Interpolation, and the probabilistic view on linguistic diversity in general

– how to convince the linguists? –

Michael Cysouw
Deutscher Sprachatlas
Philipps-Universität Marburg

Beyond the symbol-map

- Visualisation of geographic distributions in linguistics is mostly symbol-based with added lines to signal boundaries
- But: the data are empirical probabilistic distributions without real boundaries
- Why do we linguists like boundaries so much?

Händler, Harald & Carl Ludwig Naumann. 1976.
Automatisierung der Isoglossenfindung.
Germanistische Linguistik 76(3-4). 123-159.

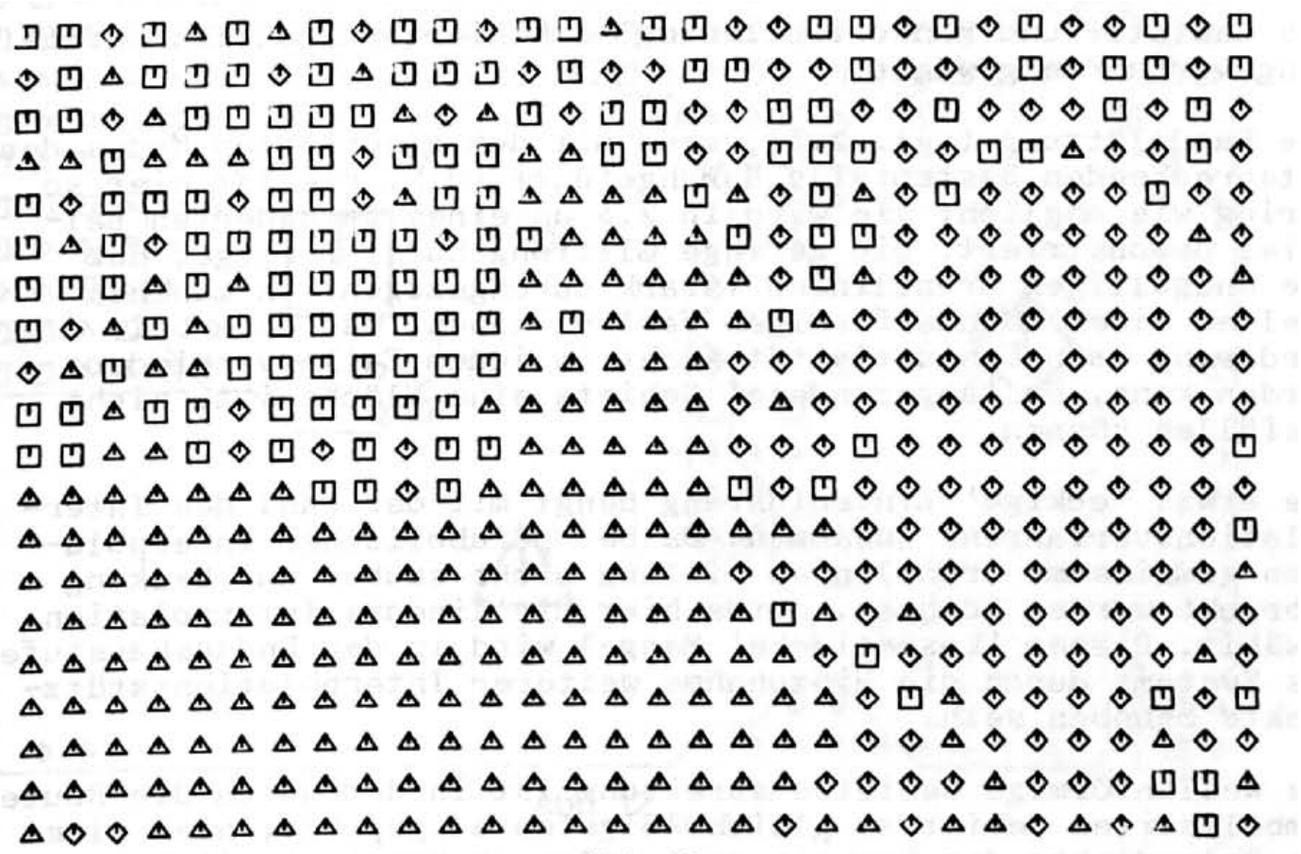


Abb. 10

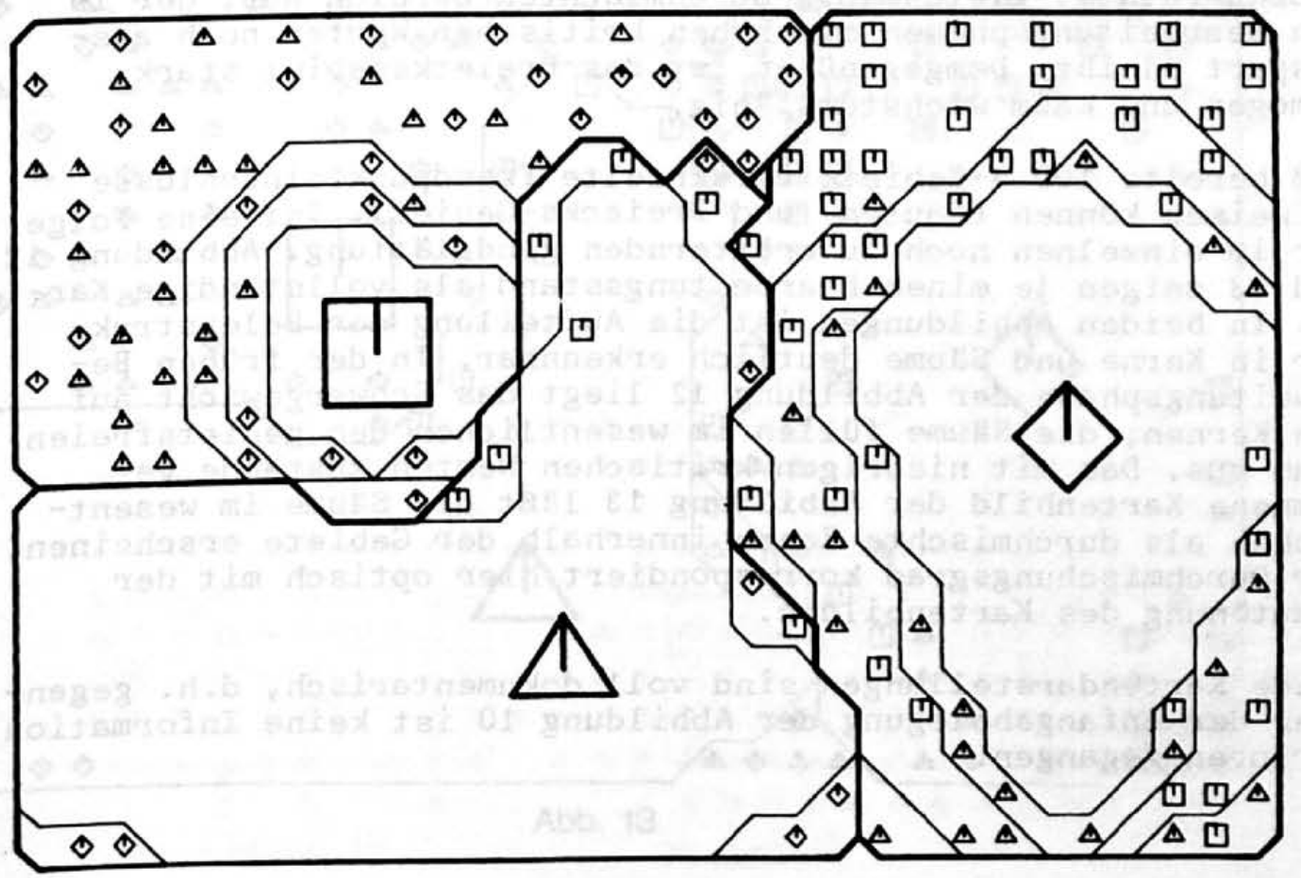


Abb. 11

Grad der Randglättung

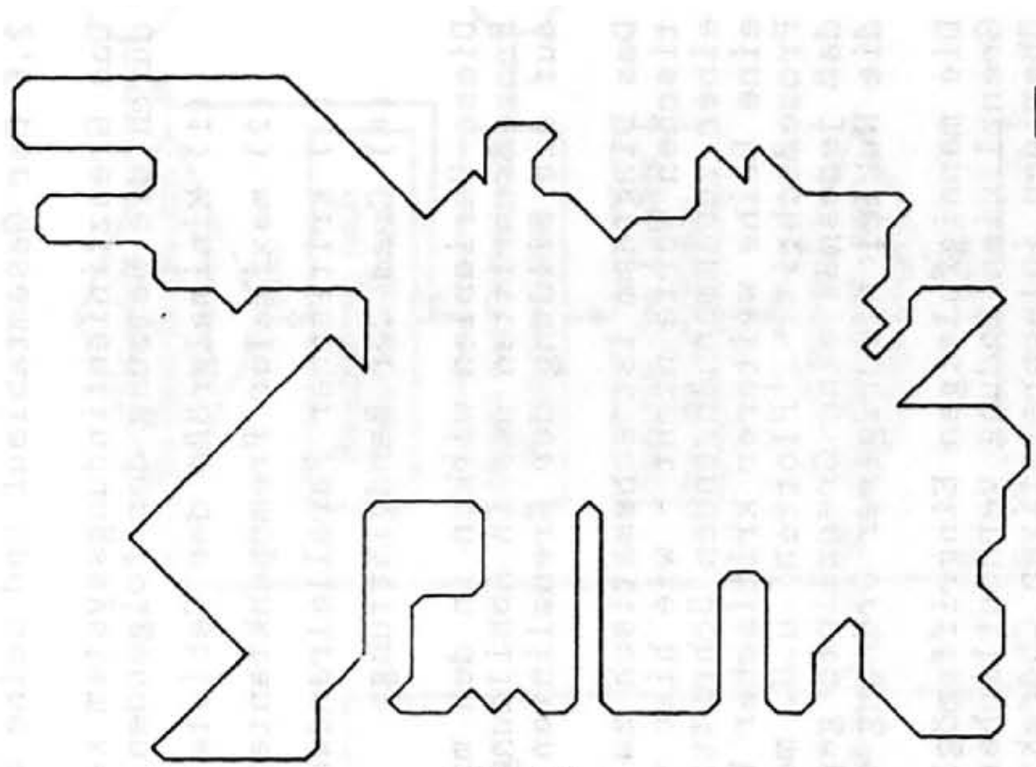


Abb. 18 a

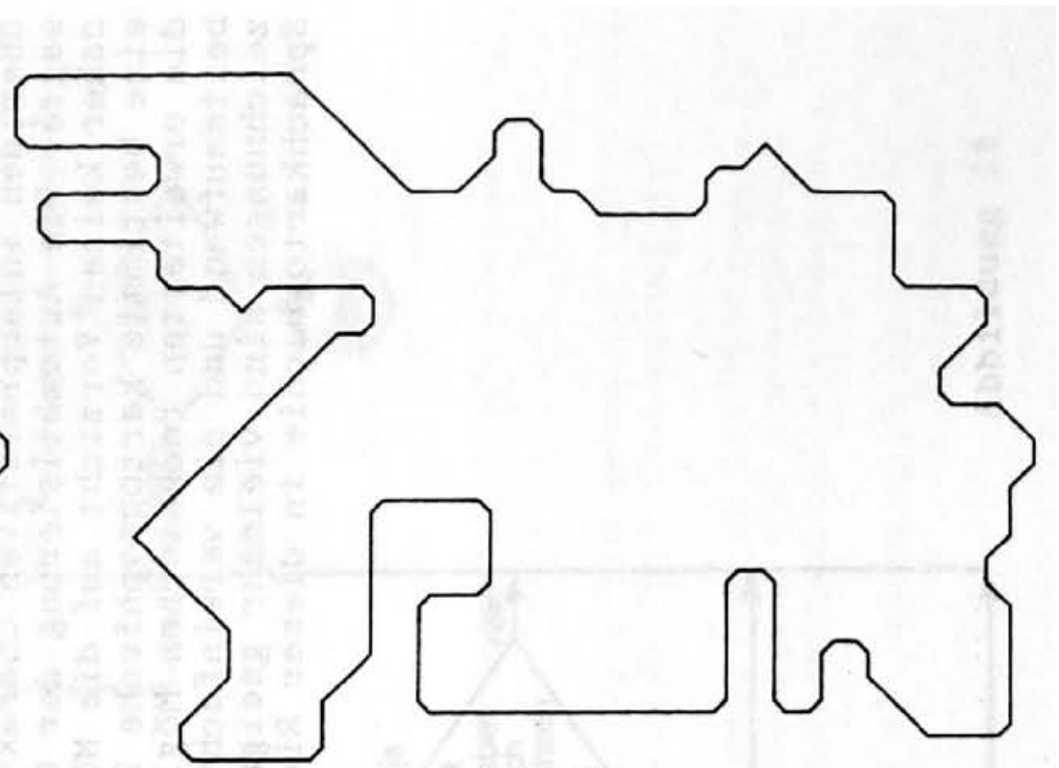


Abb. 18 b

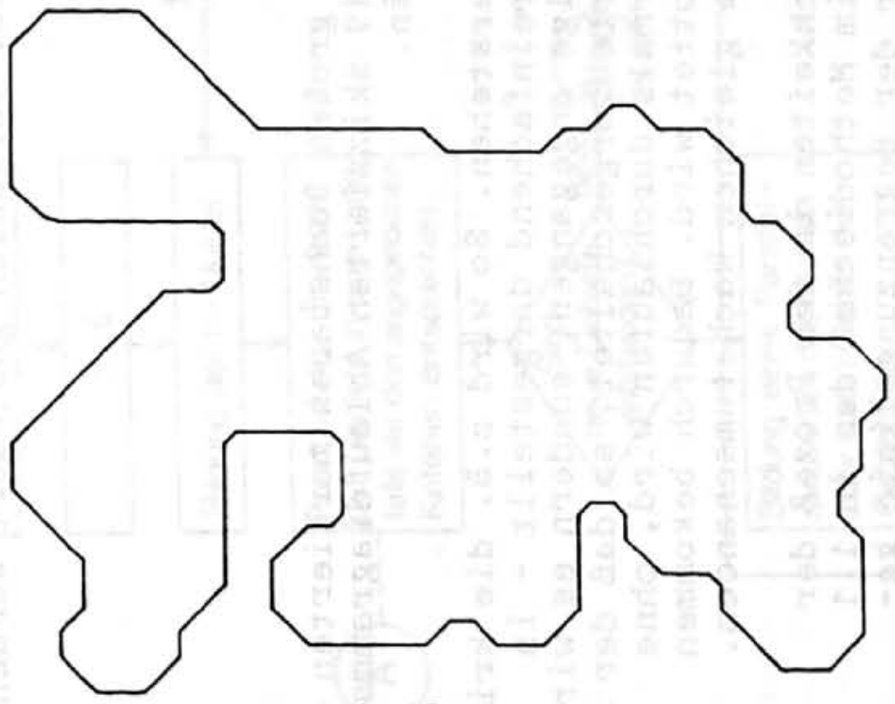


Abb. 18 c

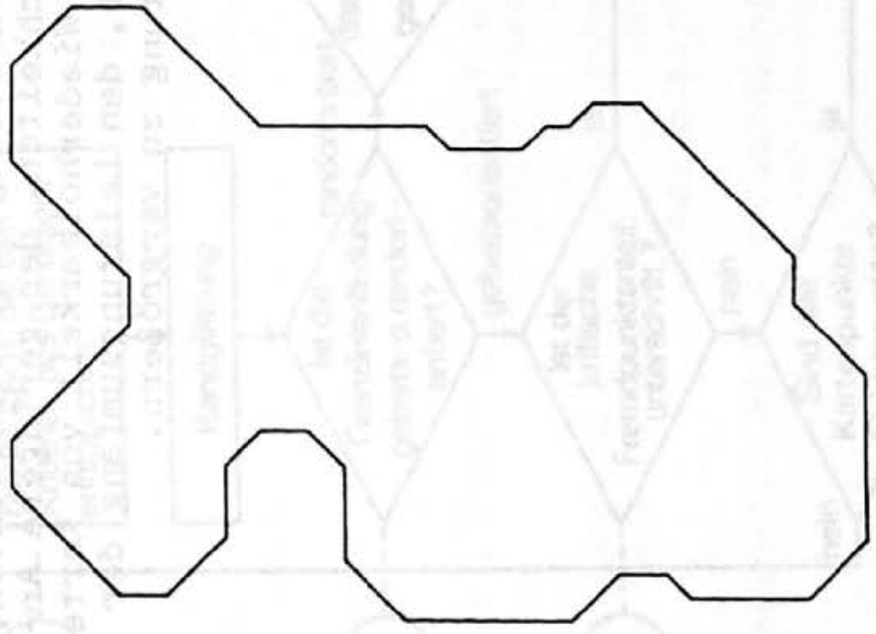
