The Problem of Representativeness of Historical Folklore Databases*

The Case of Musical Collection of Oskar Kolberg

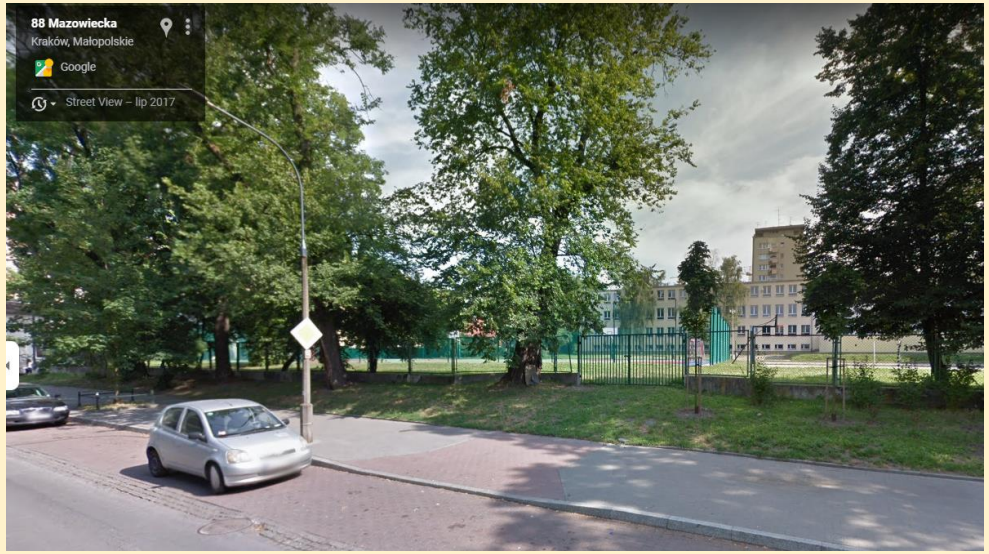
Rafał Miśta

Institute for Social Studies, University of Warsaw

---

*This research was supported by the Polish National Science Center (NCN), grant No. 2018/29/N/HS2/02973 awarded to Rafał Miśta

Common Problems with Historical Folklore Datasets



Known: what was collected (with unknown quality)

Unknown: what was not collected & what could be collected

MISSINGS or ABSENCE?

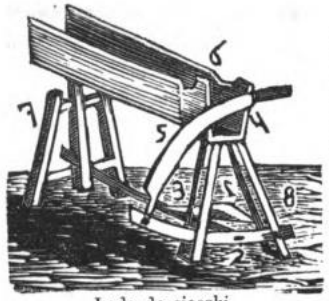
Historical ethnographic dataset

External validation

- later observations
- earlier observations
- other sources

Internal validation

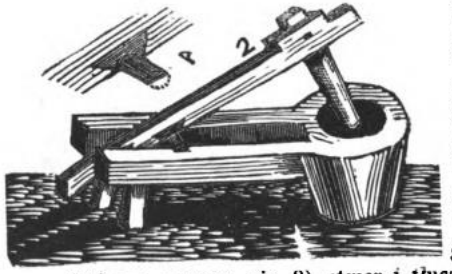
- sample biases
- internal cohesion
- testing predictions from correlates



Lada do sieczki.

W ladzie: 1) *rabsla*, *rapsla*, deszczułka do ruszania nogą; 2) *kładka*, kij krzywy; 3) *pies*, także kij węższy; 4) *stolica;* 5) *kosisko* (kosa) 6) skrzynka; 7) nogi z tyłu; 8) nogi podwójne z przodu.

Stępa (stempa) rodzaj moździerza dębowego w każdéj znajdującego się chałupie, z przyrządem poruszanym nogą służąca do spychania prosa (t. j. do jagieł) i do tłuczenia jęczmienia na *pęcak* czyli na kaszę jęczmienną (obłupywanie to zowie się: bić, tłuc czyli *chróstać* pęcak). Baba zwykle wówczas staje na téj stępie i trzymając się ręką kołka w ścianie, tłucze nogą prawą. Przy stępie jest 1) *biegun* z drzewa, za pomocą którego porusza się 2) *stępor* i tłucze w moździerzu jęczmień. Wszystko z drzewa.



Garczek *dwojak*. *Koziołek* do smażenia omasty.

Żarna uboższych.

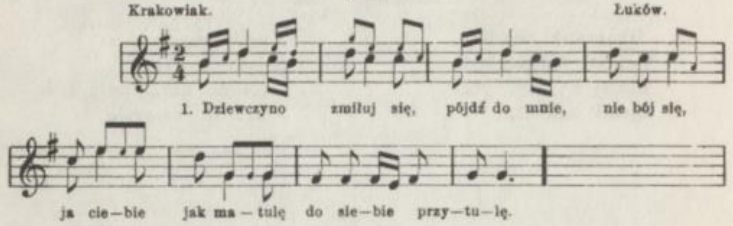
od Skulska i Ślesina.

2. Hultaje, hultaje,
zkąd wam Pan Bóg daje?
Ni z łokcia, ni z pługa,
od Boga zasługa.

3. Hultaje, hultaje,
czego wam nie staje?
Jest gorzałki flaszka,
jest dziewczyna chwacka.

546.

Krakowiak. Łuków.



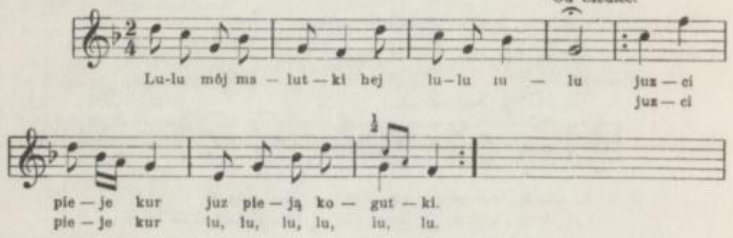
2. Stworzył Pan Bóg Ewę,
kochała Adama,
i my też ich potomki,
natura ta sama.

