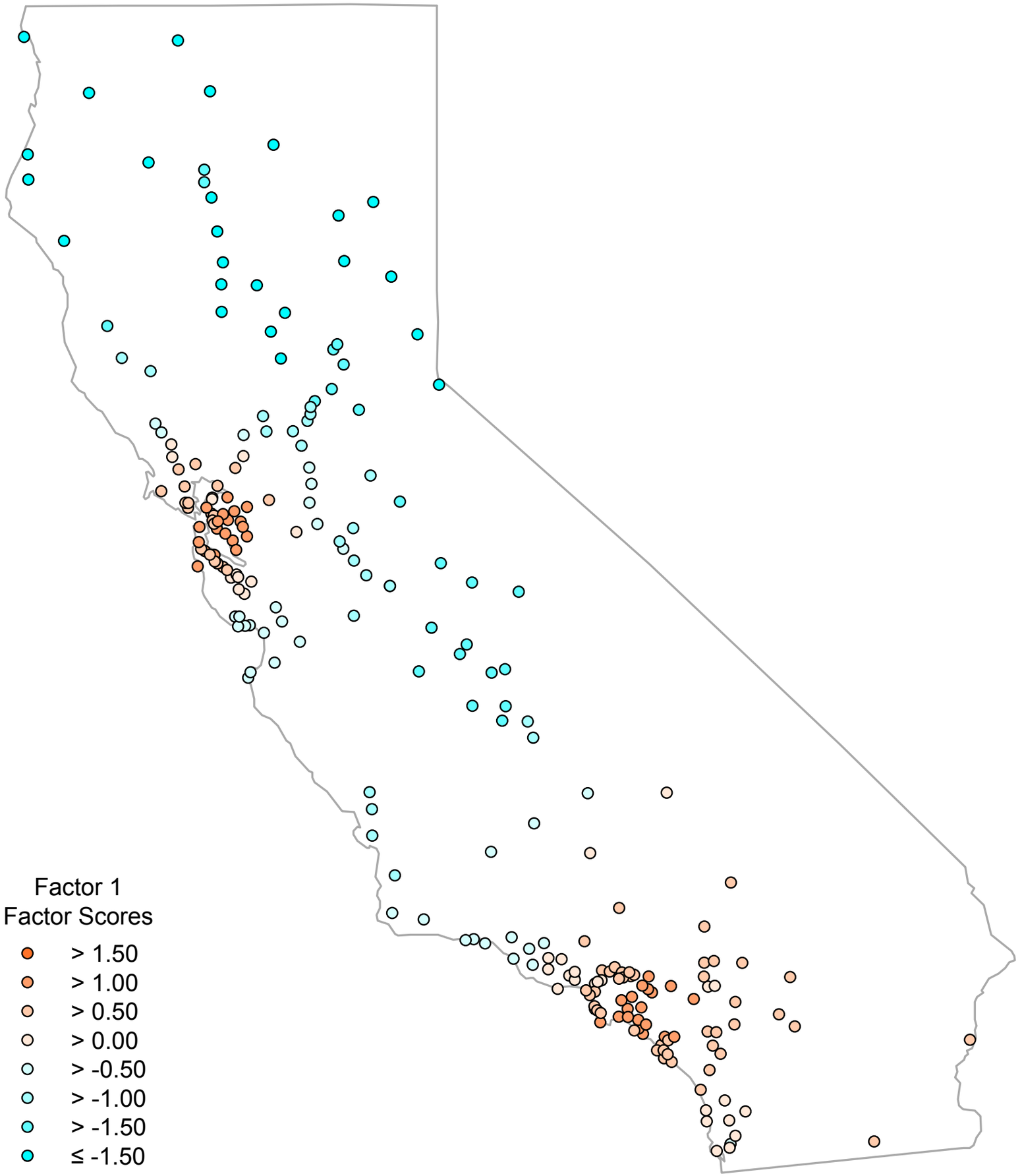
	Factor1	Factor2	Factor3
SS loadings	9.304	8.532	4.417
Proportion Var	0.227	0.208	0.108
Cumulative Var	0.227	0.435	0.543

Factor Maps Plotting

We mapped the factor scores in the locations.