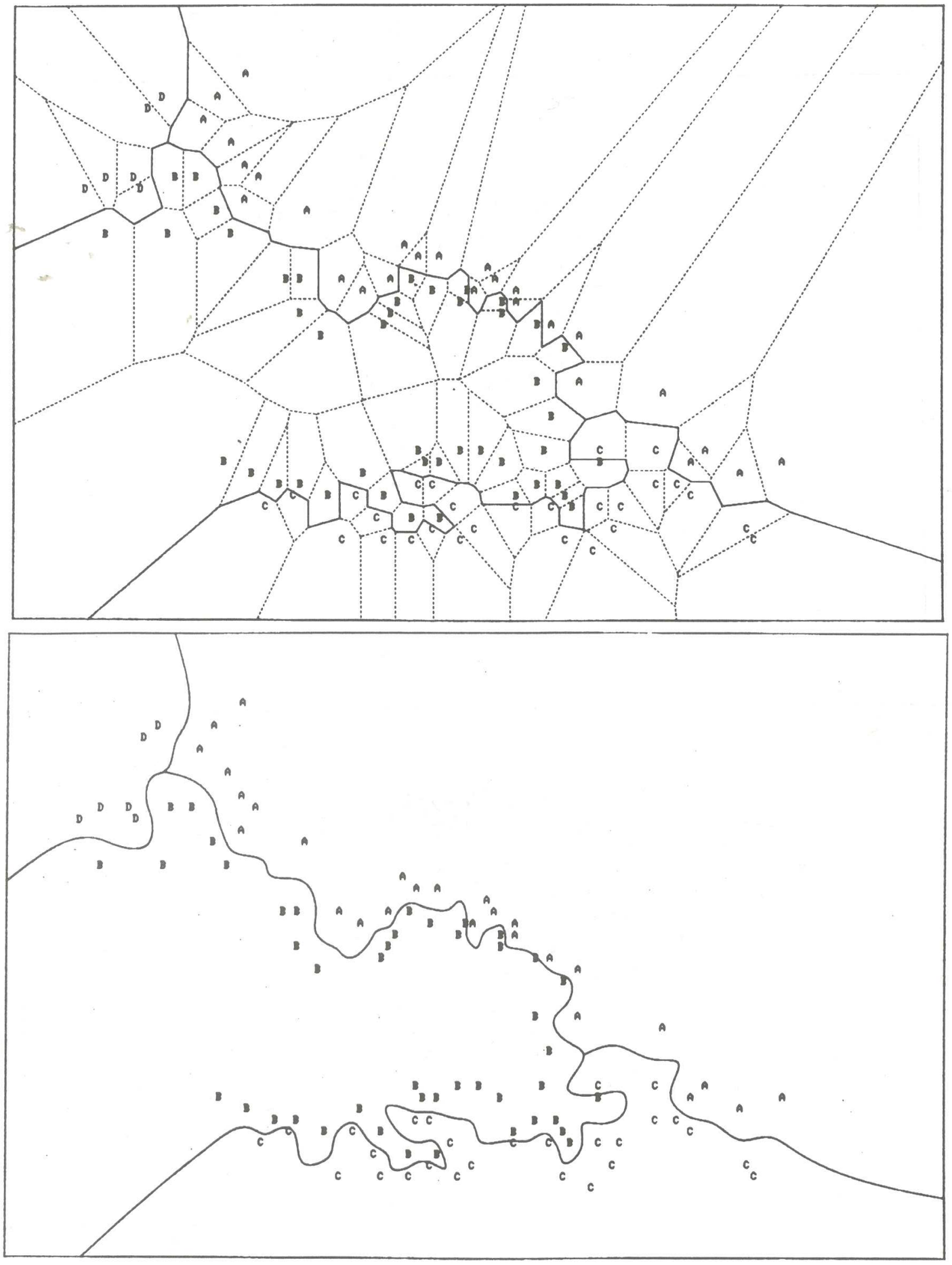
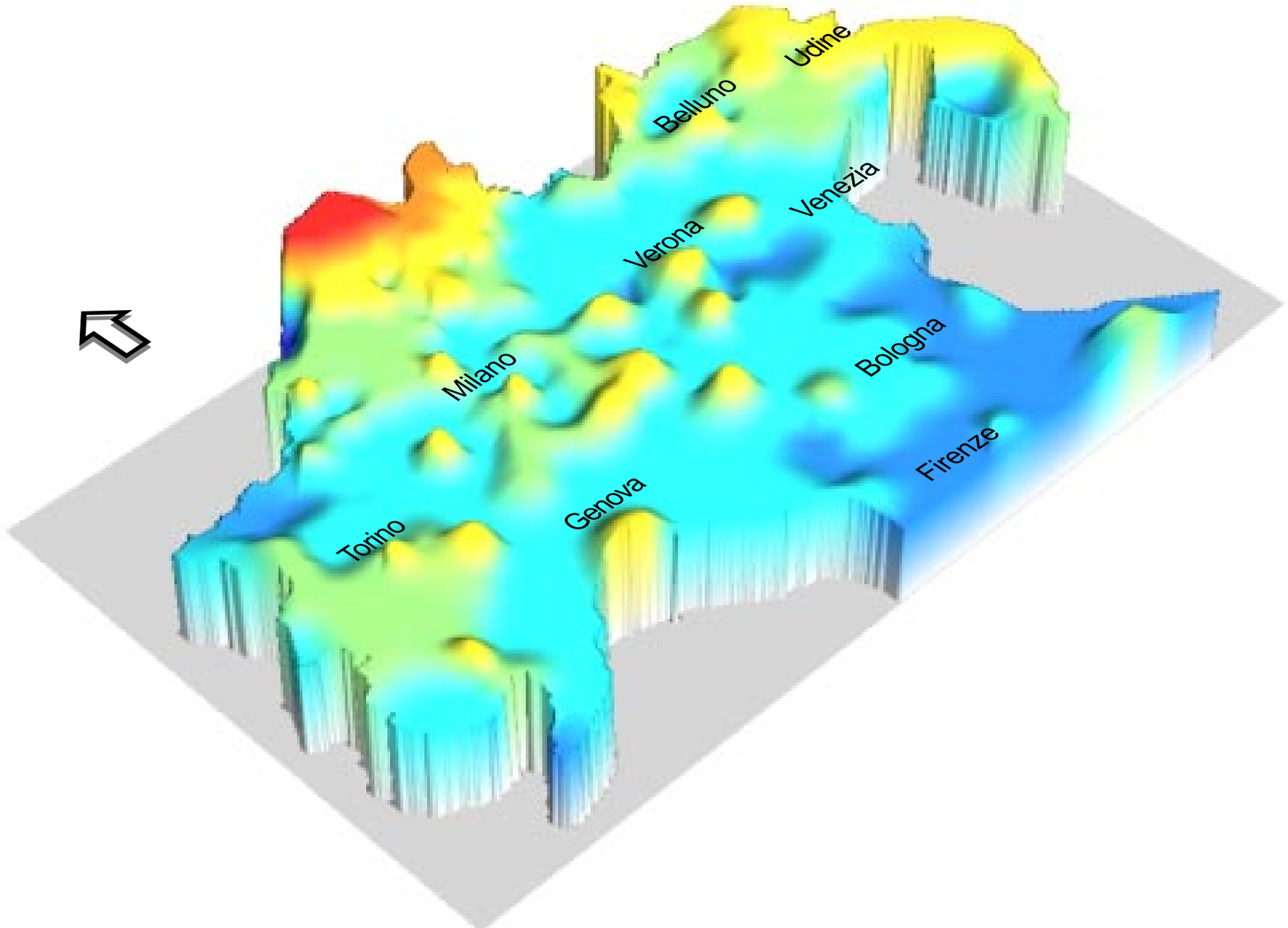


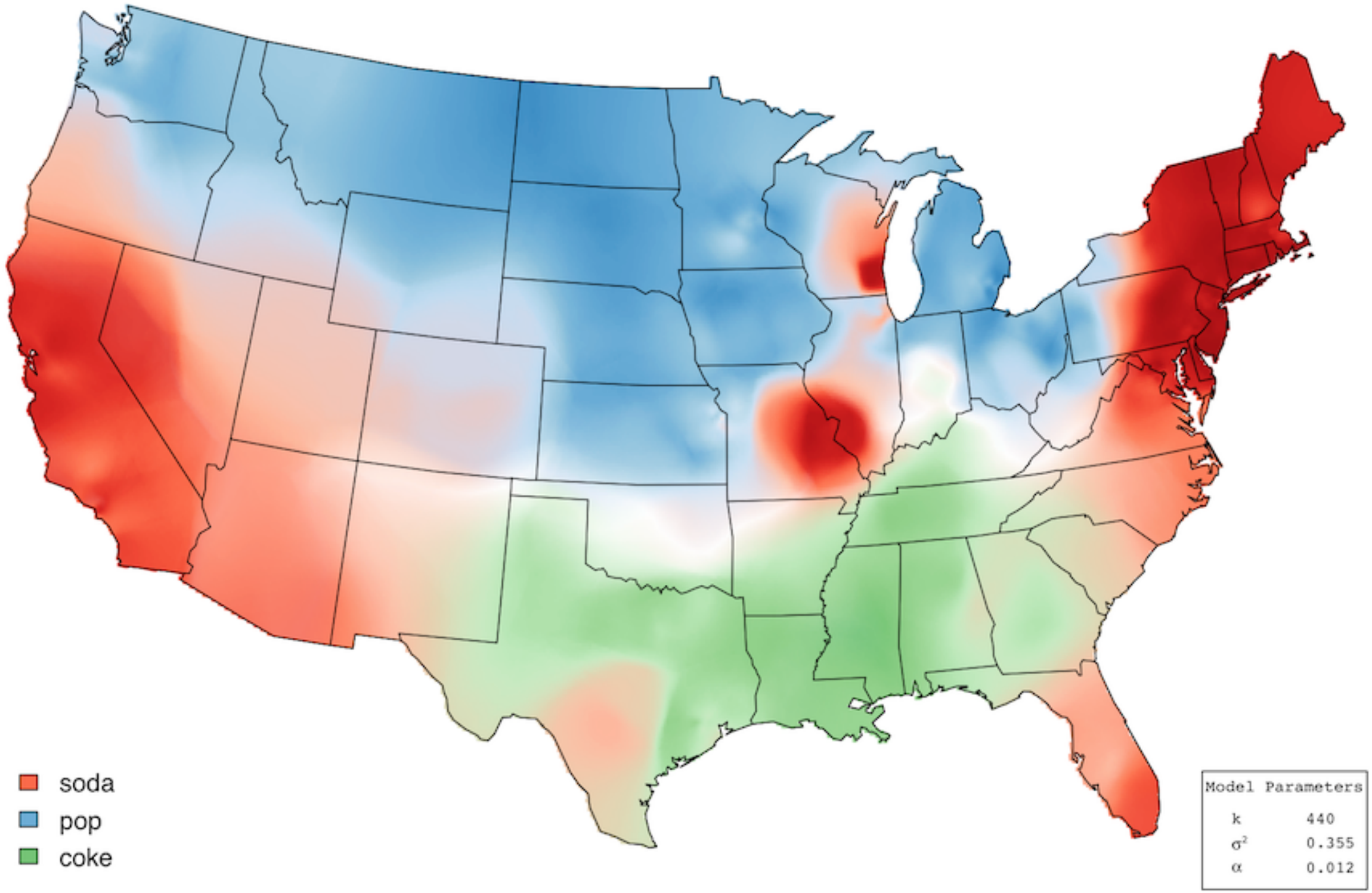
Abb. 18 d

Pudlatz, Hilmar. 1977. Automatische Erzeugung von Isoglossen auf dem Plotter mit Hilfe von Thiessen Polygone. *Germanistische Linguistik* 77(3-4), 245-258.



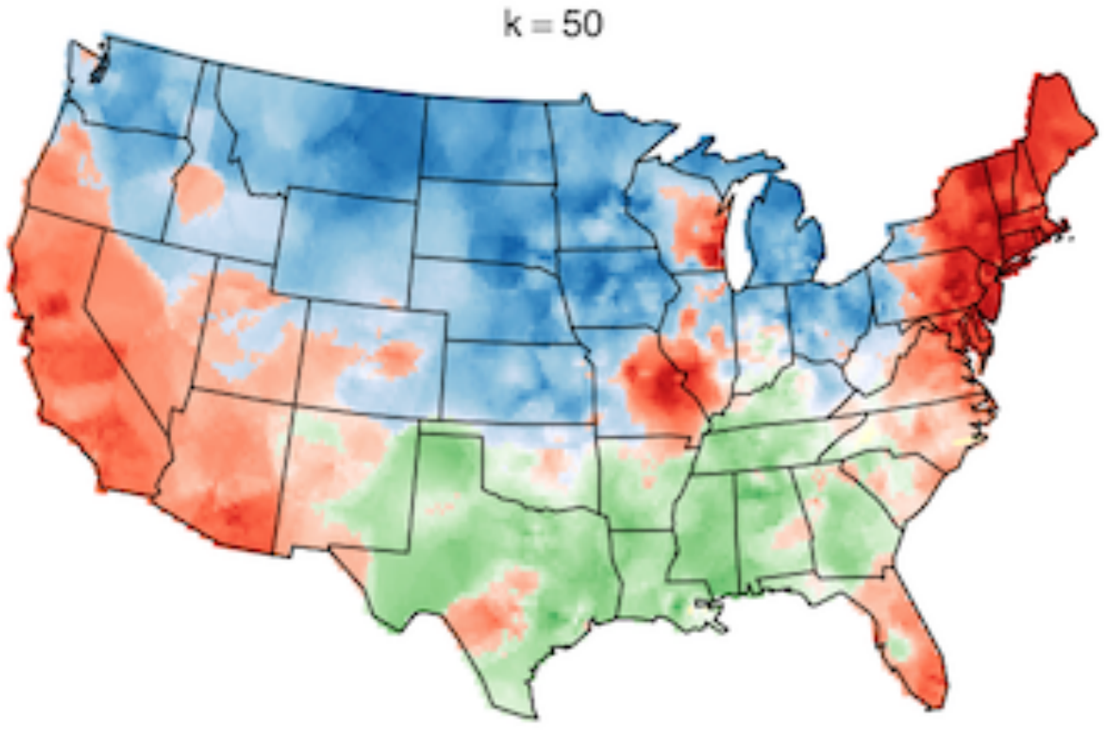


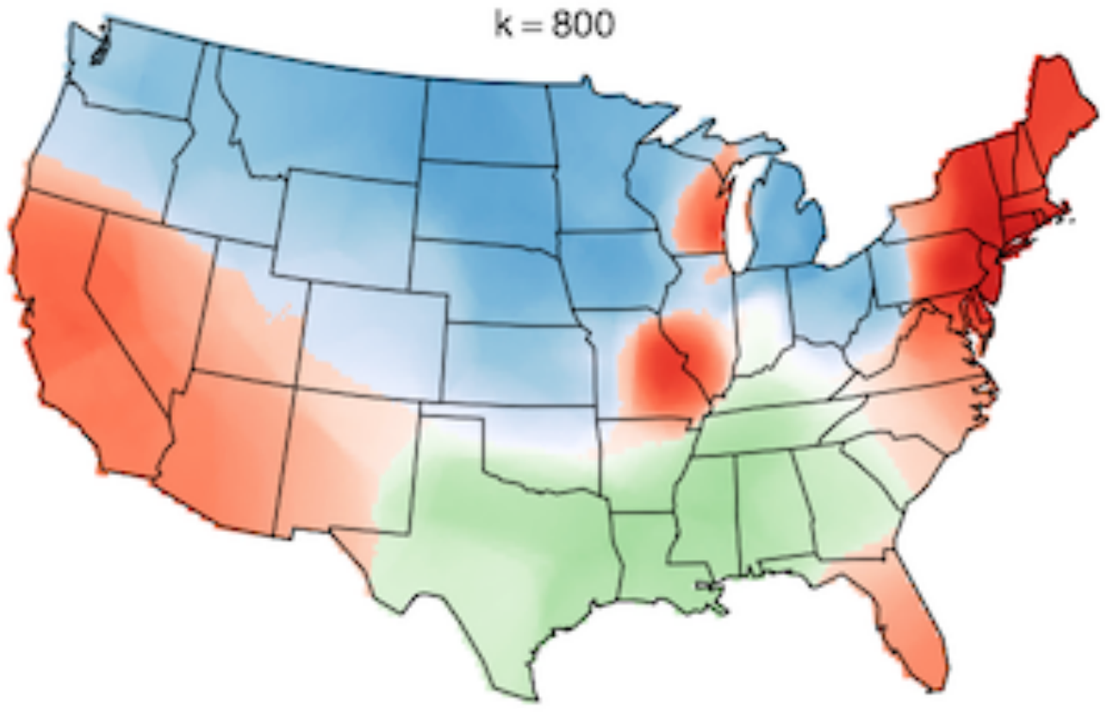
Hans Goebl (1982/2000/2012)



Katz, Joshua. 2013. Beyond "Soda, Pop or Coke".
Online at <http://www4.ncsu.edu/~jakatz2/files/dialectposter.png>

smoothing





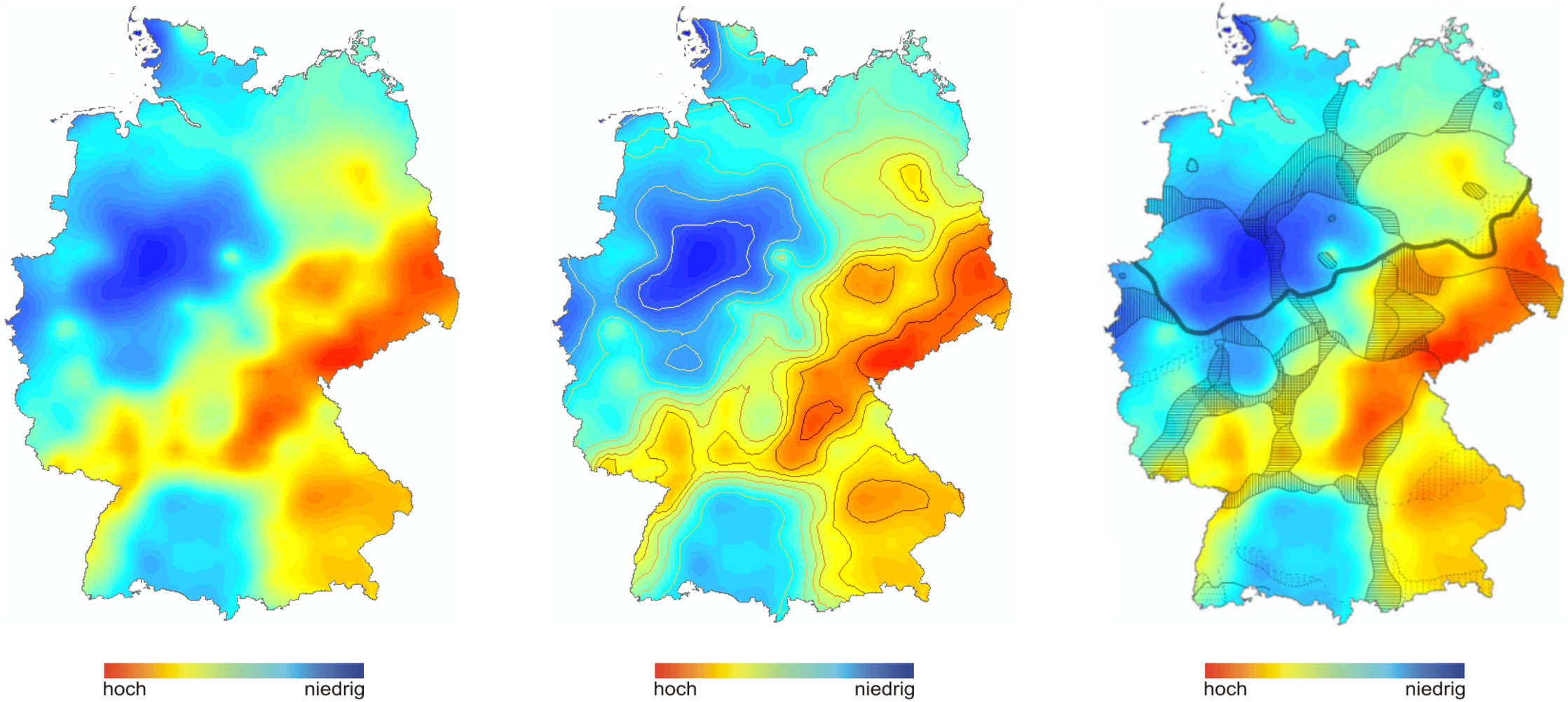
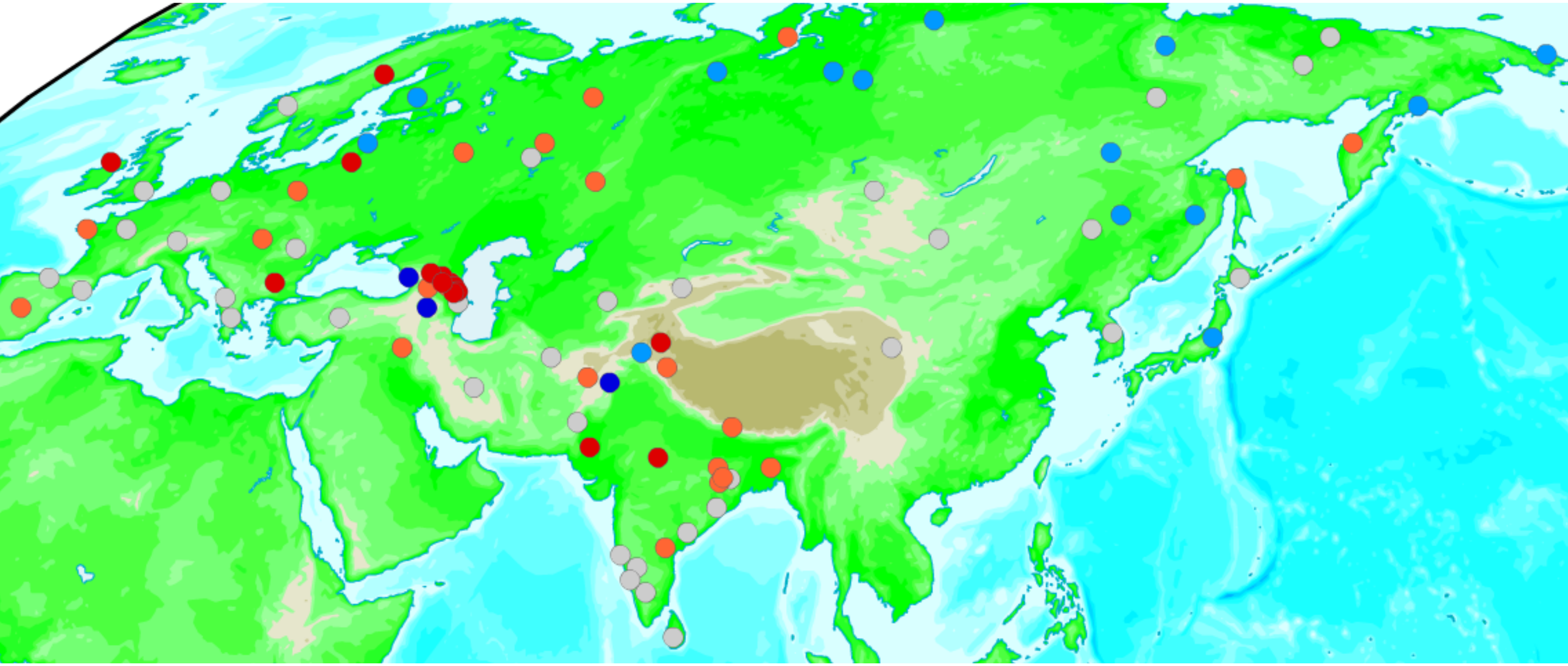


Abbildung 5-5: Geostatistische Interpolation der Similaritätswerte (Universal Kriging); links: Interpolationsergebnis; Mitte: definierte Konturlinien; rechts: Überblendung mit der Dialekteinteilungskarte von Wiesinger (1983a)

Lameli, Alfred. 2012. Strukturen im Kontinuum. Habilitationsschrift, Universität Marburg.