3. Zebym był malarzem
to bym cię malował,
co-bym na cię spojrzał,
to-bym cię całował.

Kołyska.

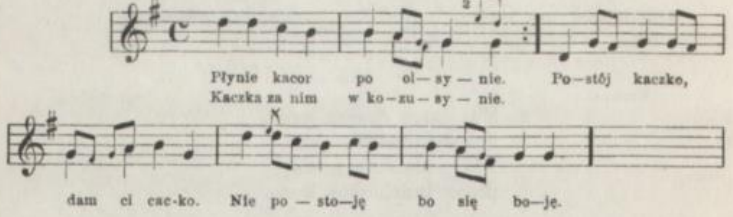
547.

Od Siedlec.

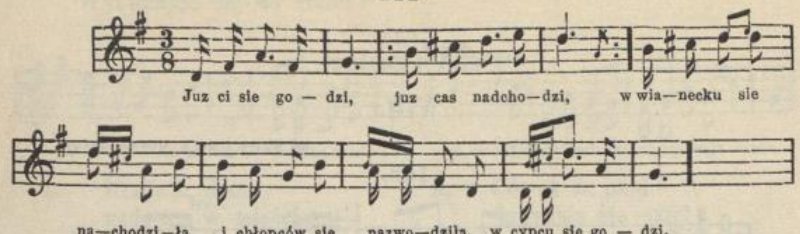


548.

Węgrów.



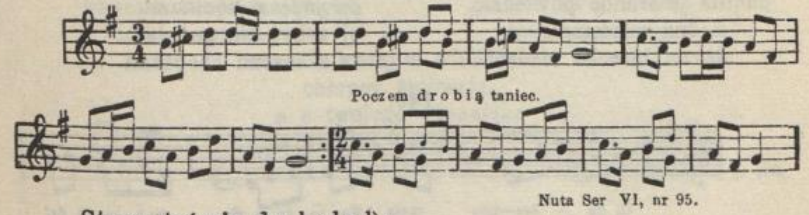
144.



Lud, Ser. VI nr 99. — Nuta Ser. II n. 137. — Ser. III n. 53. 71. — Ser. X n. 32.

Po oczepinach tańczą polskiego (Polinoza):

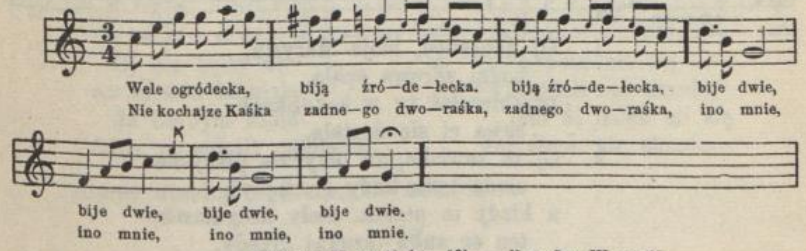
145.



Nuta Ser VI, nr 95.

Starzy tańcują do koła[1]):

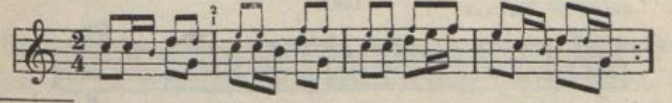
146.



Lud, Ser. II. nr 49 (str. 52). — Nuta Ser. VI, nr 45.

Prędki (przy Weselu po Drobnym):

147.



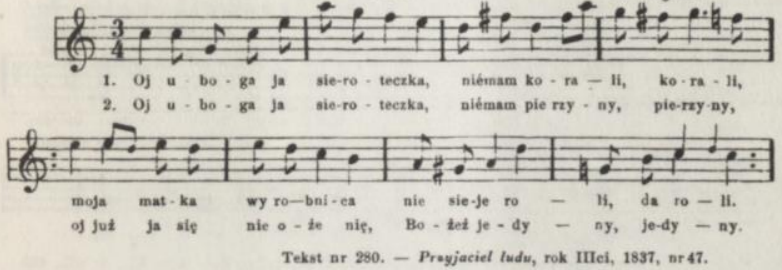
[1]) Niektórzy taniec ten nazywają Drobnym Polinozem.

Polonezy, Obertasy, Mazury i t. d.

(w rytmie $^3/_4$ i $^3/_8$).

133.

od Mogilna.

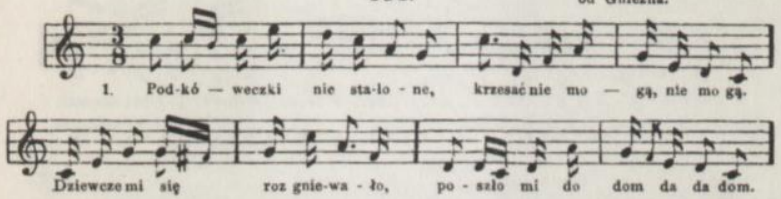


Tekst nr 280. — *Przyjaciel ludu*, rok IIIci, 1837, nr 47.

Nuta ob. Ser. IV, nr 170. 359.
Ser VI, nr. 692.

134.

od Gniezna.



2. Podkóweczki już stalone,
toć krzesają jak szalone.
Podkóweczki z saméj stali,
do Marysi daléj, daléj.

3. Podkóweczki dajcie (w) ognia,
bo Marysia tego godna.
A czy godna, czy niegodna,
podkóweczki dajcie (w) ognia.

Od Wyrzyska, Nakła (nuta nr 145):

4. Podkóweczki ognia!
czemu nie krzeszecie?
Wy Kunowskie panny.
czemu mnie nie chcecie?

5. Mybyśmy cię chciały,
gdybyś miał talary.
Myszy-ć kieszeń zjadły,
talary wypadły.

Przyjaciel ludu, 1836, rok III, nr 26. — Nuta *Lud*, Ser. IV, nr 138.
ob. Ser. I nr 275 (str. 393).