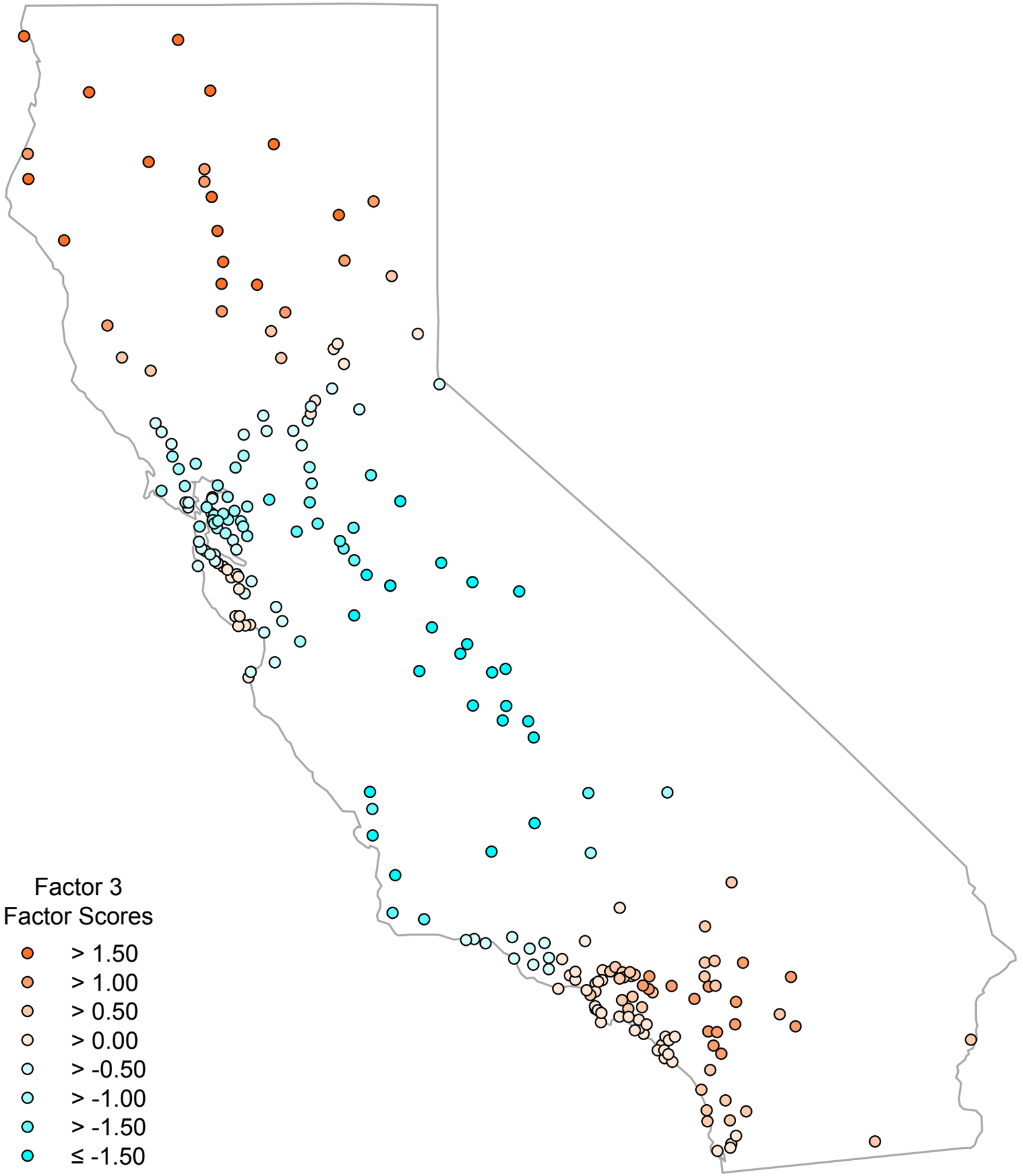
Frequency of the terms:




Factor 1
Factor Scores
> 1.50
> 1.00
> 0.50
> 0.00
> -0.50
> -1.00
> -1.50
≤ -1.50

Factor 2
Factor Scores

- > 1.50
- > 1.00
- > 0.50
- > 0.00
- > -0.50
- > -1.00
- > -1.50
- ≤ -1.50


Factor 3
Factor Scores
> 1.50
> 1.00
> 0.50
> 0.00
> -0.50
> -1.00
> -1.50
≤ -1.50