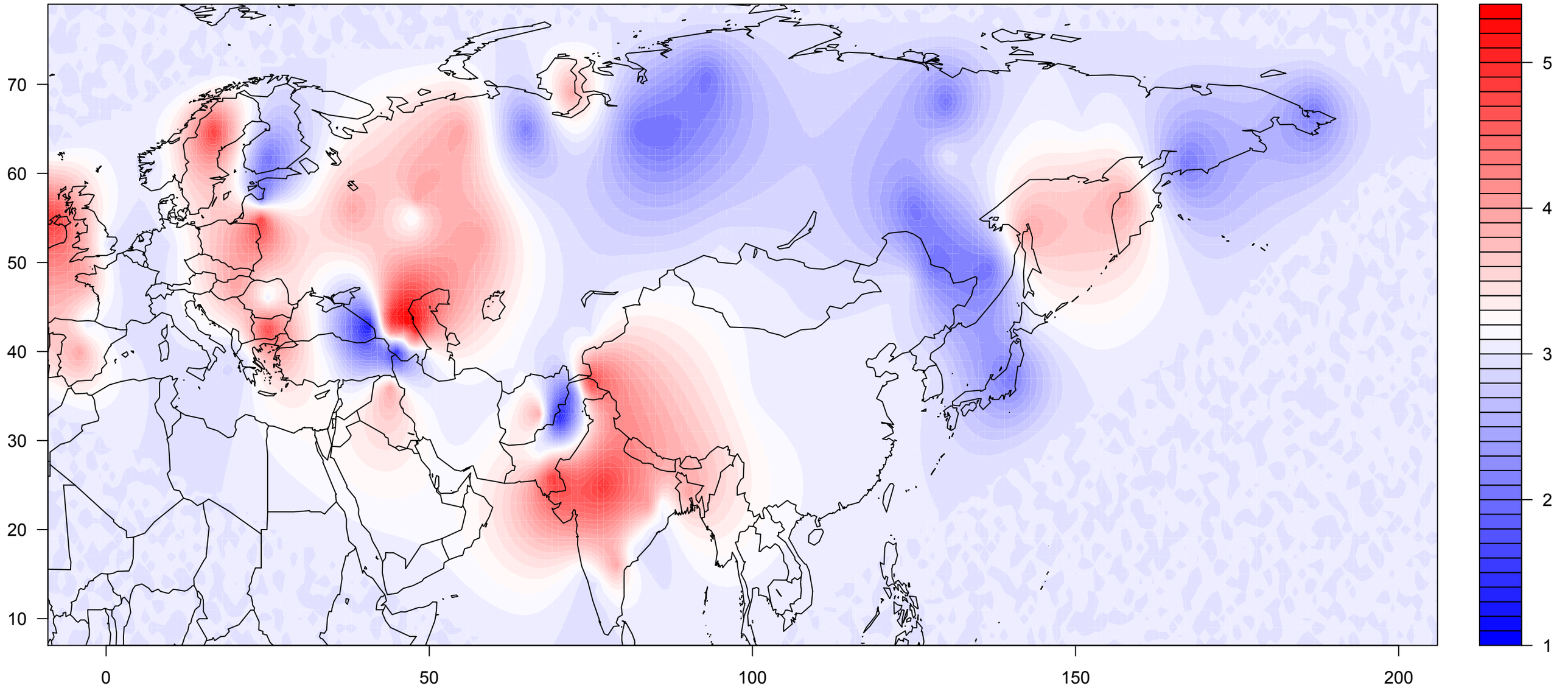
Consonant Inventories



few consonants

many consonants

Maddieson, Ian (2005) ‘Consonant Inventories’ in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press.


70
60
50
40
30
20
10
0
50
100
150
200
5
4
3
2
1

- Interpolation (Zwischenräume ausfüllen)
  - Glättung ist notwendig
  - Implizite “Vorhersage” über die Zwischenräume

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
Astfeld/ Gandersheim	I/62	Angelika Braun	A
Planrechteck X 29	Aufnahmedatum 20.11.1965	Transkribiert von 14.6.1985 bis 24.7.1985	

Besprochen von 24.07.1985 25.07.1985 L/SK

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪnt^{h}əʁ	ʁ = kontinuant Oberzb velarisiert
2	fliegen	56	ˈflaɪ̯ɐt^{h}ə	'fliegen die', Sequenzierung unklar - kein gemeintes [L]
3	Blätter	23	ˈblɛ̝ːdəʁ̥	
4	Luft	103	lʊˈft^{h}	ʁ = kontinuant
5	hört	89	hɪʁt	ʁ = kontinuant
6	gleich	78	kl$^{ɛ}_{+}$ɪk	folgt P
7	schneien	130	ˈʃnɛɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʁʊŋə
9	tu	151	dɛʊ	

Phonetischer Atlas von Deutschland

- Wenker-sentences recorded in the 1960s (with additions in the 1970s)
- Selected words from the recordings were transcribed on paper in the 1980s
- A joint project between Marburg and Groningen digitised the data in the 2000s
- In total 29530 words distributed over 183 locations and 186 cognate sets

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
Astfeld/ Gandersheim	I/62	Angelika Braun	A
Planrechteck X 29	Aufnahmedatum 20.11.1965	Transkribiert von 14.6.1985 bis 24.7.1985	

Besprochen von 24.07.1985 25.07.1985 L/SK

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪntʰəɐ̯	ɐ̯ = kontinuant Übergang velarisiert
2	fliegen	56	ˈflaɪ̯ɔtʰə	'fliegen die', Sequentierung unklar - kein gemeintes [L]
3	Blätter	23	ˈblɛːdəɐ̯̊	
4	Luft	103	lʊˈftʰ‿	ɐ̯ = kontinuant
5	hört	89	hɪɐ̯t‿	ɐ̯ = kontinuant
6	gleich	78	‿klɛɪ̯ç‿	folgt P
7	schneien	130	ˈʃnɛɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʀʊŋə
9	tu	151	dɛʊ	

Phonetischer Atlas von Deutschland

- Digitised in X-SAMPA, converted back to match original transcriptions, minor corrections for consistency of encoding
- The data is transcribed in high phonetic detail (3786 different phonetic segments)
- We will make the complete data available
  - electronically, separated by phonetic segments
  - as close as possible to the original source
  - including all idiosyncrasies

Multiple Sequence Alignment

- Just a fancy name for sound correspondences
- Each sound correspondence is "aligned" in a column, possibly adding empty cells
- It is a useful and consistent way to represent comparative data (both between languages or dialects)

LOCATION	WORD
Aachen	a̠ːph
Adorf	a̬ːbʱə
Ahrbergen	o̬→ɔ̬phə̬
Albersloh	ɑːpᵍ̈ə̈
Allna	ɑɸh
Altenberg	ʌfɛ̈
Altentrüdin	a̠f
Altlandsberg	ɑʾfəʾ
Altwarp	oːph
Astfeld	pˤːpʰə
Atzendorf	afɛ̠
Ballhausen	ʌʾfə
Bardenfleth	ɔːp͡ɸ
Barssel	ɒːpʲə
Bempflingen	afː
Bennin	ɔ̬pʰ
Billingsbach	a̠f
Bockelwitz	ʌ̬və̈
Bonn	ɑːpʼ
Borstendorf	ɤ̬fː
Breddin	ɒːph
Brelingen	ɑ̬f͡βə
Bremscheid	pˤːpḫə̬
…	…

A	FF	E
a̠ː	ph	-
a̬ː	bʱ	ə
o̬→ɔ̬	ph	ə̬
ɑː	pᵍ̈	ə̈
ɑ	ɸh	-
ʌ	f	ɛ̈
a̠	f	-
ɑʾ	f	əʾ
oː	ph	-
pˤː	pʰ	ə
a	f	ɛ̠
ʌʾ	f	ə
ɔ̬ː	p͡ɸ	-
ɒː	pʲ	ə
a	fː	-
ɔ̬	pʰ	-
a̠	f	-
ʌ̬	v	ə̈
ɑː	pʼ	-
ɤ̬	fː	-
ɒː	ph	-
ɑ̬	f͡β	ə
pˤː	pḫ	ə̬
…	…	…

- **Workflow:**
  - Tokenisation of segments
  - Automatic alignment using LingPy (github.com/lingpy)
  - Manual correction
  - Separation of cognates (e.g. *Samstag* vs. *Sonnabend*)
  - Annotation of columns (e.g. many-to-one alignments, metathesis)
  - Merging of complex columns and removing boundaries

Automatic Isoglossing

- Each column represents an isogloss
- Plotting one column gives a dialect map (after some interpretation of the data)
- Areas can be added by using 3D-interpolation (or any other of the many methods being developed currently)

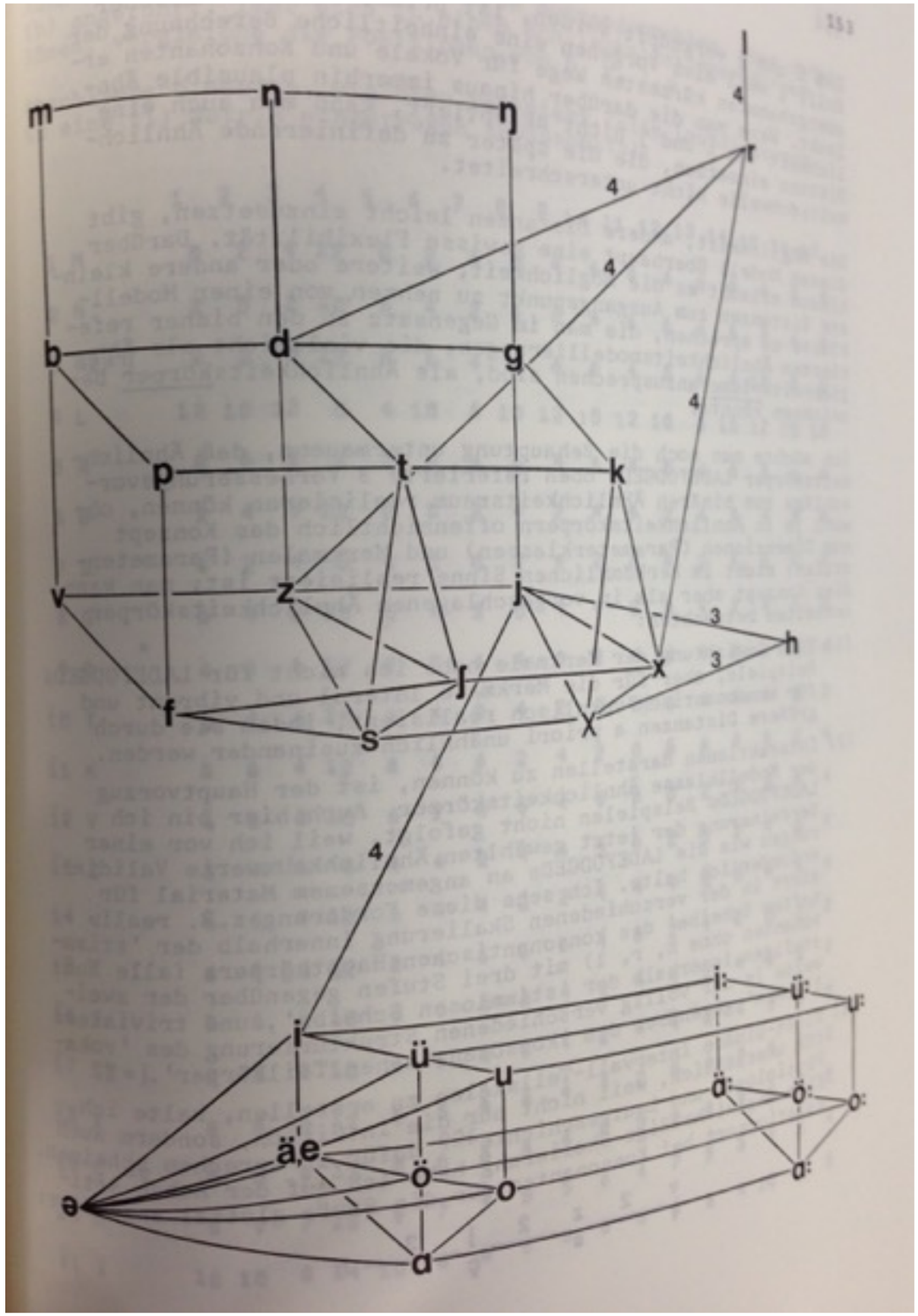
Example 1

- **Consonant** in German Perfect-prefix ge-
- Ordered to ‘strength’, visualised as ‘height’
- (NULL) ʏ̟ ɪ j̩ j j̲ ʔ c͡ɟ z̪ ʒ̪ ʒ ɕ ɣ̪ ç ɟ͡j ɣ ɟ̟ ç̲ ɣ̲ t͡ɕ g͡ɣ x͡ɣ g̪ x̪ c g k͡ç k͡j x ç̲ x̲ kʰ k͡x k̪ ʀ̥ χ k k̲ q k’

Character Model

- Needleman, SB & CD Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443-453.
- Widely used in bioinformatics ever since
- Multiple time reinvented in linguistics
- I will use a 'linear' character model, which is easier to visualise. For 'nominal' character models, multiple maps have to be combined

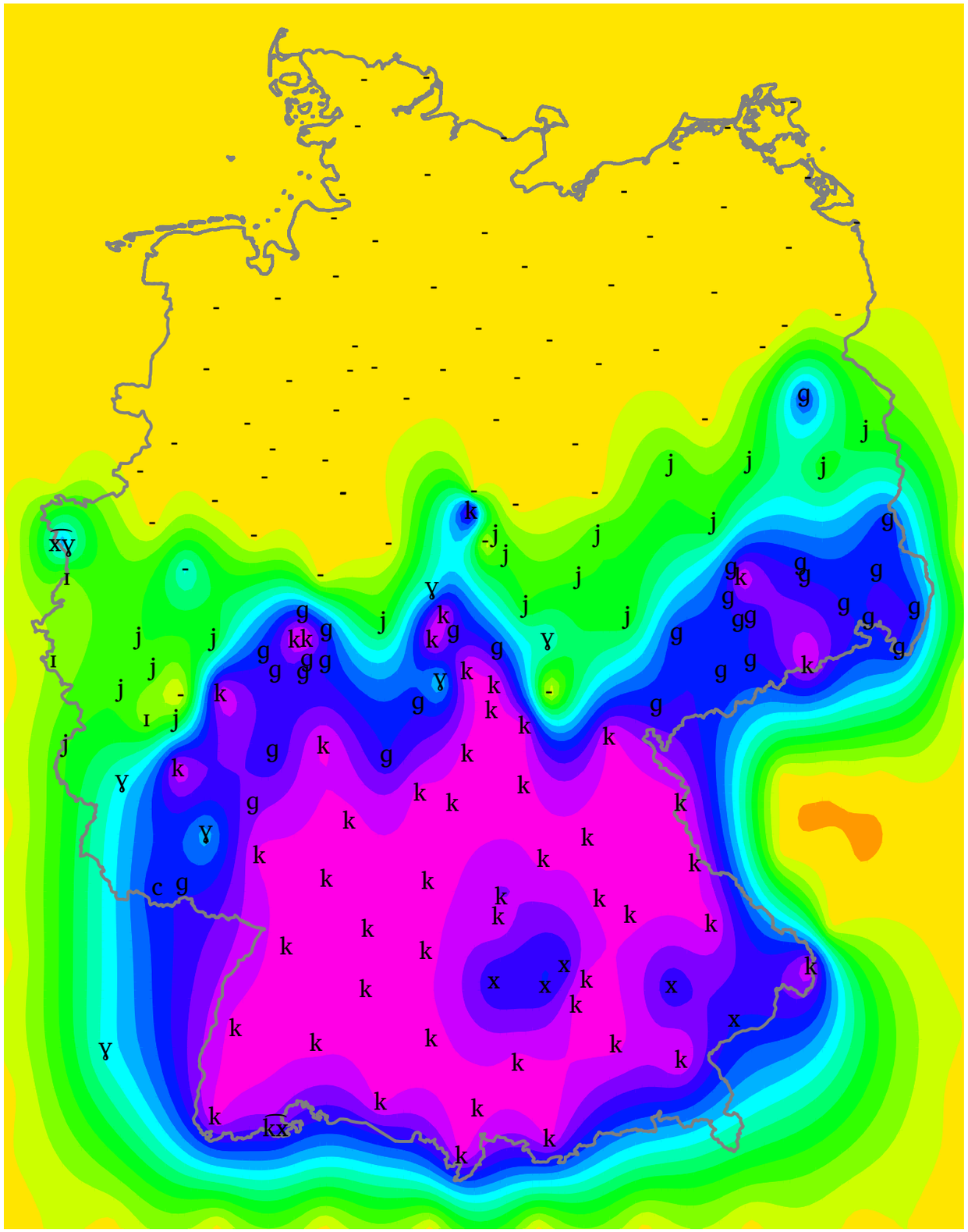
Naumann, Carl Ludwig. 1976. Grundzüge der Sprachkartographie und ihrer Automatisierung. *Germanistische Linguistik* 76(1-2). 1-285.