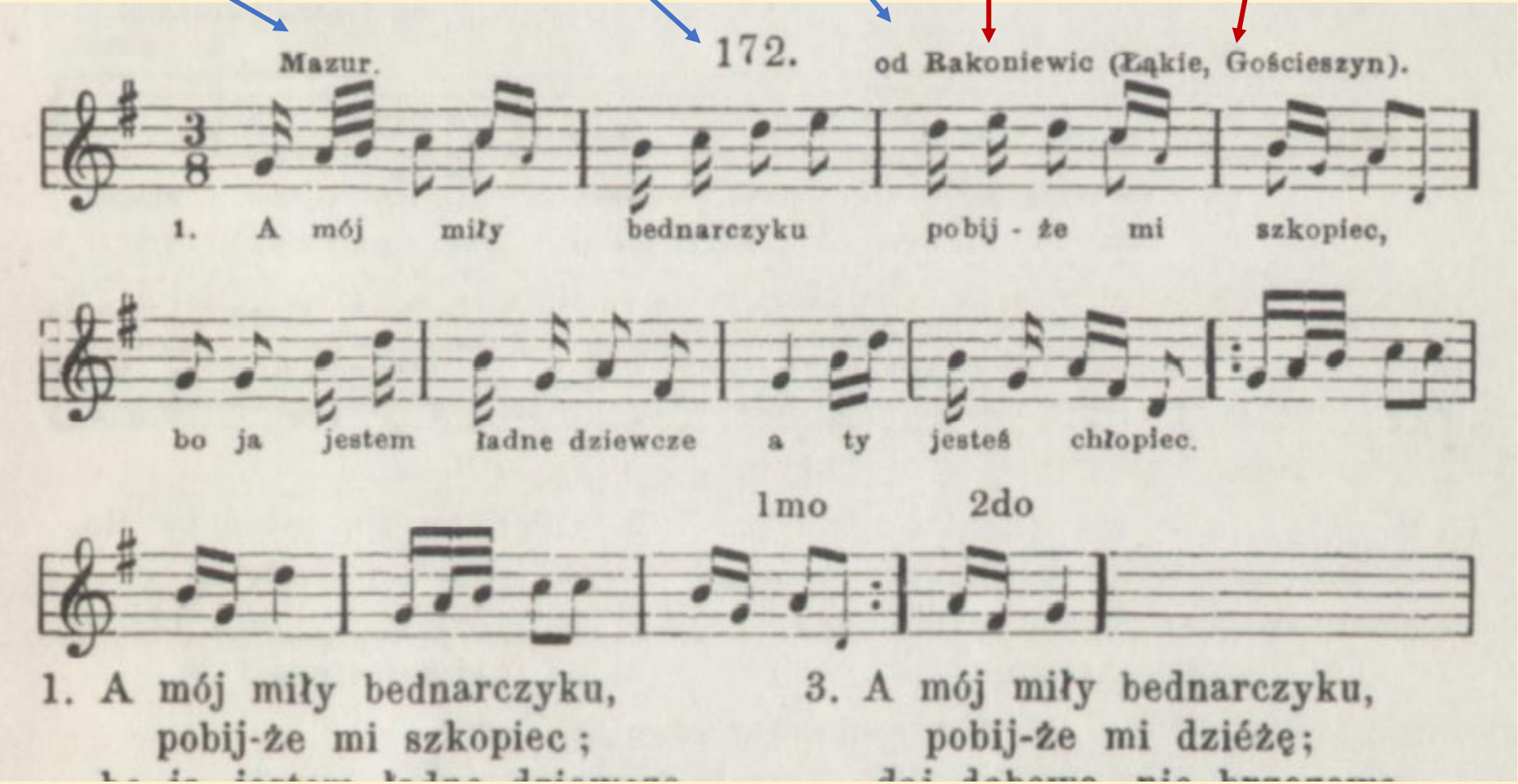
information about geographical location

location (general)