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

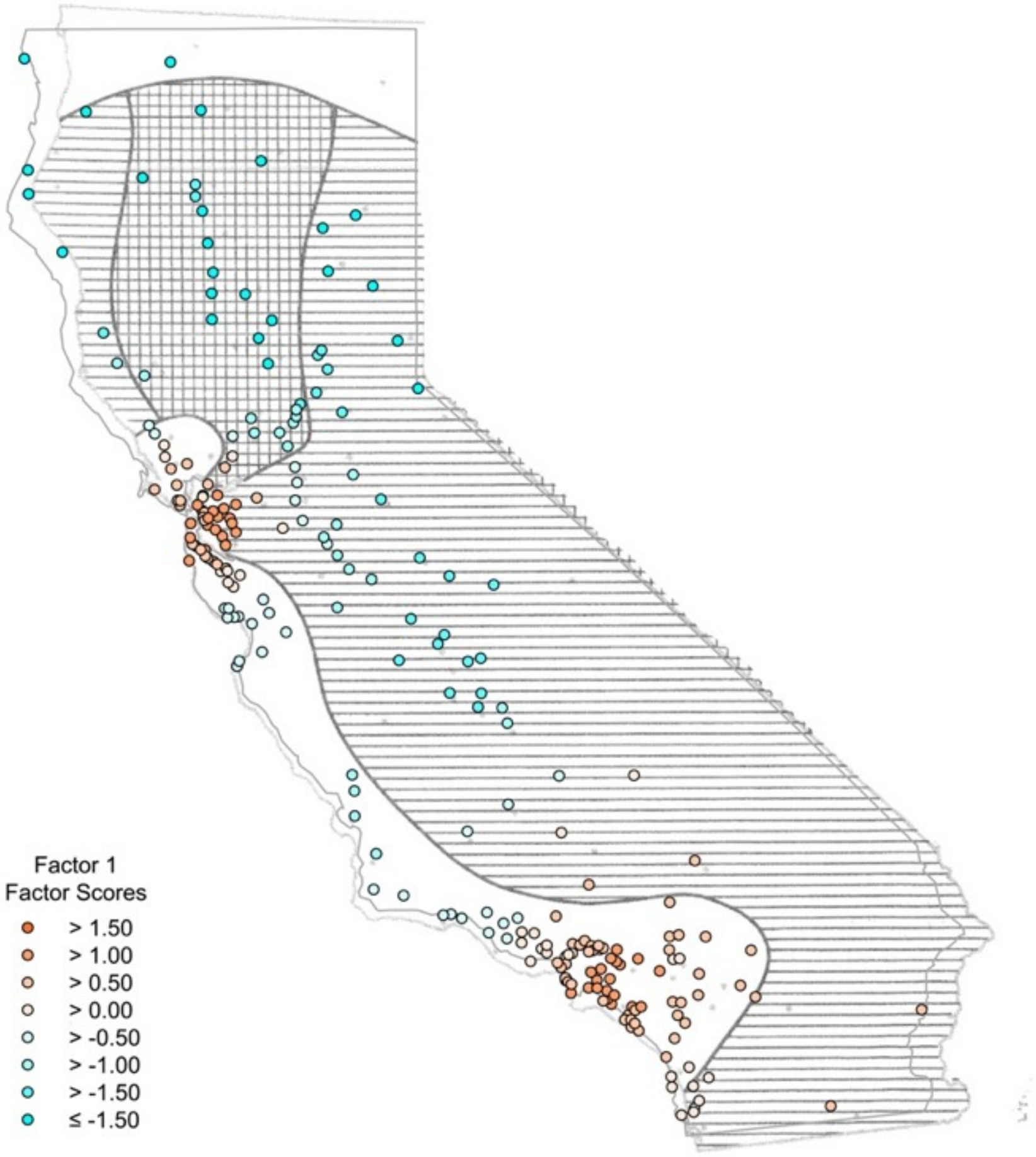
Evaluation

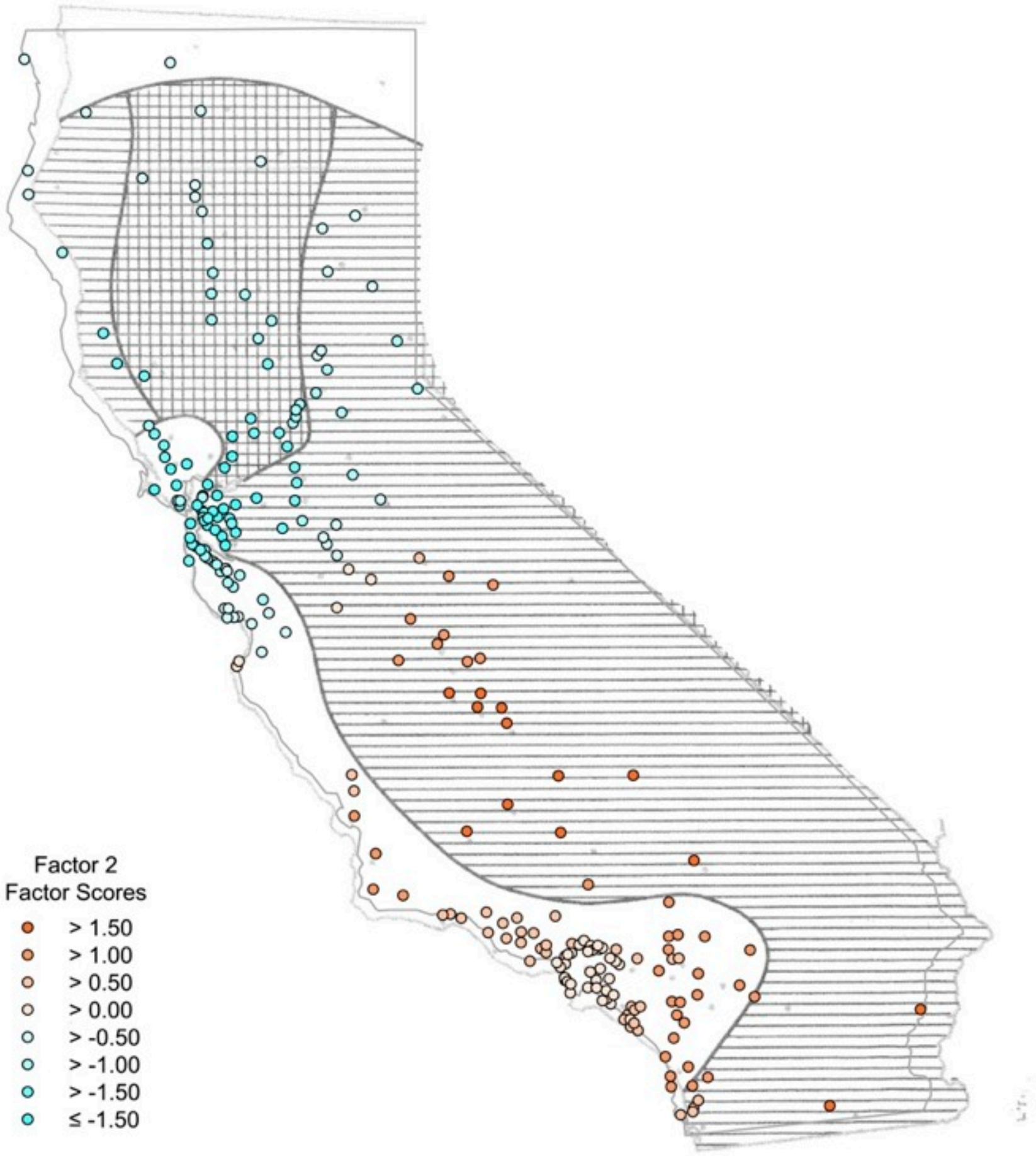
Conclusions

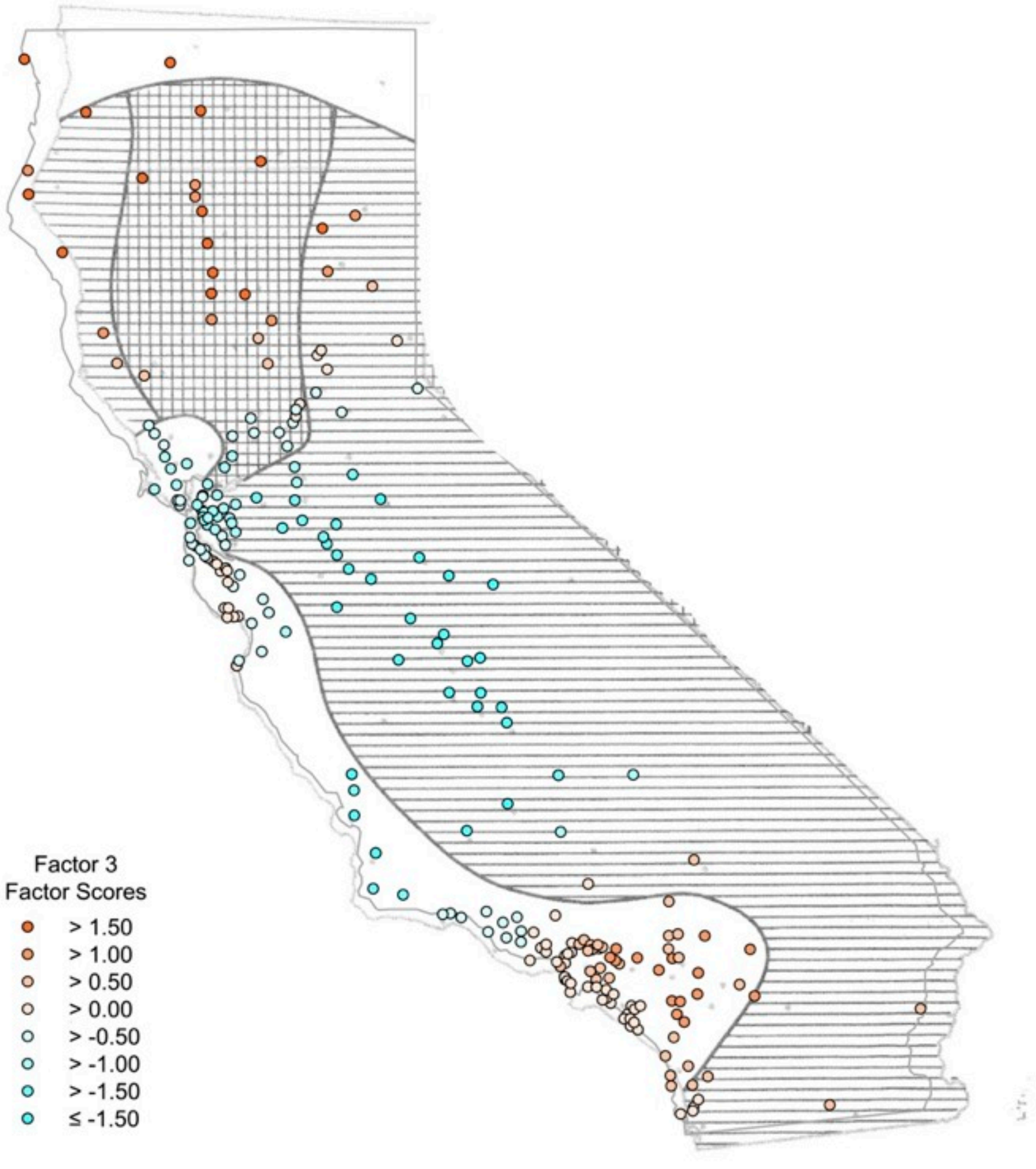
Evaluation

Our results do not align with Bright's work (1971)
except for the similarities found between Bright's map and Factor 1 map
(metropolitan areas);

Additional patterns:
North/South
Rural/Urban


Factor 1
Factor Scores
> 1.50
> 1.00
> 0.50
> 0.00
> -0.50
> -1.00
> -1.50
≤ -1.50


Factor 2
Factor Scores
> 1.50
> 1.00
> 0.50
> 0.00
> -0.50
> -1.00
> -1.50
≤ -1.50


Factor 3
Factor Scores
> 1.50
> 1.00
> 0.50
> 0.00
> -0.50
> -1.00
> -1.50
≤ -1.50

Structure

Why study Californian English?

Specific Research Questions

Method

Multivariate Statistics

Evaluation

Conclusions

Importance of Findings

We found regional lexical variation in standard written Californian English.

We completed a state-wide survey of Californian dialects.

Future Research

Did settlement patterns affect California dialects?

Are our results comparable with patterns of diffusion of languages other than English?

How is travel time relevant to predict dialect differences? (especially important in California due to the Sierra Nevada)

Thank you!

An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches

Costanza Asnaghi, Jack Grieve

thanks to Prof. Maggioni and Prof. Speelman