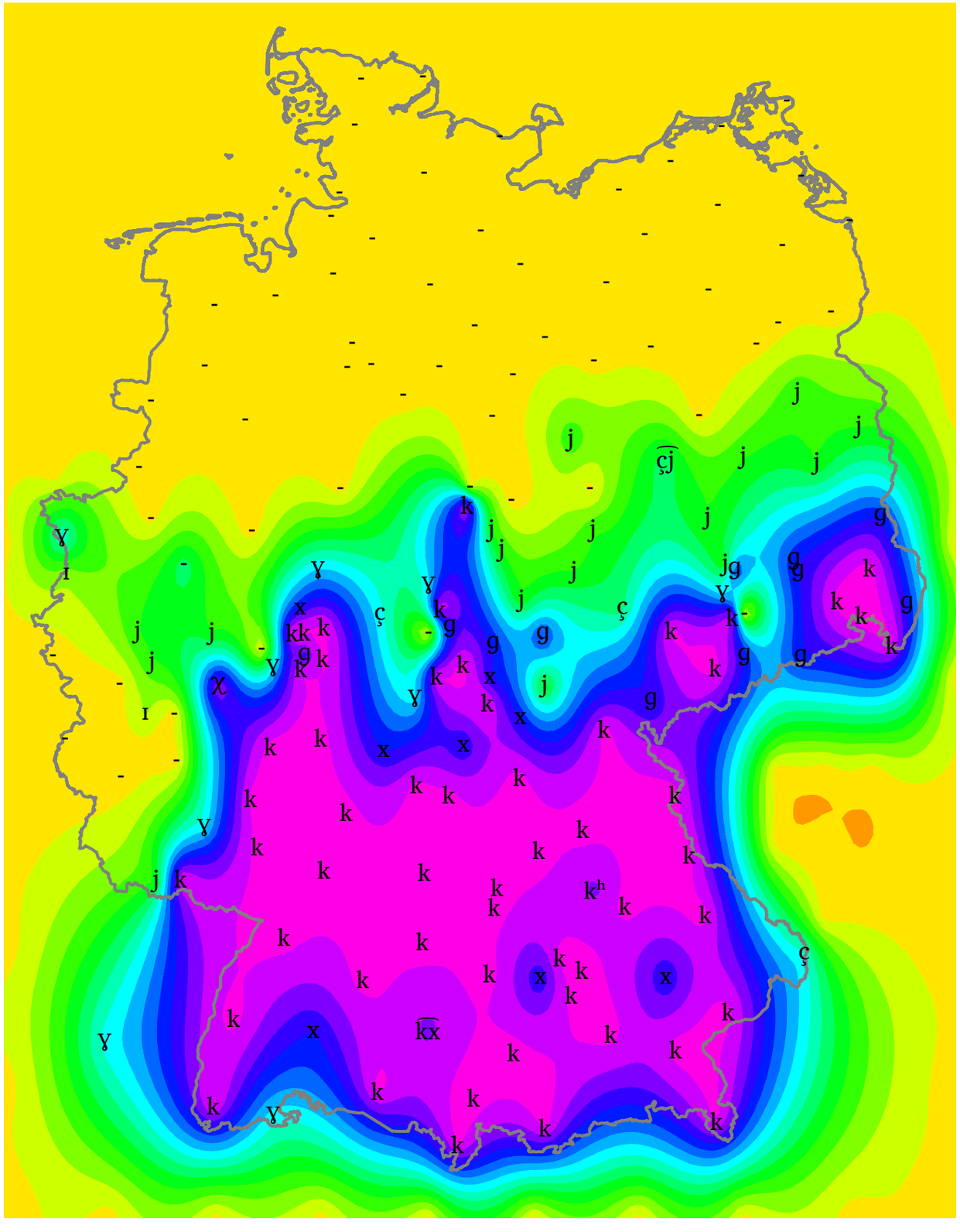
Example 1

- **Consonant** in German Perfect-prefix ge-
- Ordered to ‘strength’, visualised as ‘height’
- (NULL) ʏ̟ ɪ j̩ j j̲ ʔ c͡ʝ̟ z̩ ʑ̩ ʒ ɕ ɣ̩ ç̩ c͡j̟ ɣ ɟ ç̲ ɣ̲ t͡ɕ̩ g͡ɣ x͡ɣ̞ g̩ x̩ c g k͡ɕ̩ k͡j̟ x c̲ x̲ kʰ k͡x k̩ ʀ̥ χ k k̲ q k’