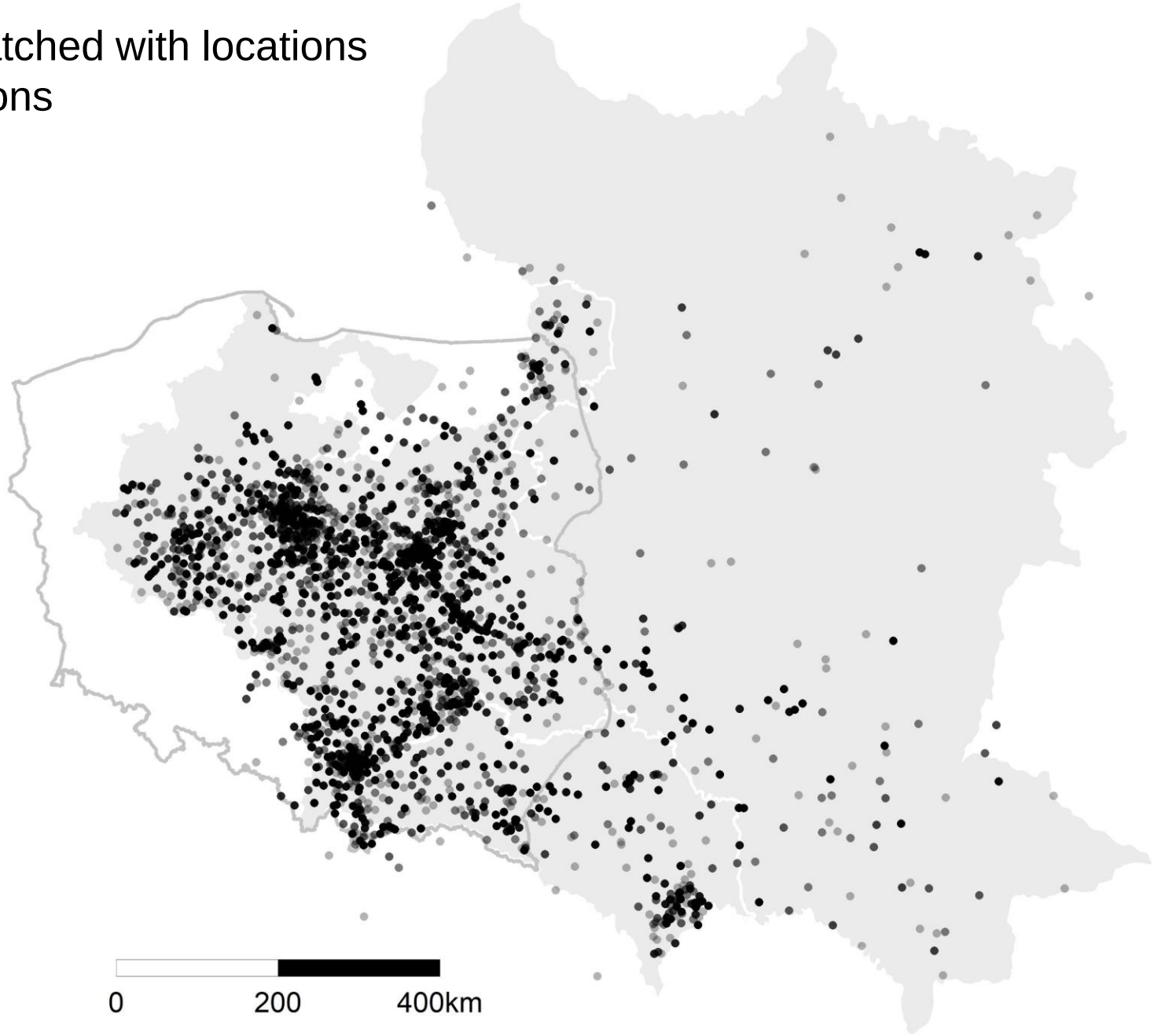
id

location (precise)

remarks (e.g. tune type)