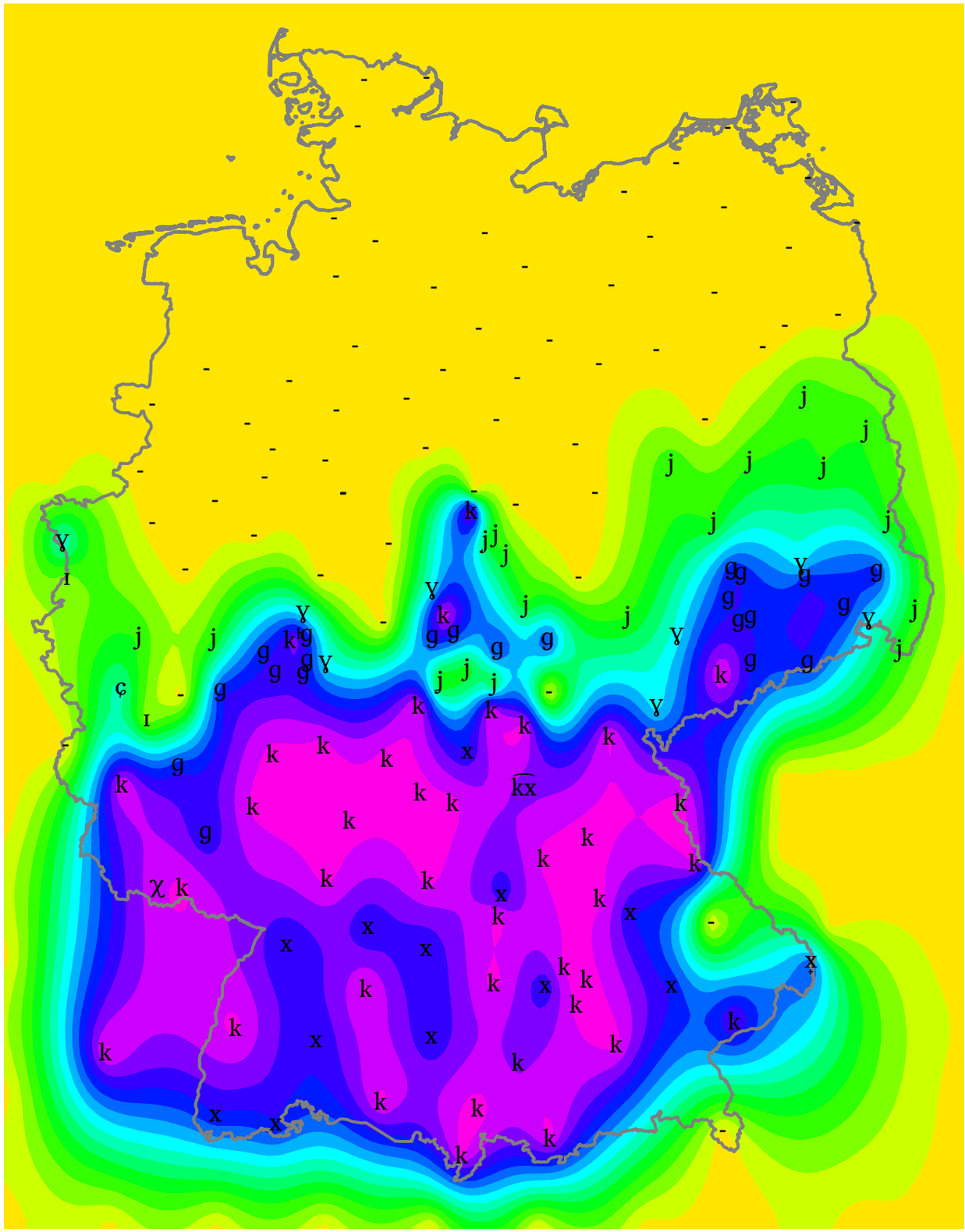
gefahren



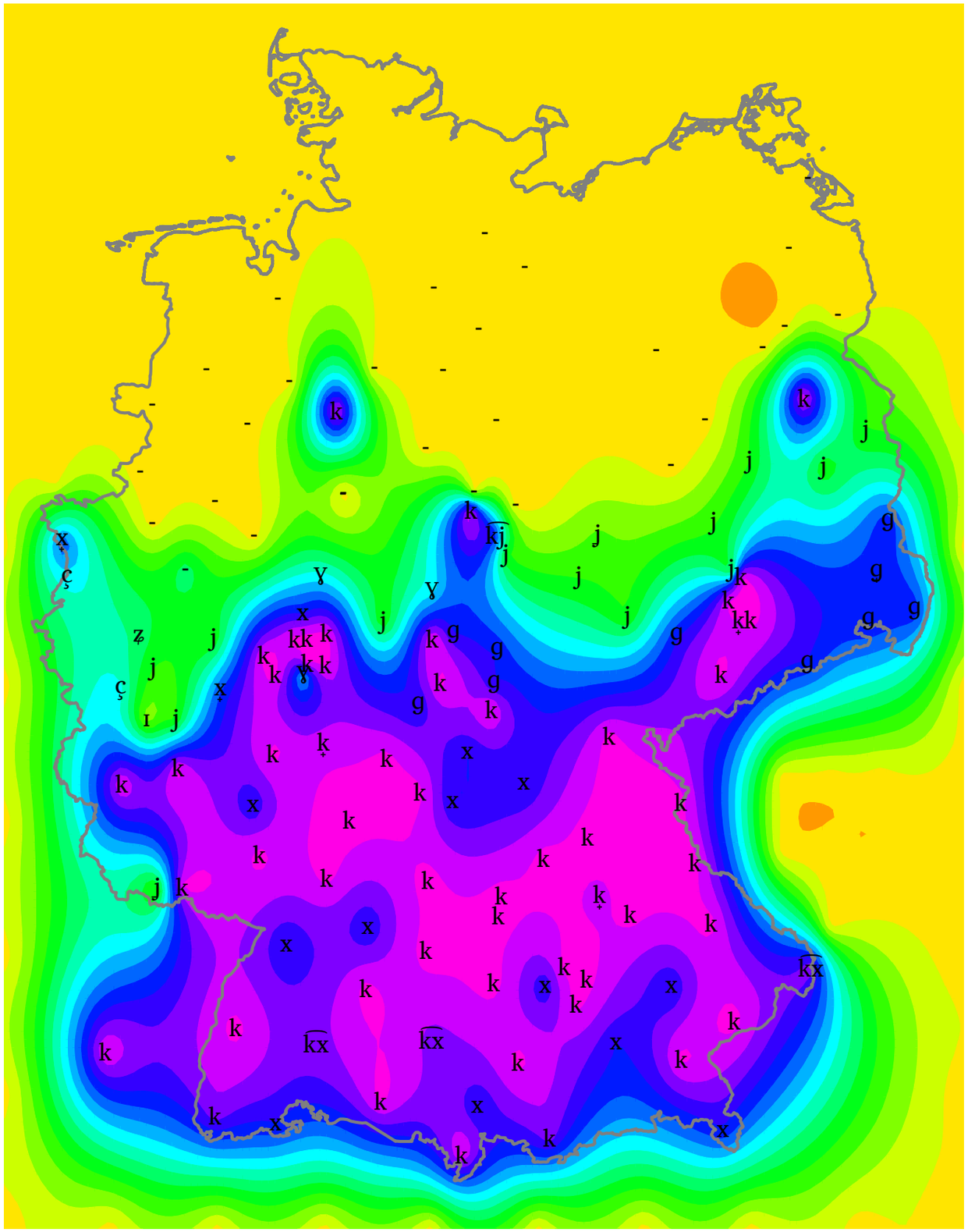
gefunden



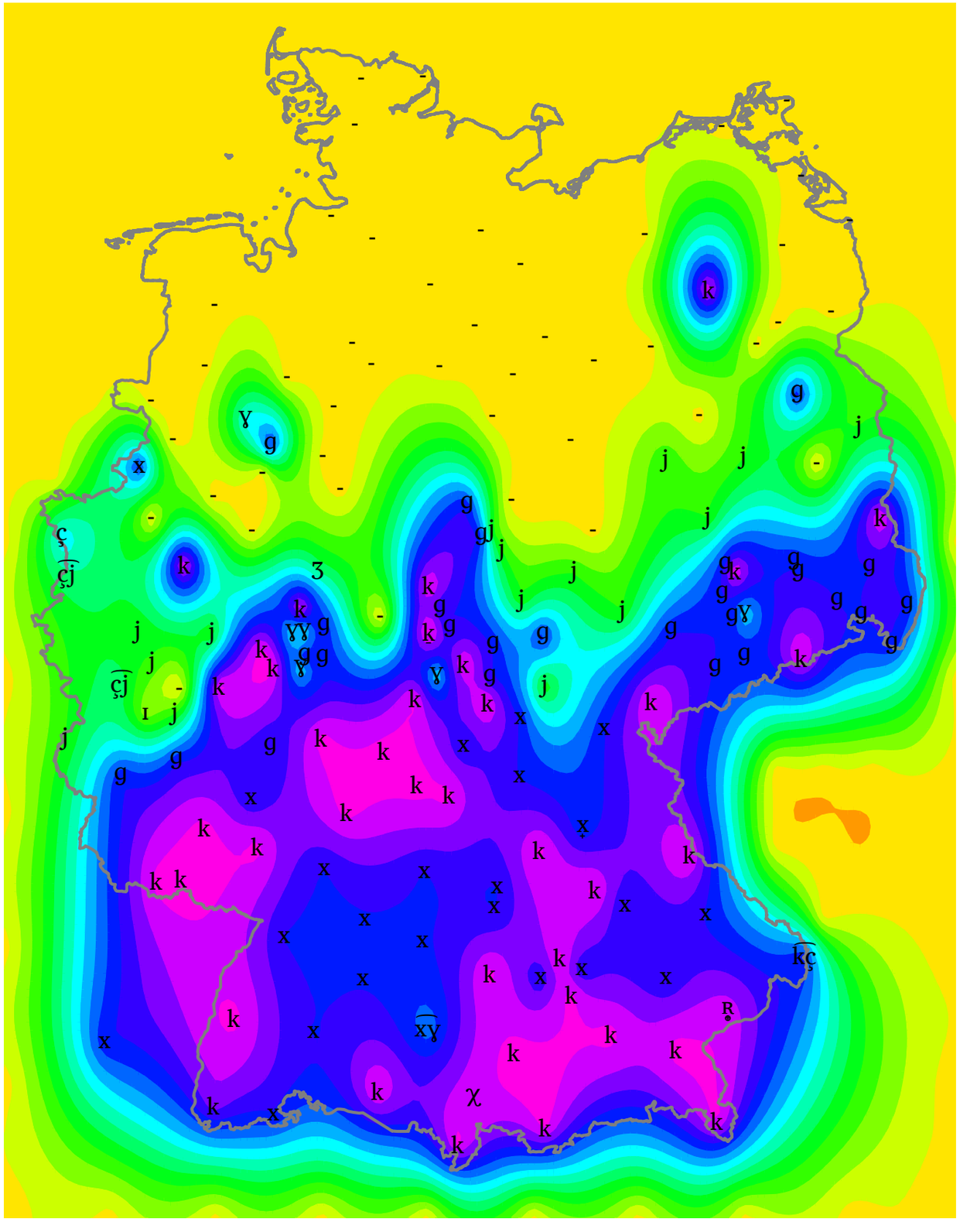
eingeschlafen



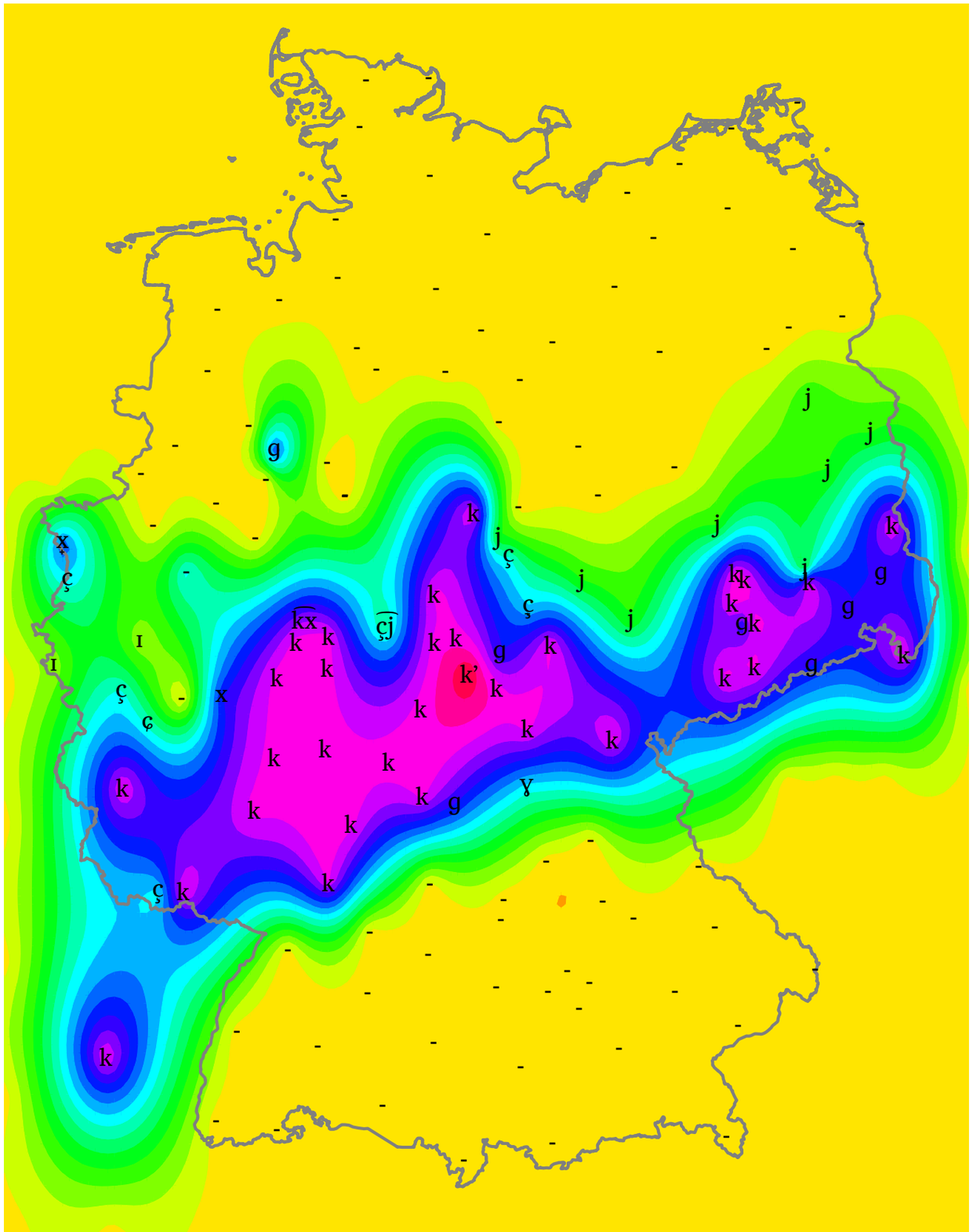
gestohlen



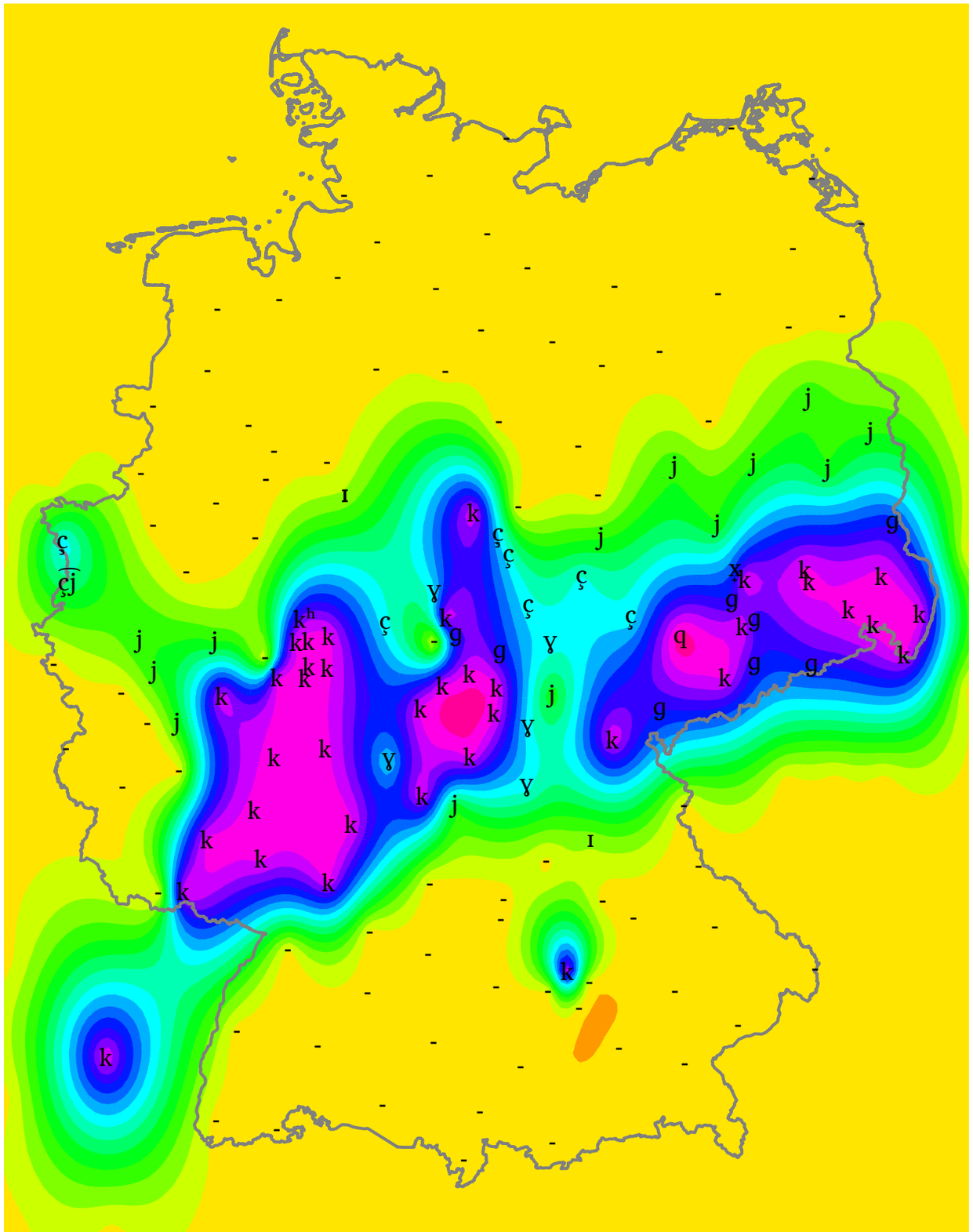
gestorben



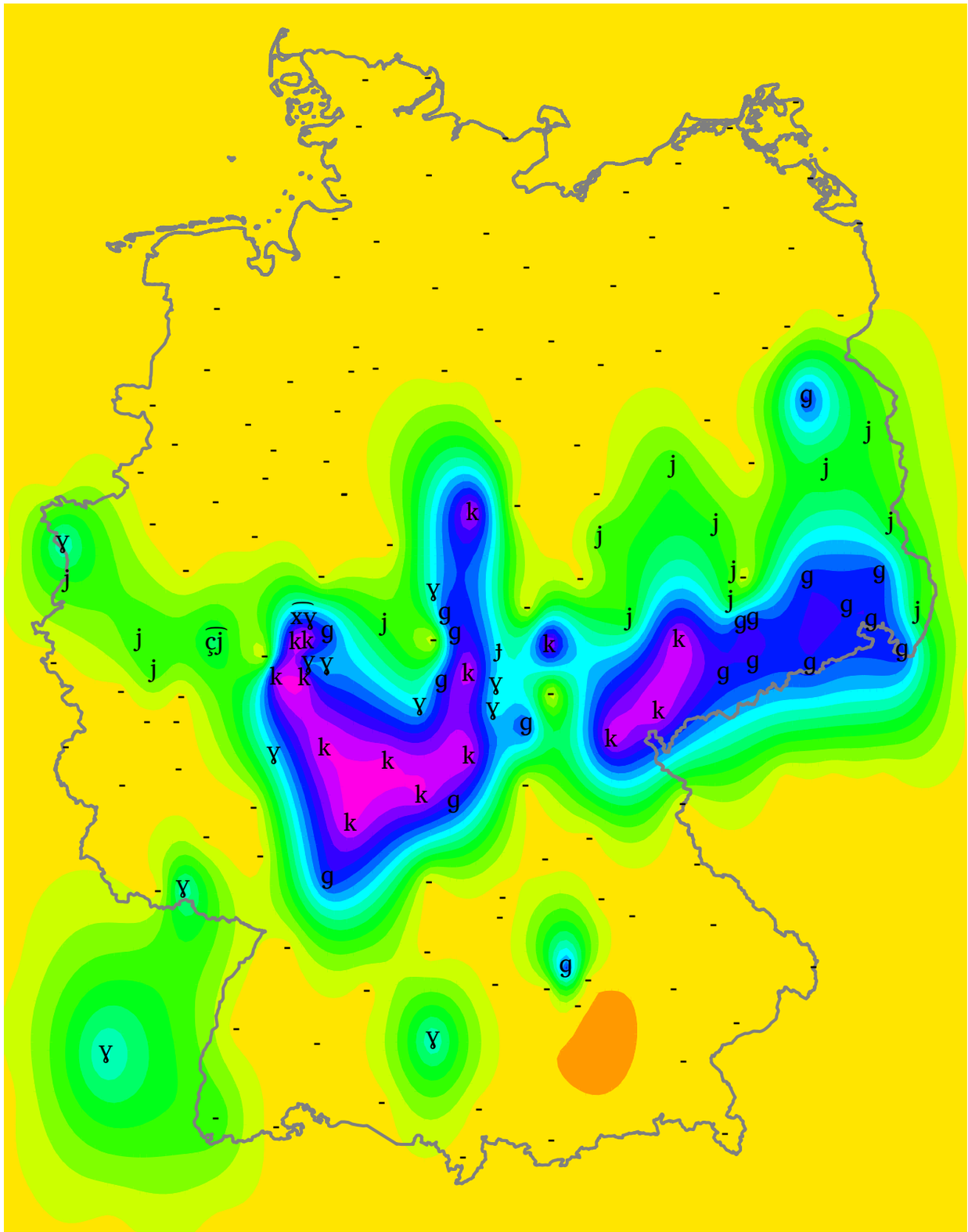
gebrannt



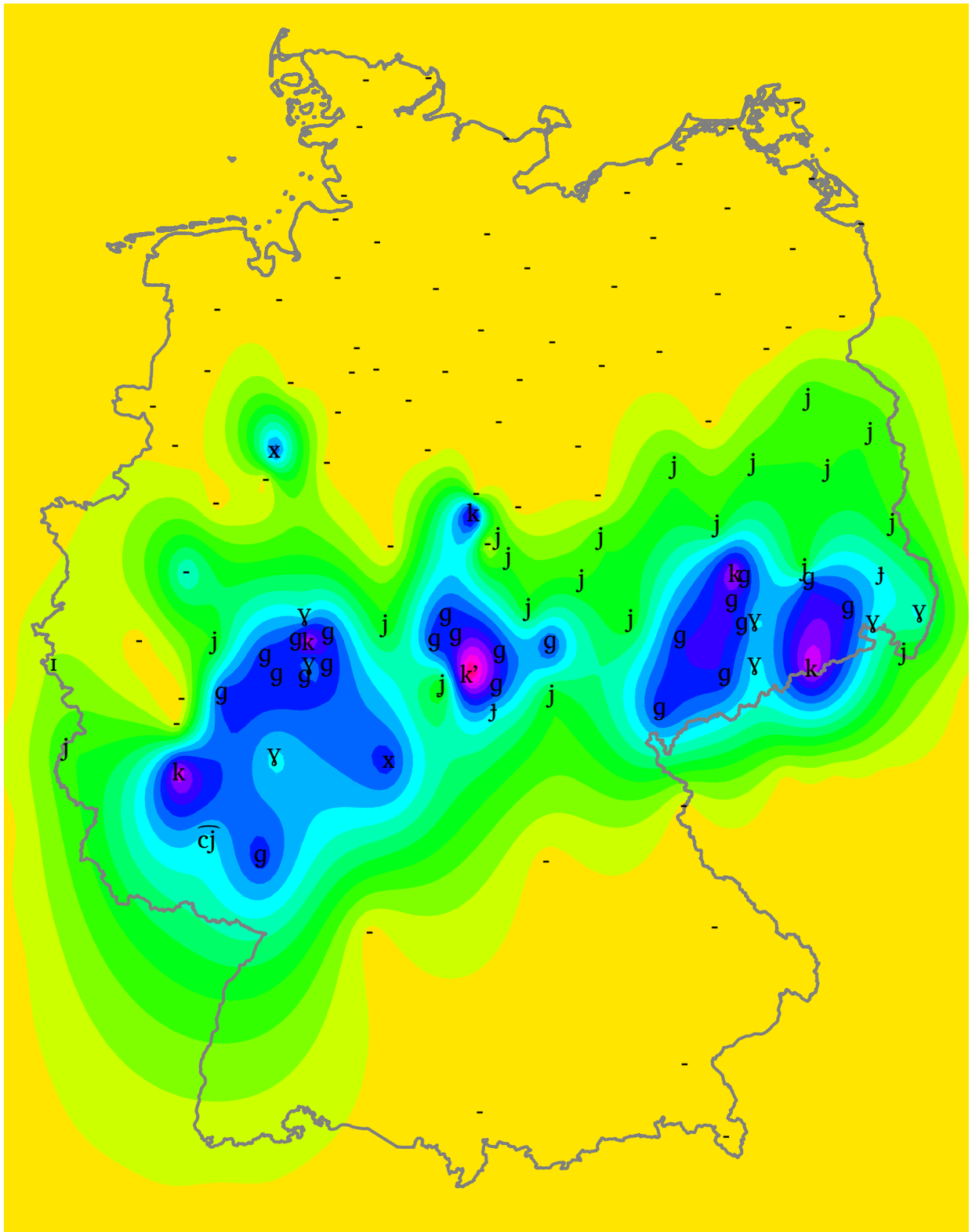
gebracht



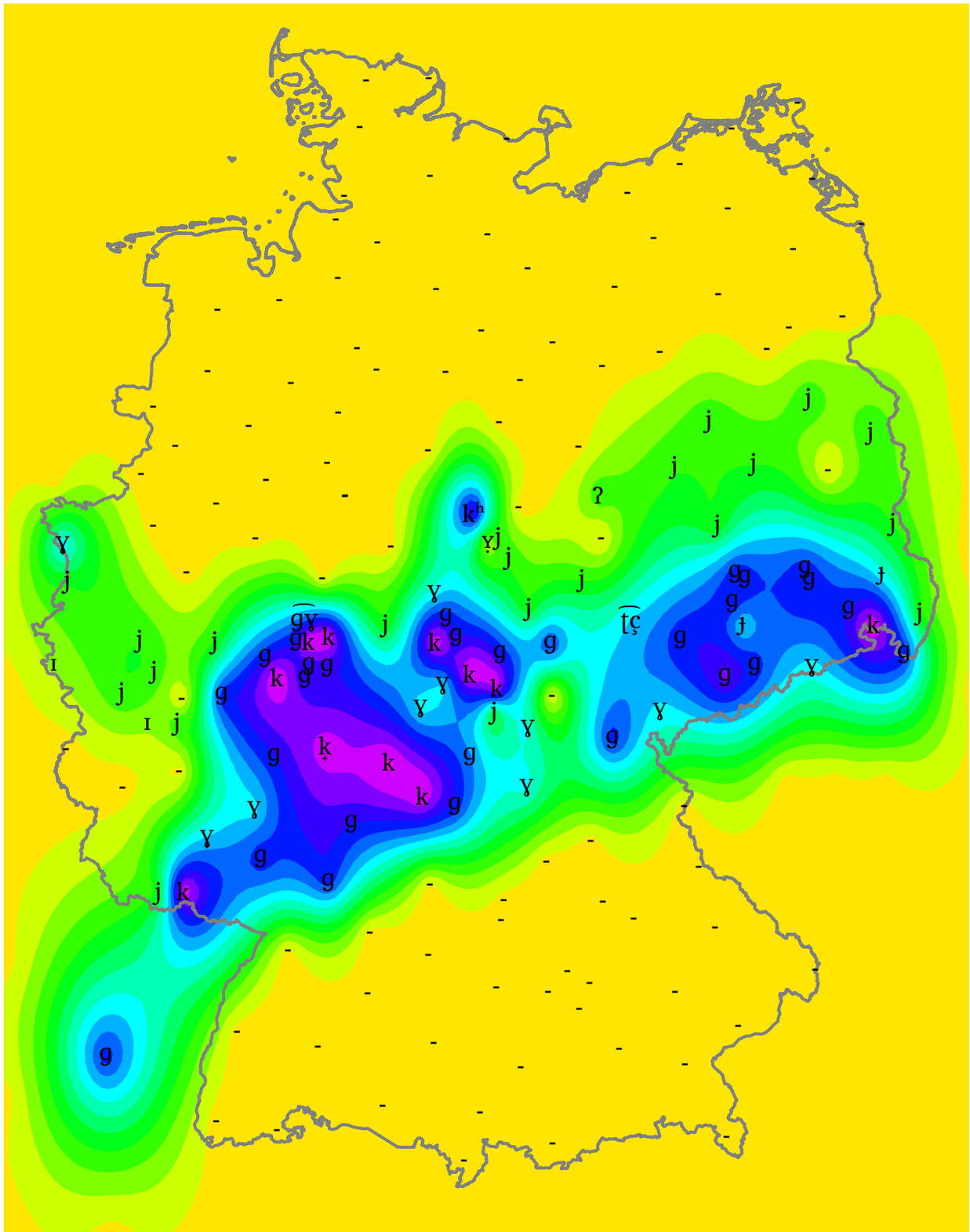
geblieben



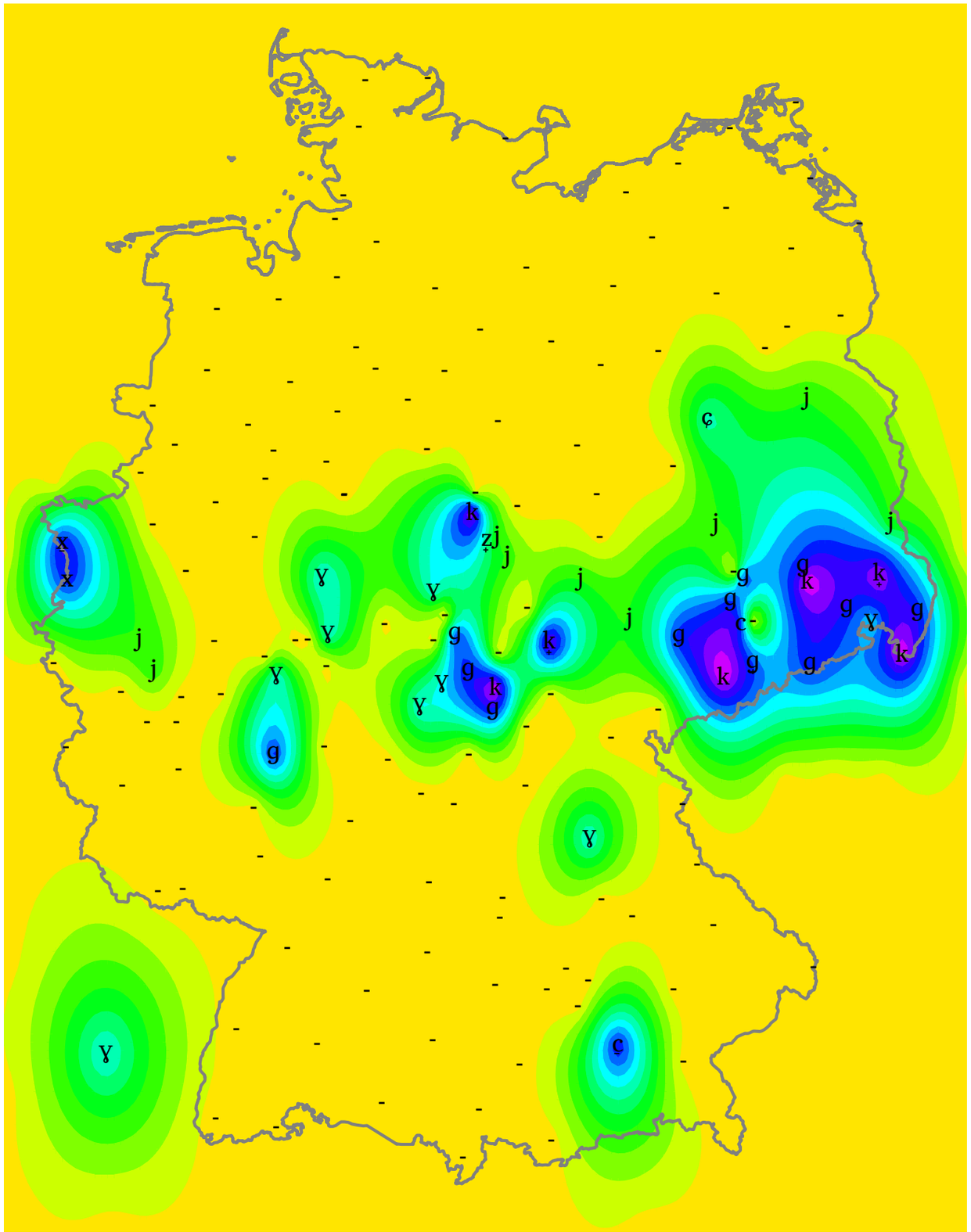
eingebrochen



gekannt



gekommen

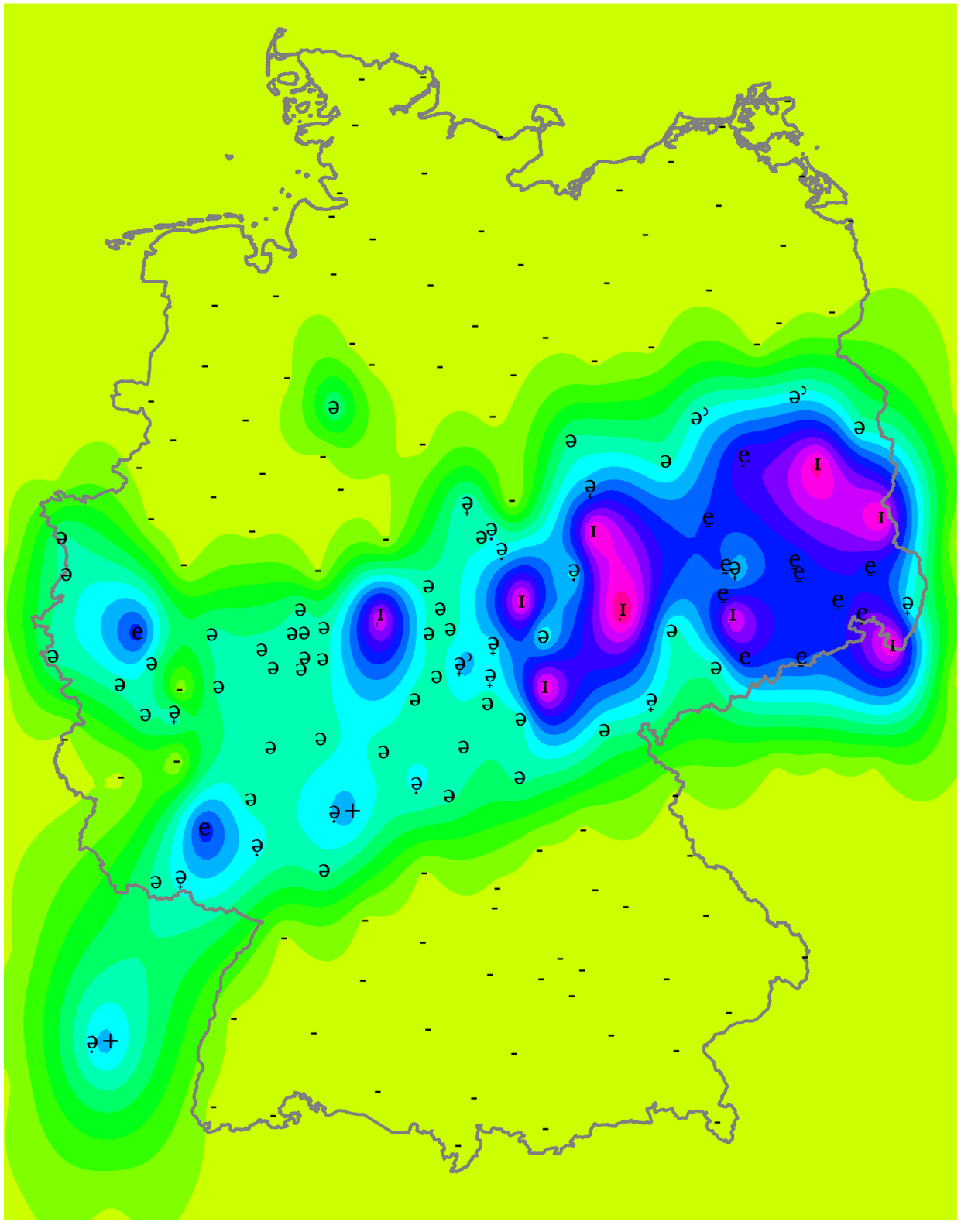


Example 2

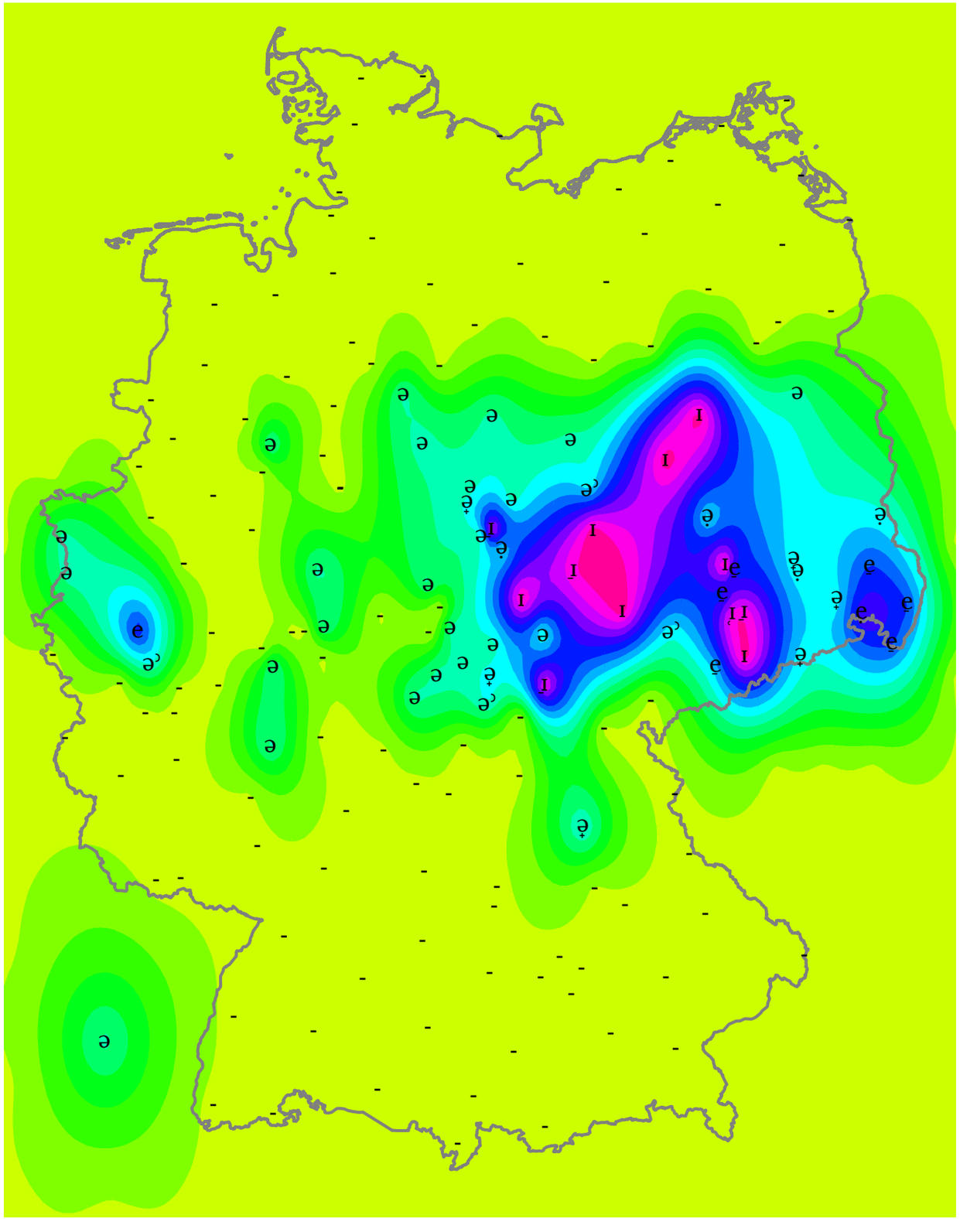
- **Vowel** in German Perfect-prefix ge-
- Ordered to ‘strength’, visualised as ‘height’
- (CLUSTER) (NULL) ə̹ ə̄ ə əˁ ə̟ ə̗ əˀ ə̟ˀ ə̗ˀ