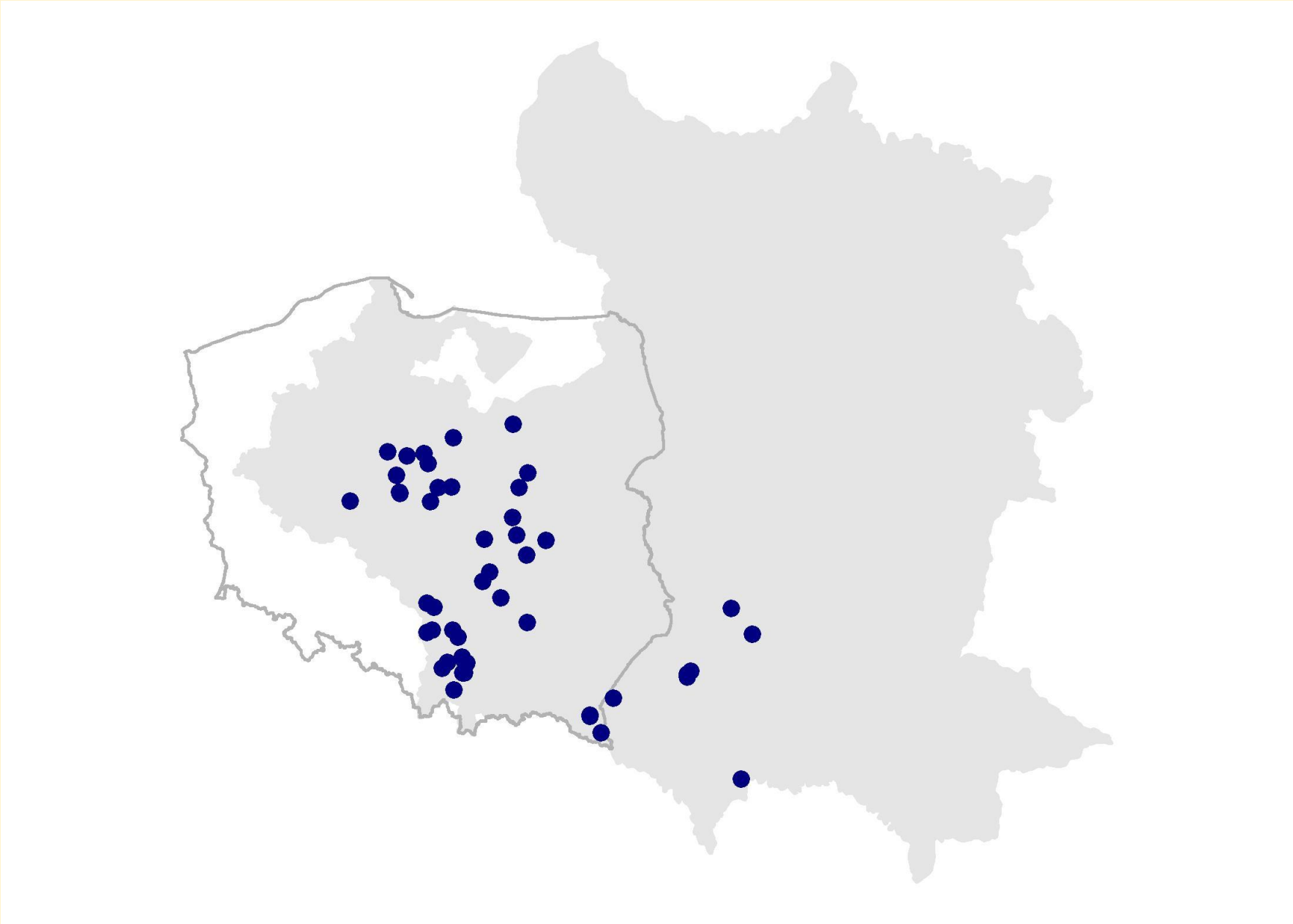


17 358 melodies matched with locations
2 259 various locations