  ə+ ə̗+ ə̟ˀ ə̗ˀ ē e ēˀ ẹ œˁ œ ɛ̄ ɛ̣ I̹ Iˁ Ī I̗ I

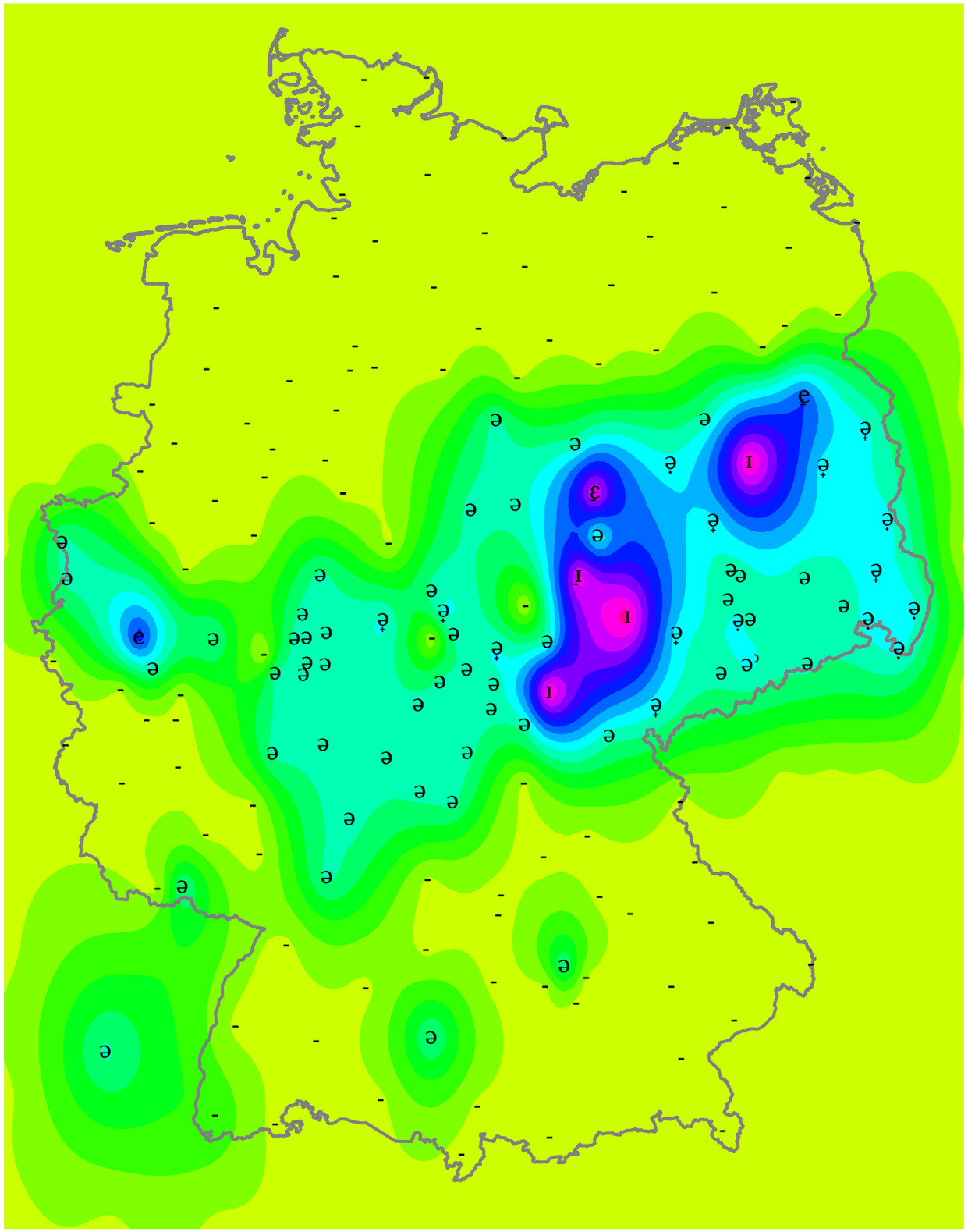
gekannt



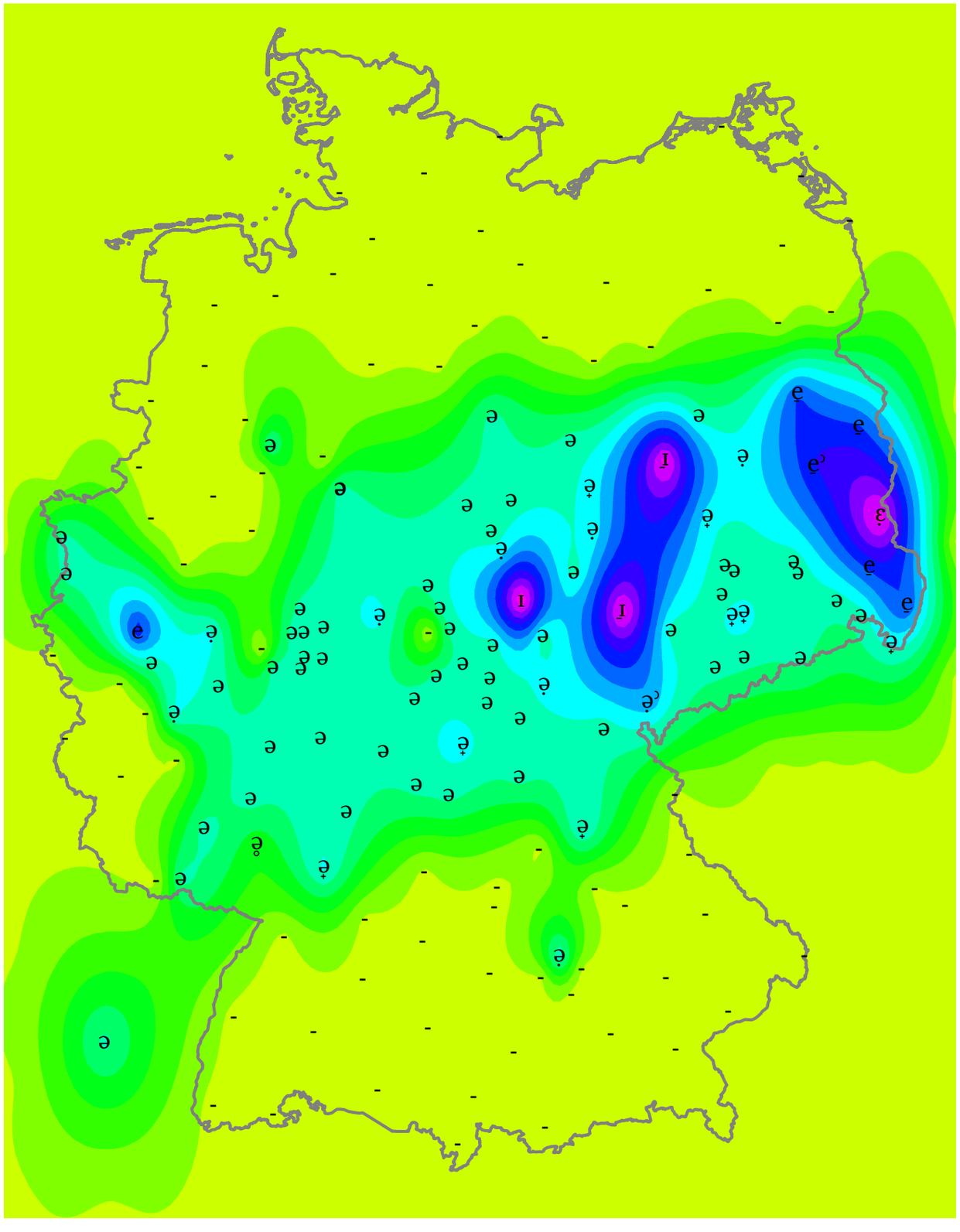
gekommen



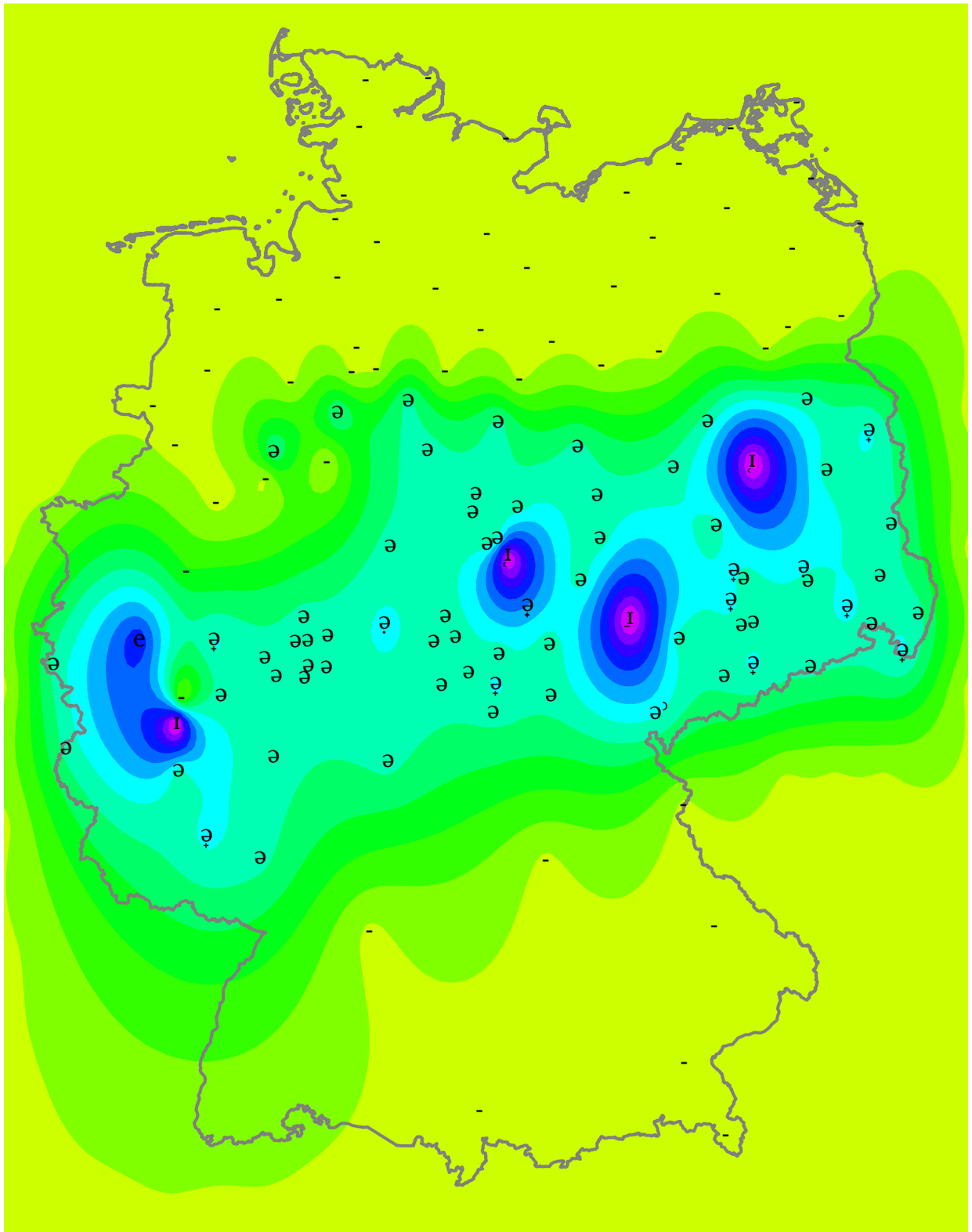
geblieben



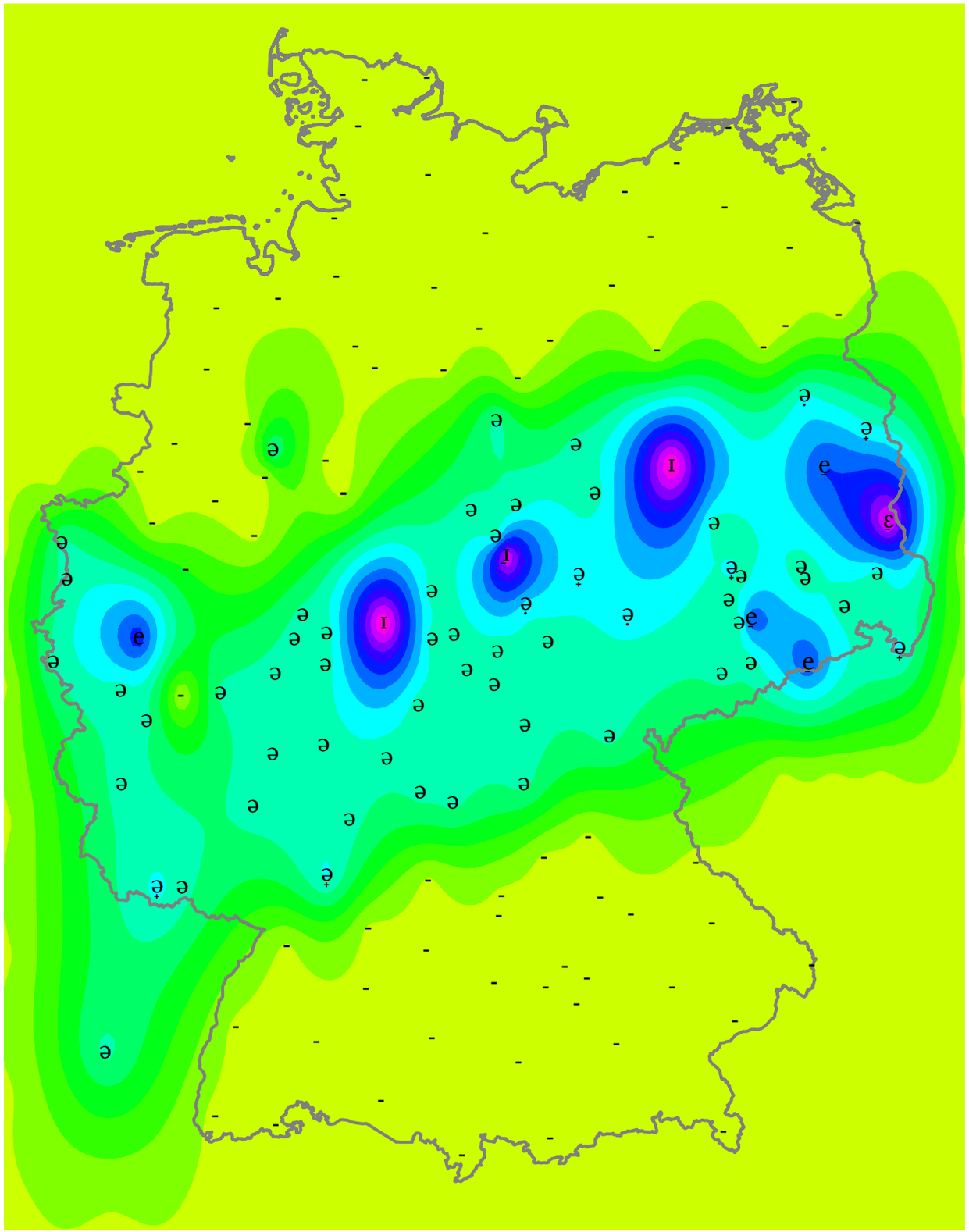
gebracht



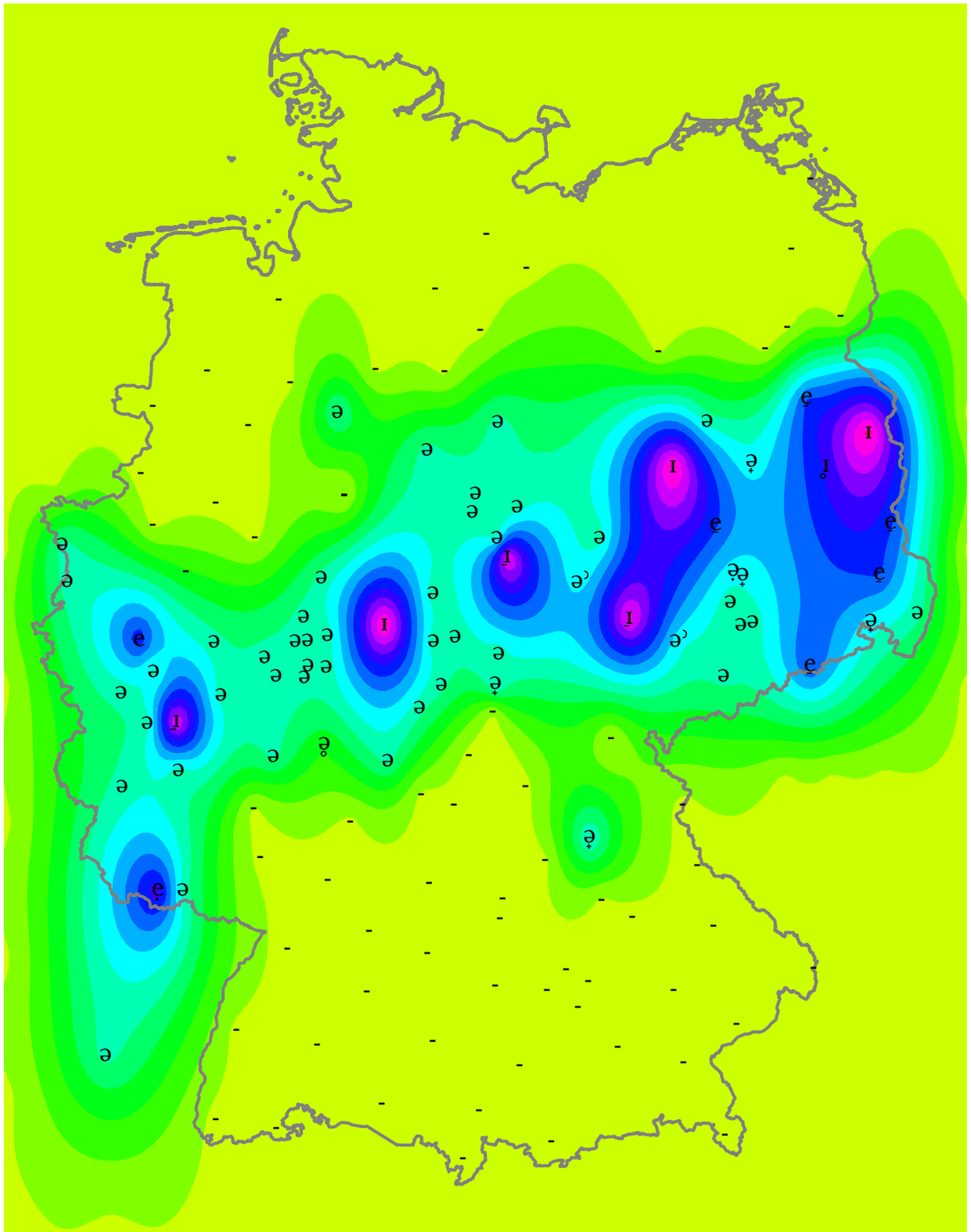
eingebrochen



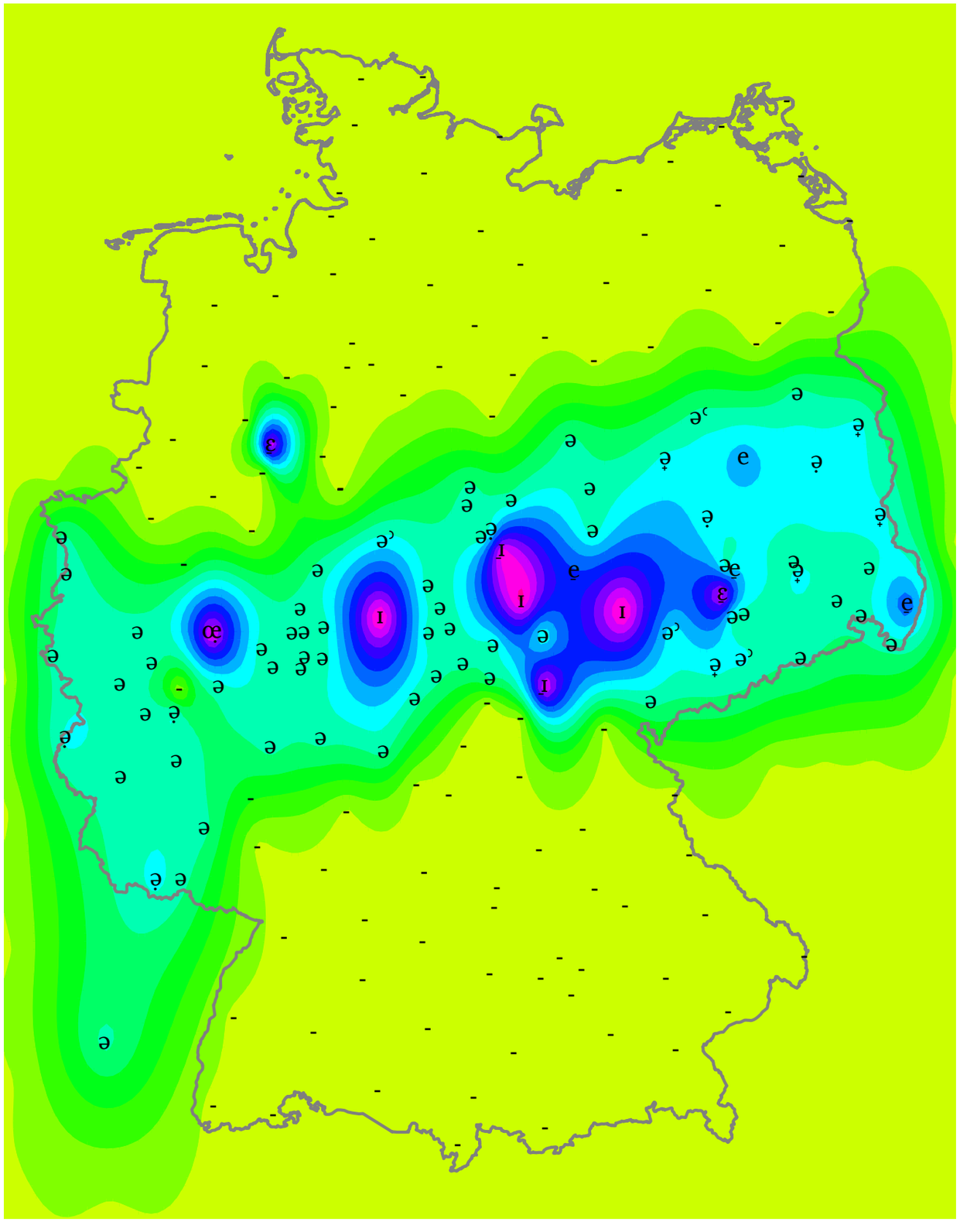
gebrannt



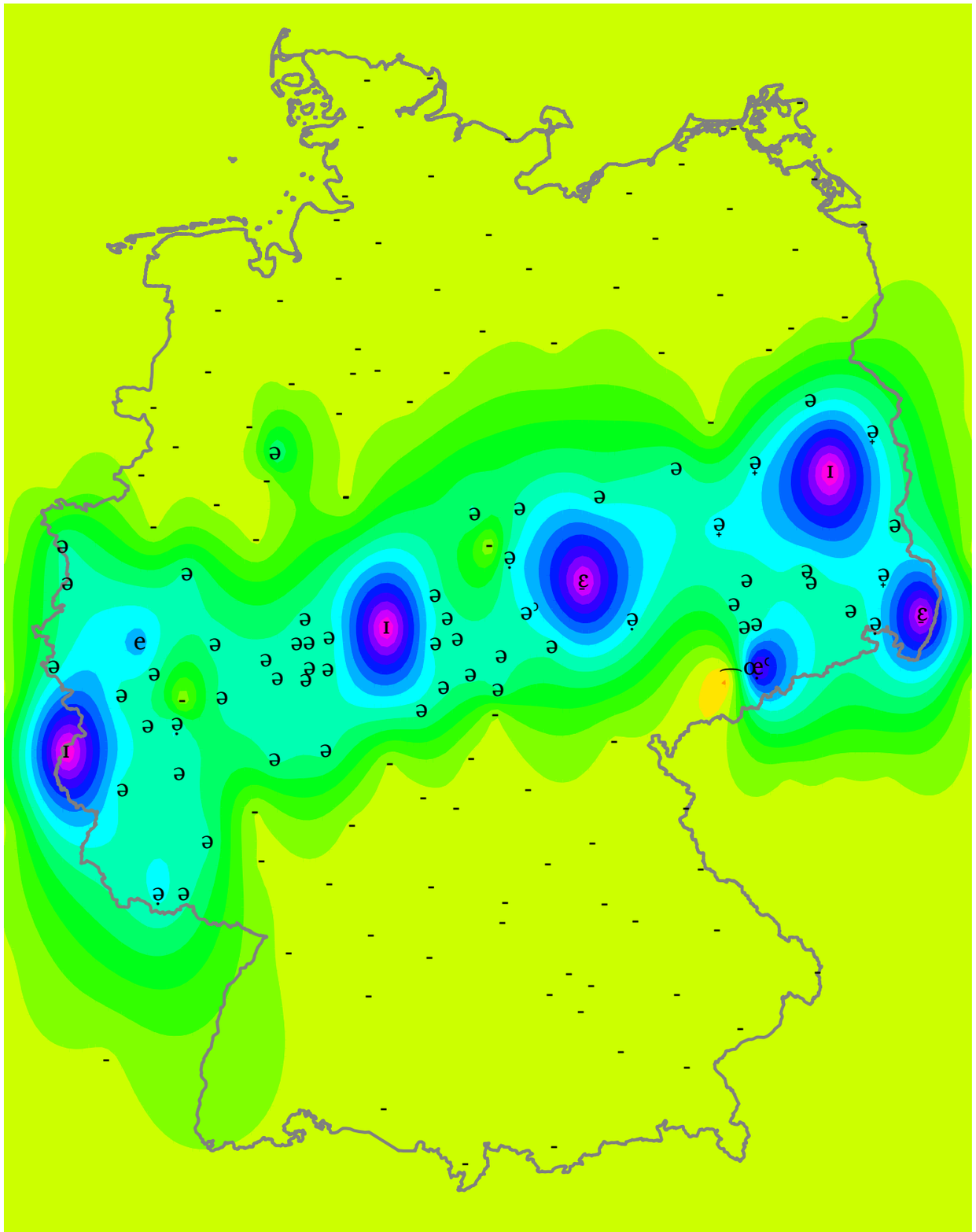
gestohlen



gefahren



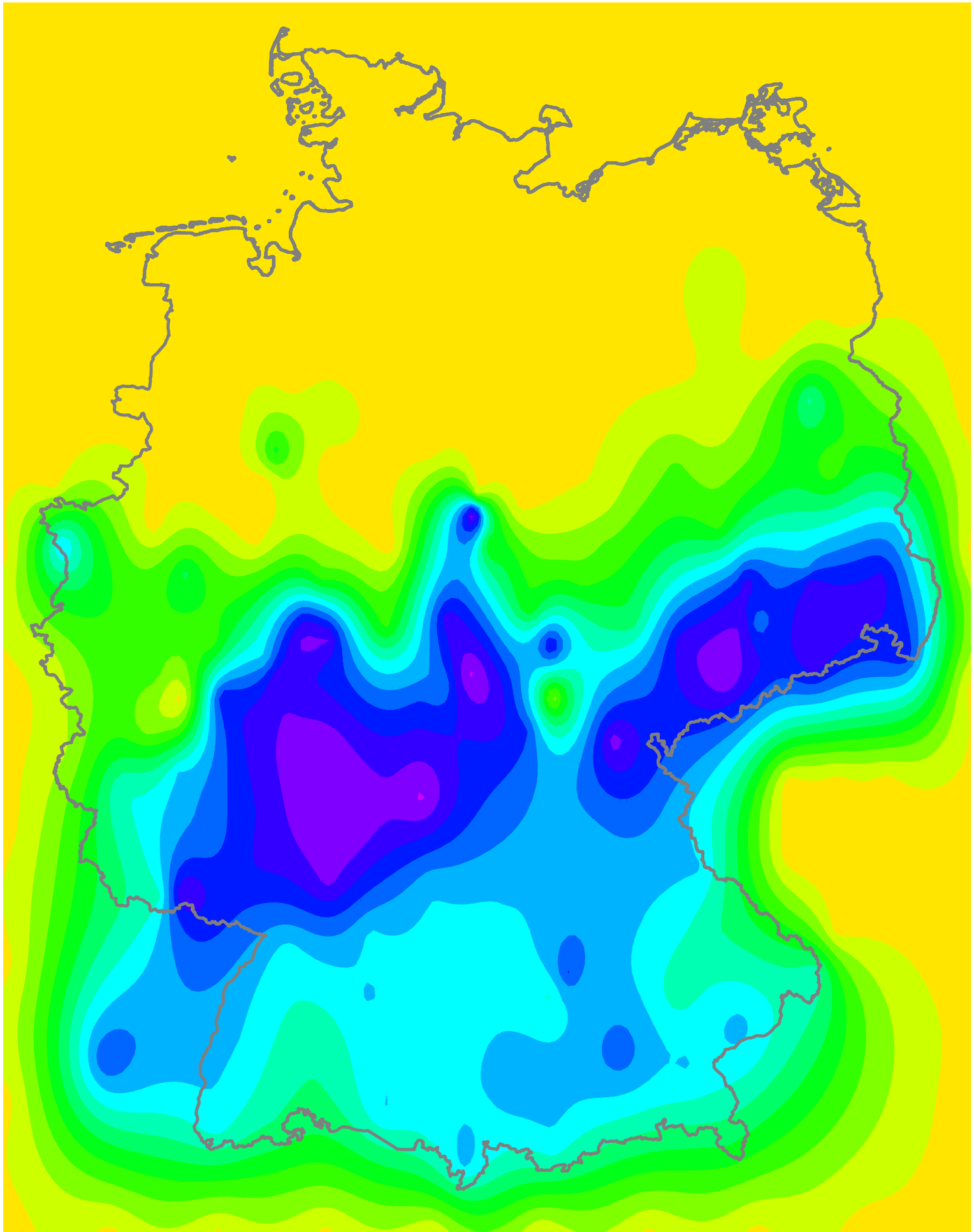
gefallen



Averaging

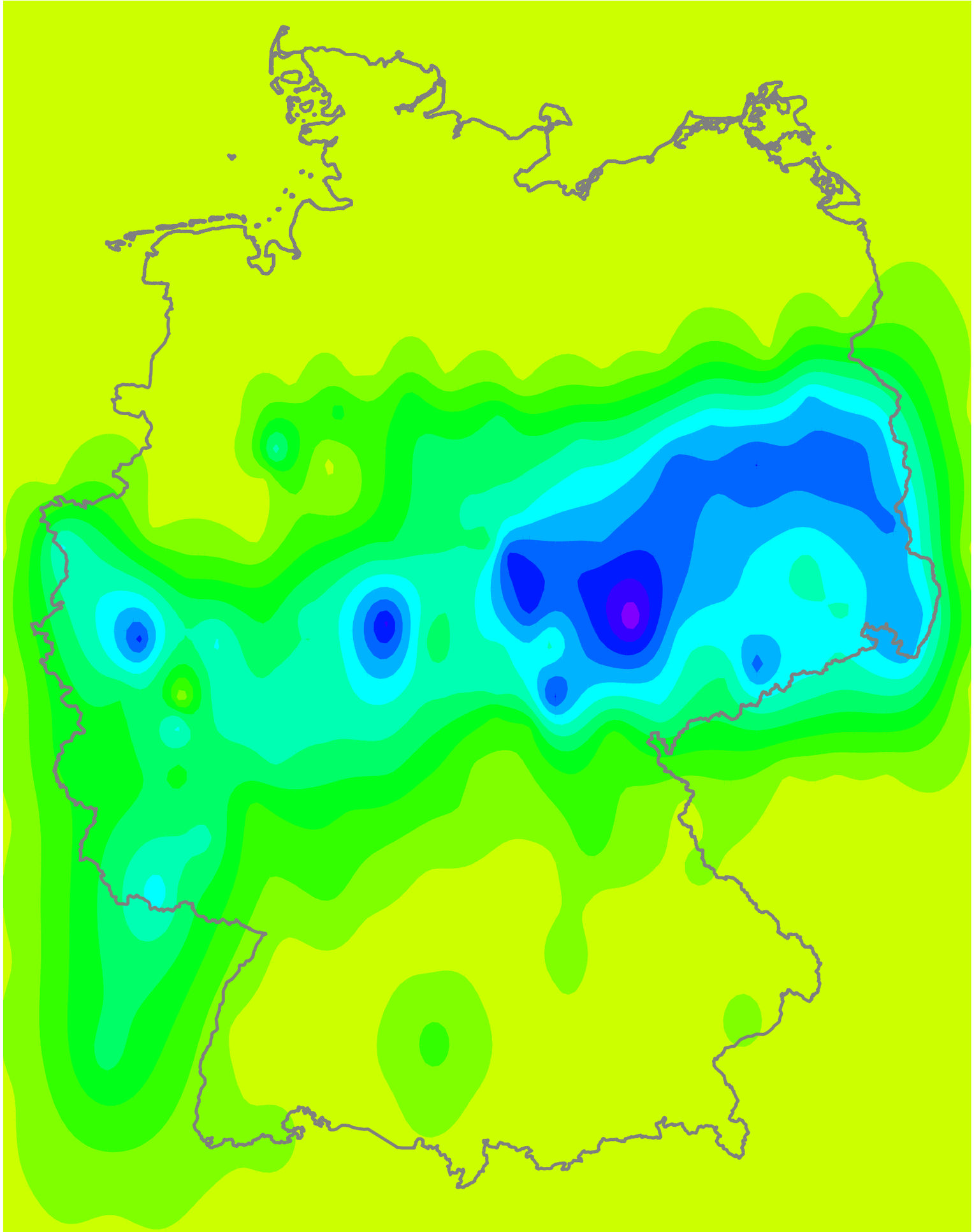
- Combining all consonants and all vowels into a single map

Consonants
of ge- prefix
(averages over
14 lexemes)

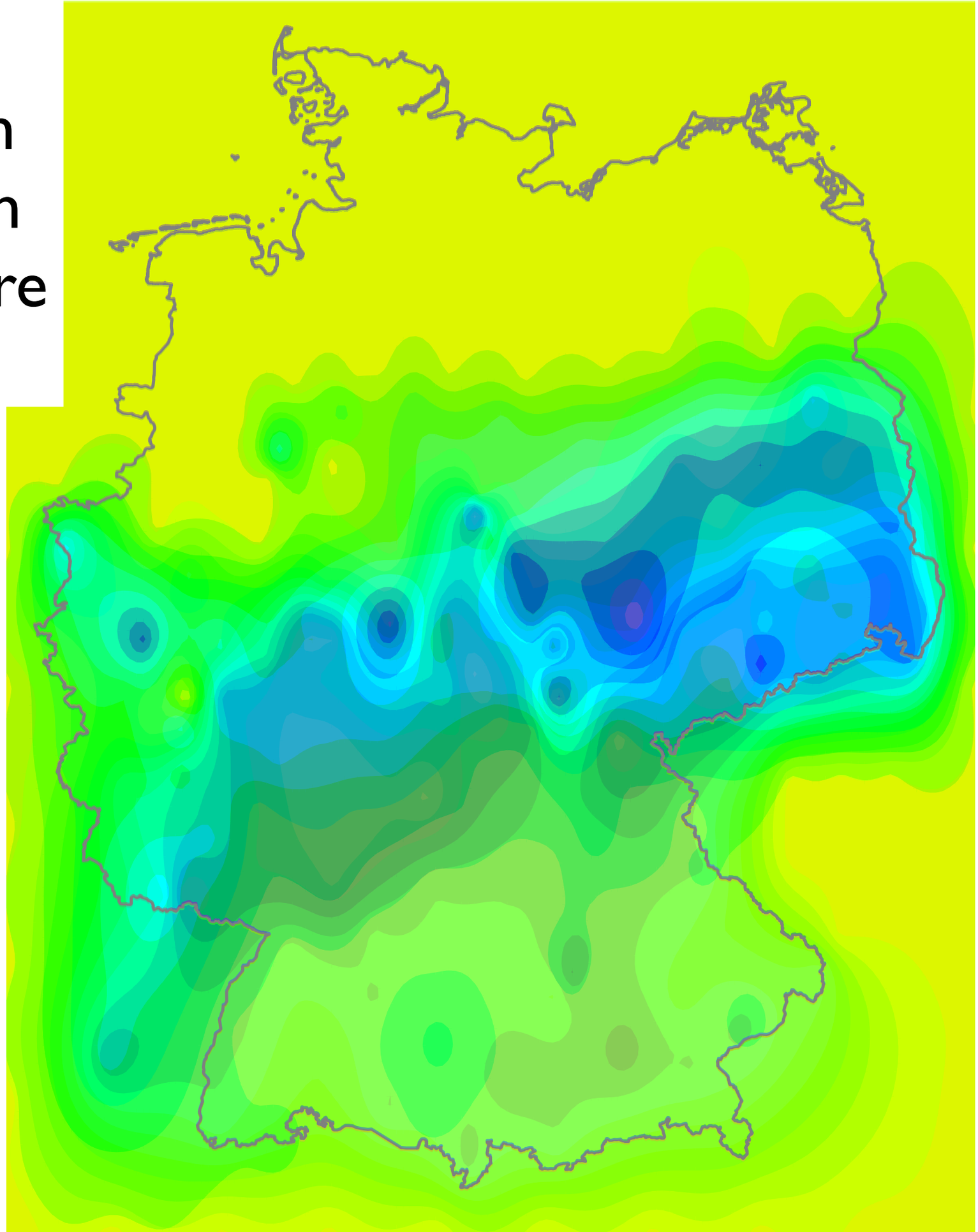


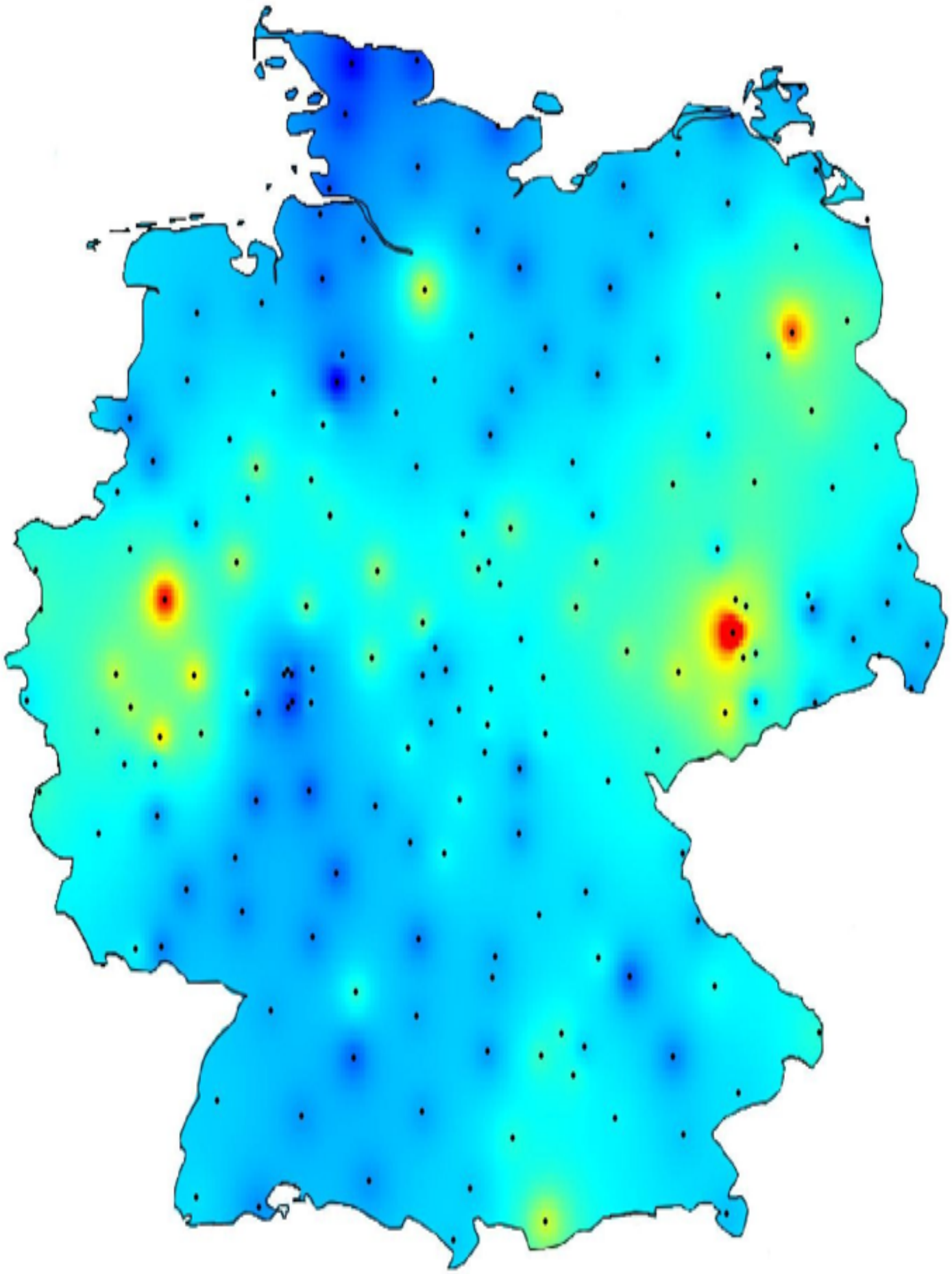
Vowels

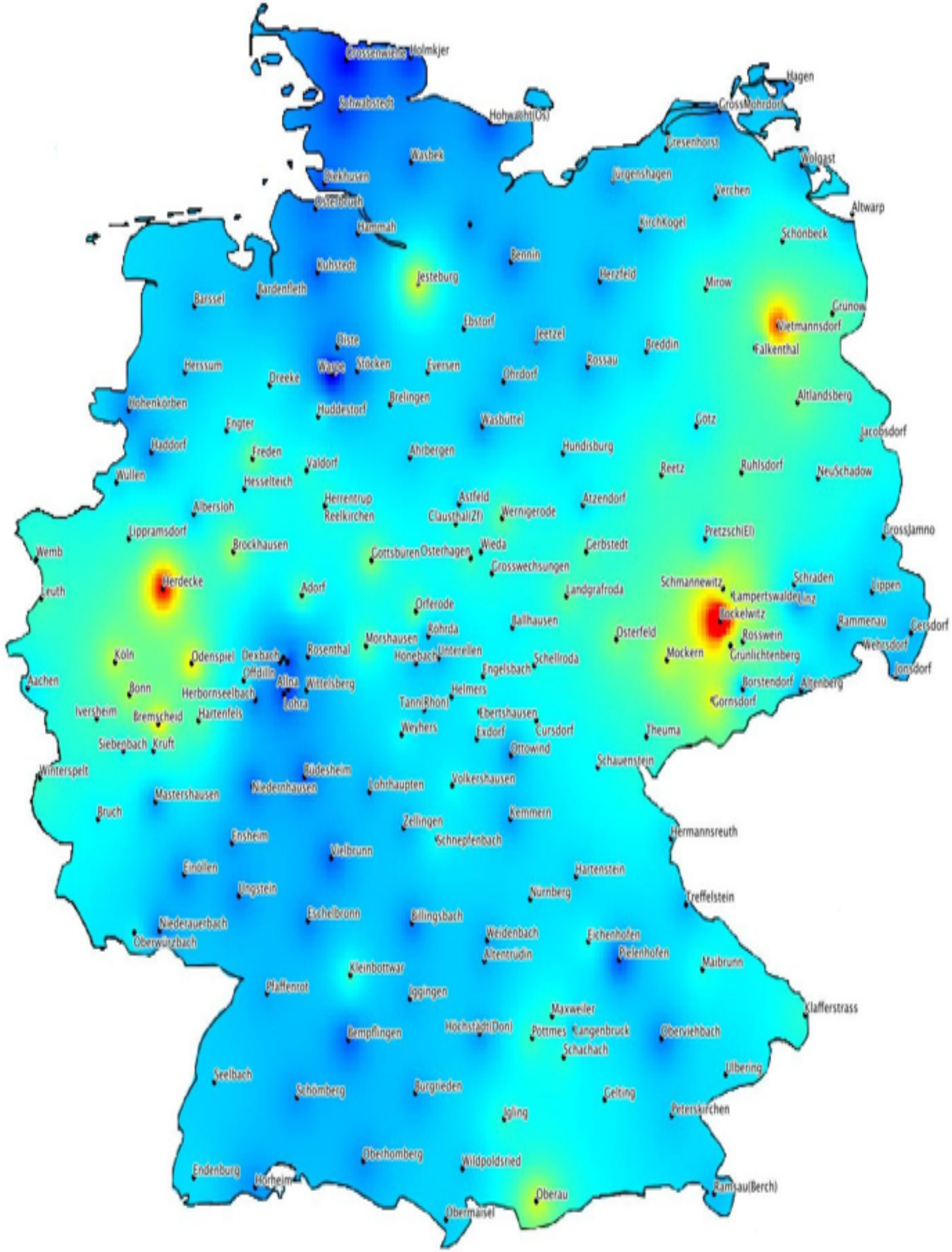
of ge- prefix (averages over 14 lexemes)



Vowels are ‘strong' in an area in which consonants are ‘weakening’







Grossenwiehe
Holmkjer
Hagen
Schwabstedt
Hohwacht(Os)
GrossMohrdorf
Gresenhorst
Wasbek
Wolgast
Jürgenshagen
Dekhusen
Verchen
Altwarp
Hammah
KirchKogel
Schönbeck
Bennin
Kuhstedt
Jesteburg
Herzfeld
Mirow
Barssel
Bardenfleth
Grunow
Vietmannsdorf
Ebstorf
Jeetzel
Falkenthal
Breddin
Elste
Herssum
Warpe
Stöcken
Eversen
Rossau
Ohrdorf
Dreeke
Altlandsberg
Hohenkörben
Brelingen
Huddestorf
Wasbüttel
Götz
Engter
Haddorf
Jacobsdorf
Hundisburg
Freden
Ahrbergen
Valdorf
Reetz
Ruhlsdorf
NeuSchadow
Wullen
Hesselteich
Astfeld
Herrentrup
Atzendorf
Wernigerode
Reelkirchen
Clausthal(Zf)
Albersloh
Lippramsdorf
Pretzsch(El)
Grossjamno
Brockhausen
Wieda
Gerbstedt
Wemb
Cottsbüren
Osterhagen
Grosswechsungen