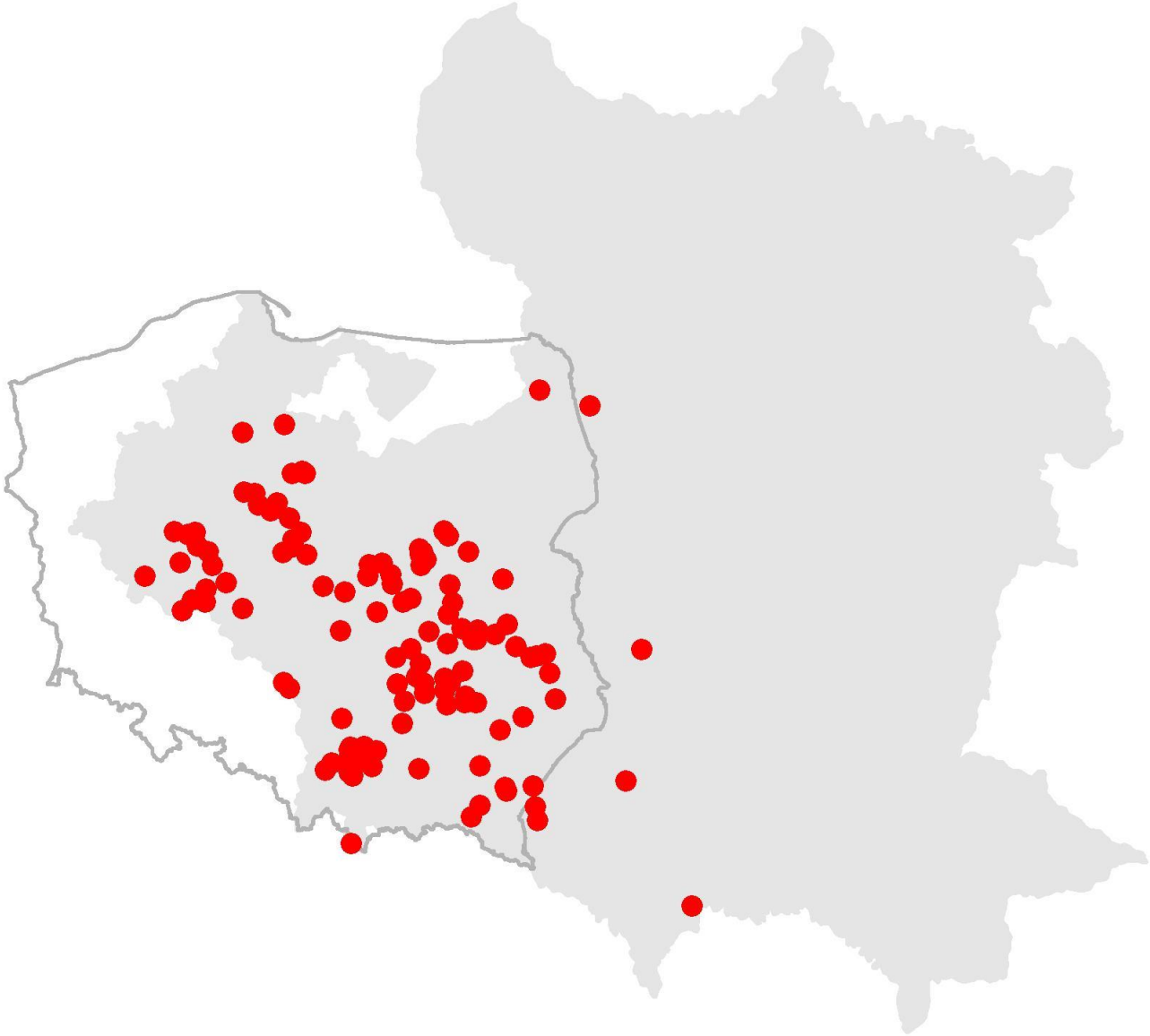


WALC

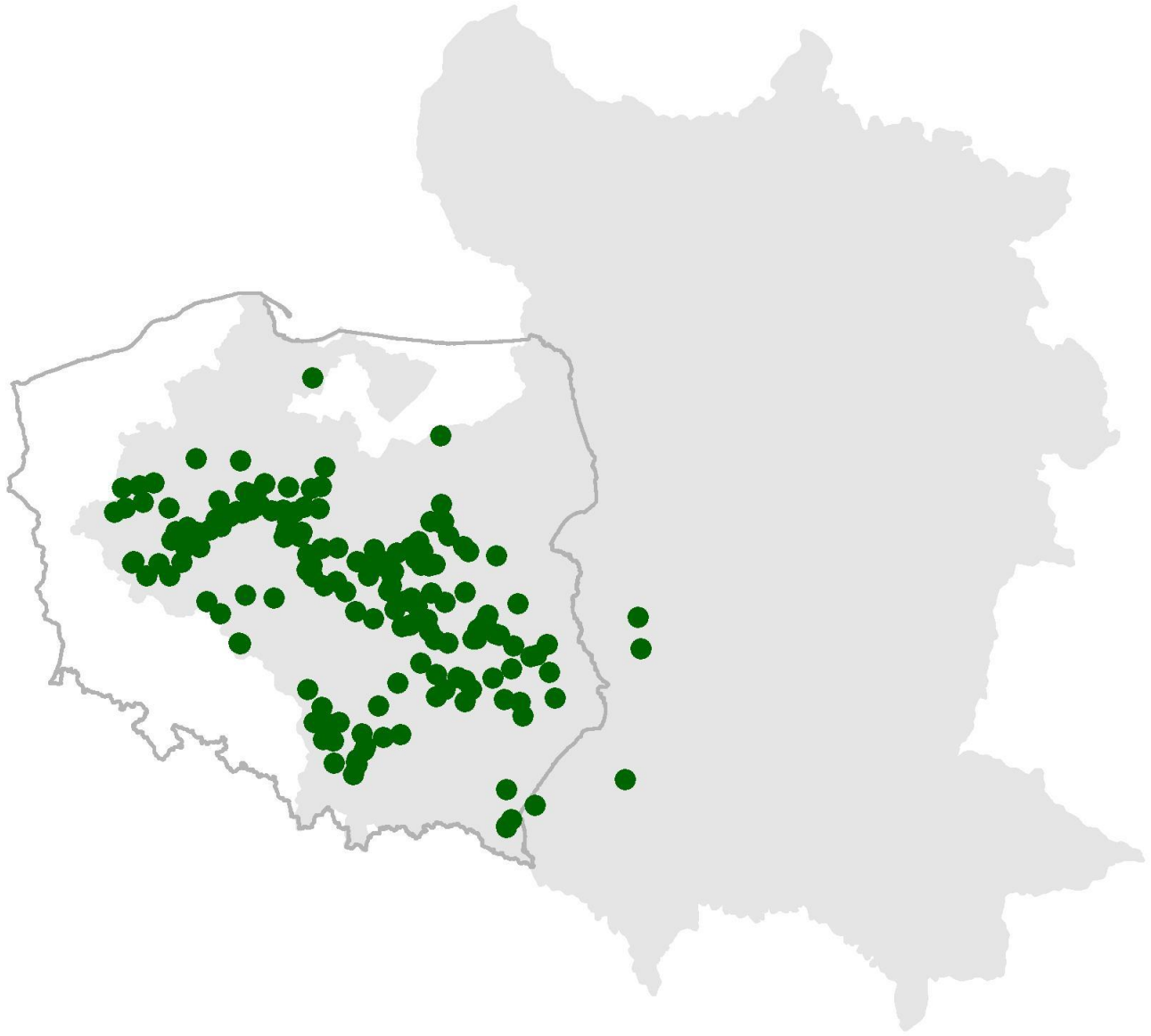
POLKA



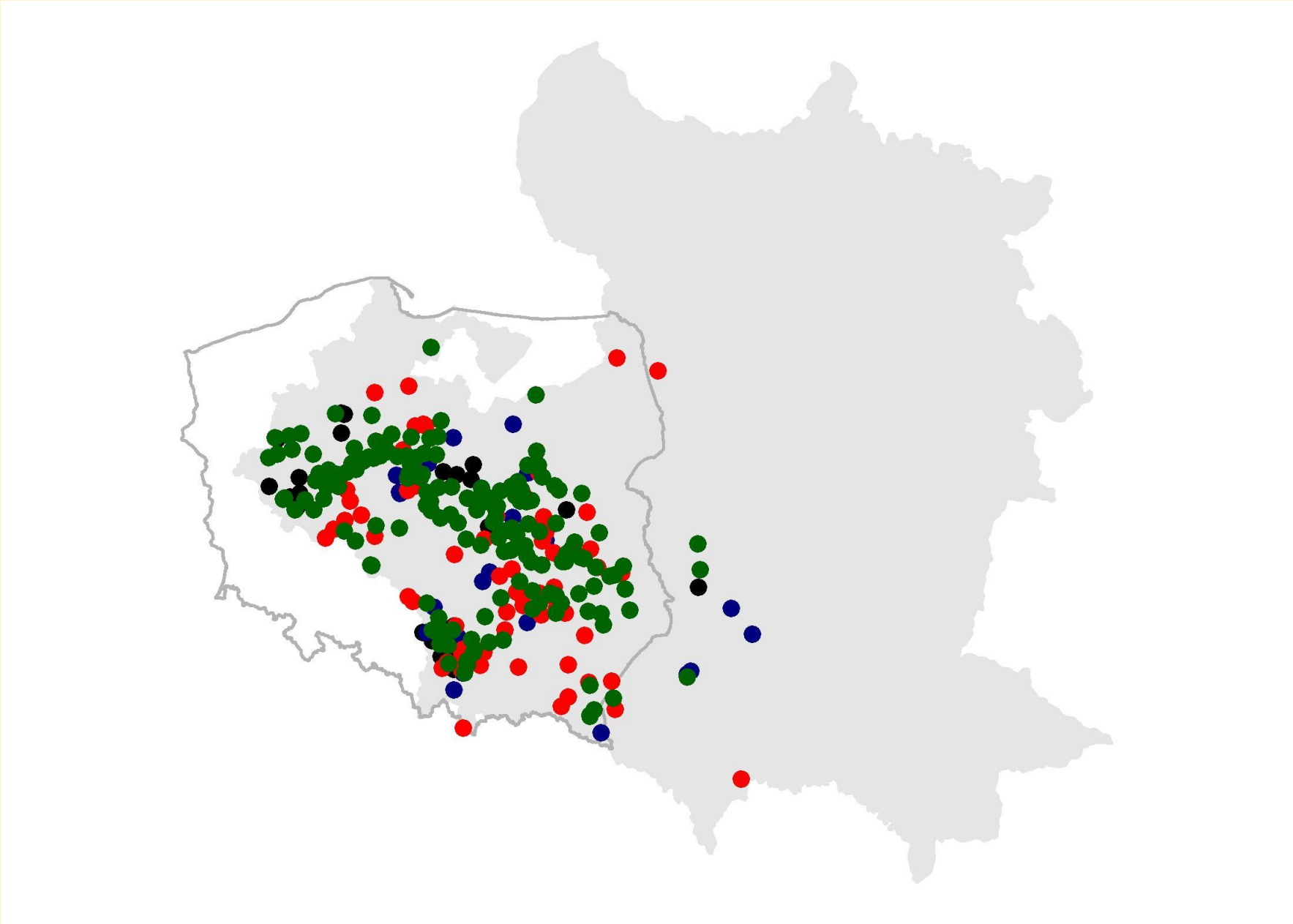