Schraden
Herdecke
Schmannewitz
Adorf
Landgrafroda
Leuth
Lampertswalde
Linz
Lippen
Orferode
Zockelwitz
Rammenau
Gersdorf
Röhrda
Ballhausen
Osterfeld
Rosswein
Morshausen
Wehrsdorf
Köln
Odenspiel
Dexbach
Rosenthal
Honebach
Unterellen
Mockern
Grunlichtenberg
Schellroda
Jonsdorf
Offdilln
Engelsbach
Aachen
Allna
Wittelsberg
Borstendorf
Altenberg
Bonn
Herbornseelbach
Helmers
Lohra
Tann(Rhön)
Gornsdorf
Iversheim
Hartenfels
Ebertshausen
Weyhers
Cursdorf
Theuma
Exdorf
Siebenbach
Kruft
Ottowind
Schauenstein
Winterspelt
Büdesheim
Lohrhaupten
Volkershausen
Niedernhausen
Mastershausen
Kemmern
Bruch
Zellingen
Hermannsreuth
Ensheim
Schnepfenbach
Vielbrunn
Einöllen
Hartenstein
Nürnberg
Treffelstein
Ungstein
Eschelbronn
Billingsbach
Niederauerbach
Weidenbach
Eichenhofen
Oberwürzbach
Altentrüdin
Pielenhofen
Maibrunn
Kleinbottwar
Pfaffenrot
Iggingen
Maxweiler
Klafferstrass
Hochstädt(Don)
Pöttmes
Langenbruck
Oberviehbach
Bempflingen
Schachach
Ulbering
Seelbach
Schömberg
Burgrieden
Gelting
Igling
Peterskirchen
Oberhomberg
Wildpoldsried
Endenburg
Horheim
Ramsau(Berch)
Oberau
Obermaisel

The git-approach

- git is a versioning control system: keeping track of who changed what
- You can run it privately on your own computer, or use services like github or bitbucket
- Text-based: so use simple formats! (CSV in UTF-8, NFC, LF)

Conclusions

- Interpolation results in nice pictures
- Trends can be made visible
  - watch out with sparse data
  - be aware of the many possible interpolations
- Boundary-lines can be drawn (“Isohypses”), but should not be interpreted as traditional boundaries