KRAKOWIAK



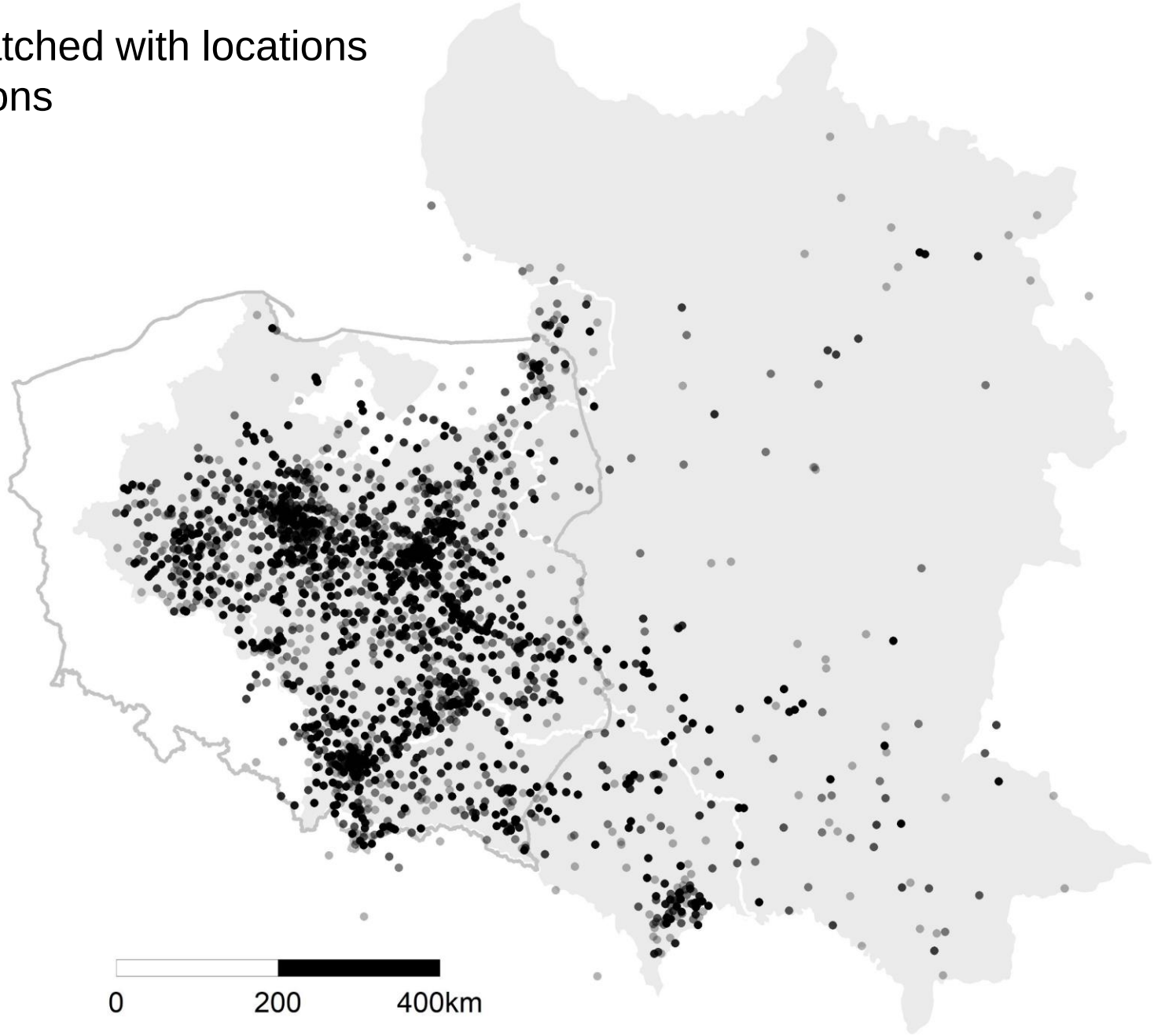
MAZUR

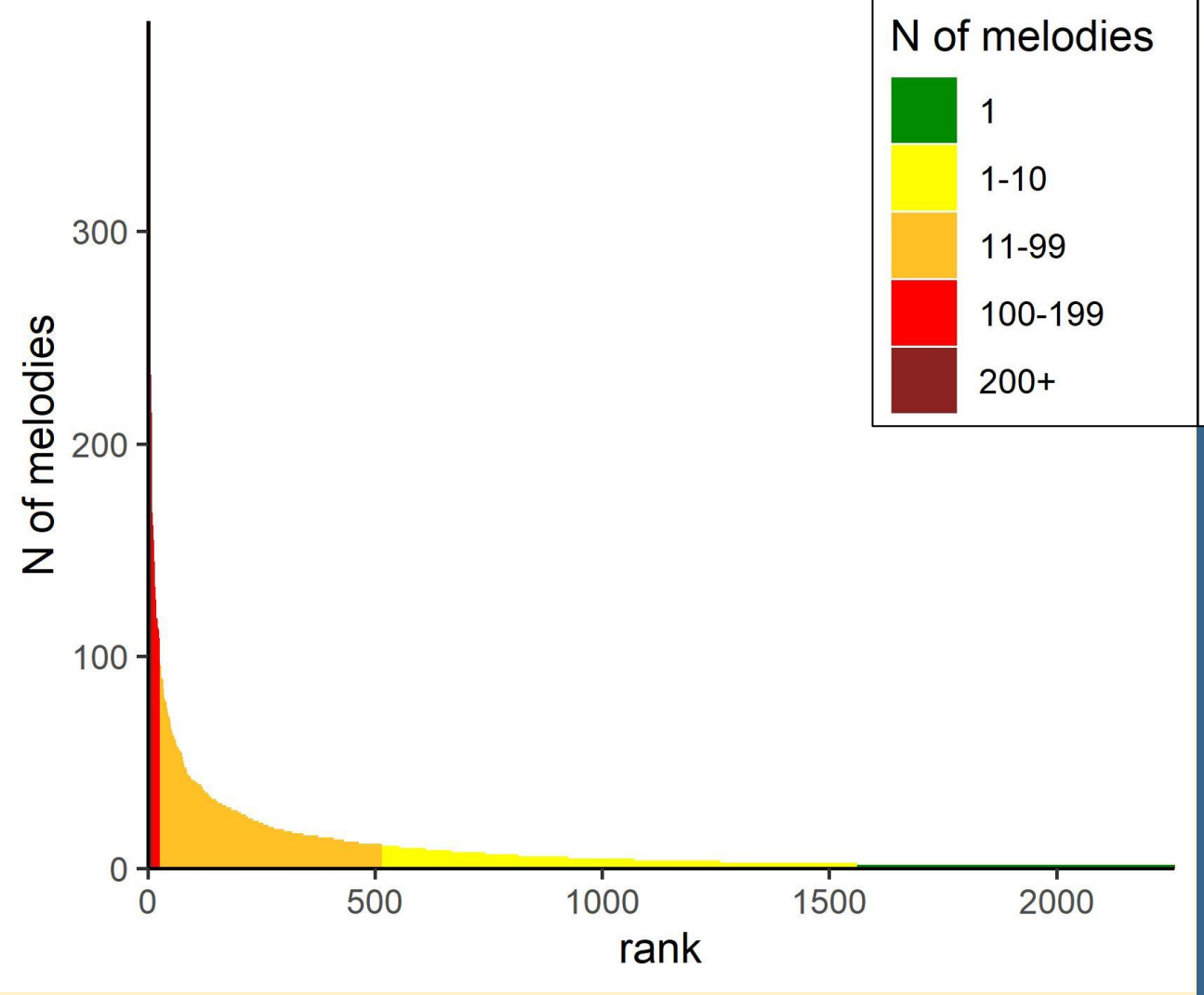


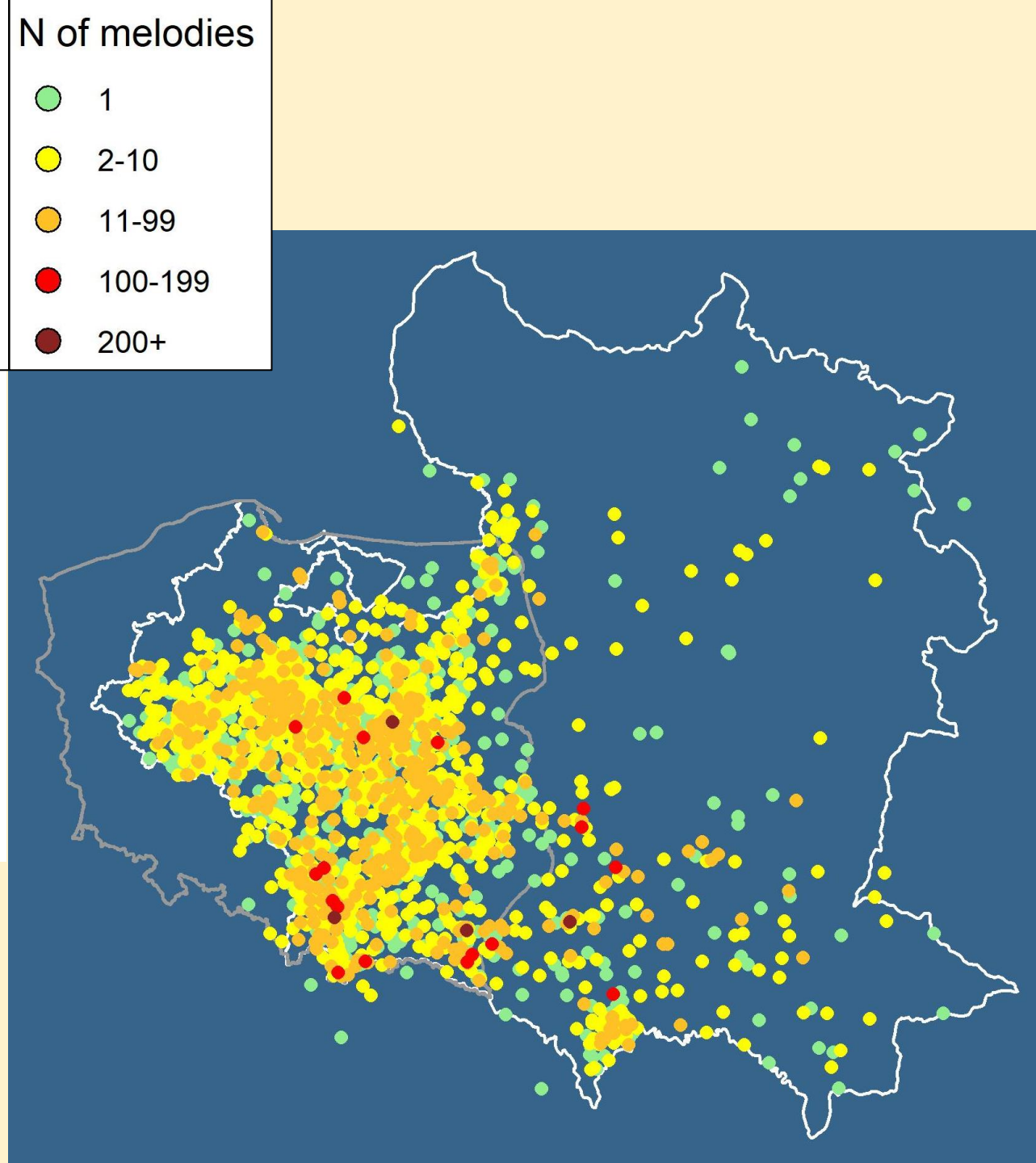
TOGETHER



17 358 melodies matched with locations
2 259 various locations





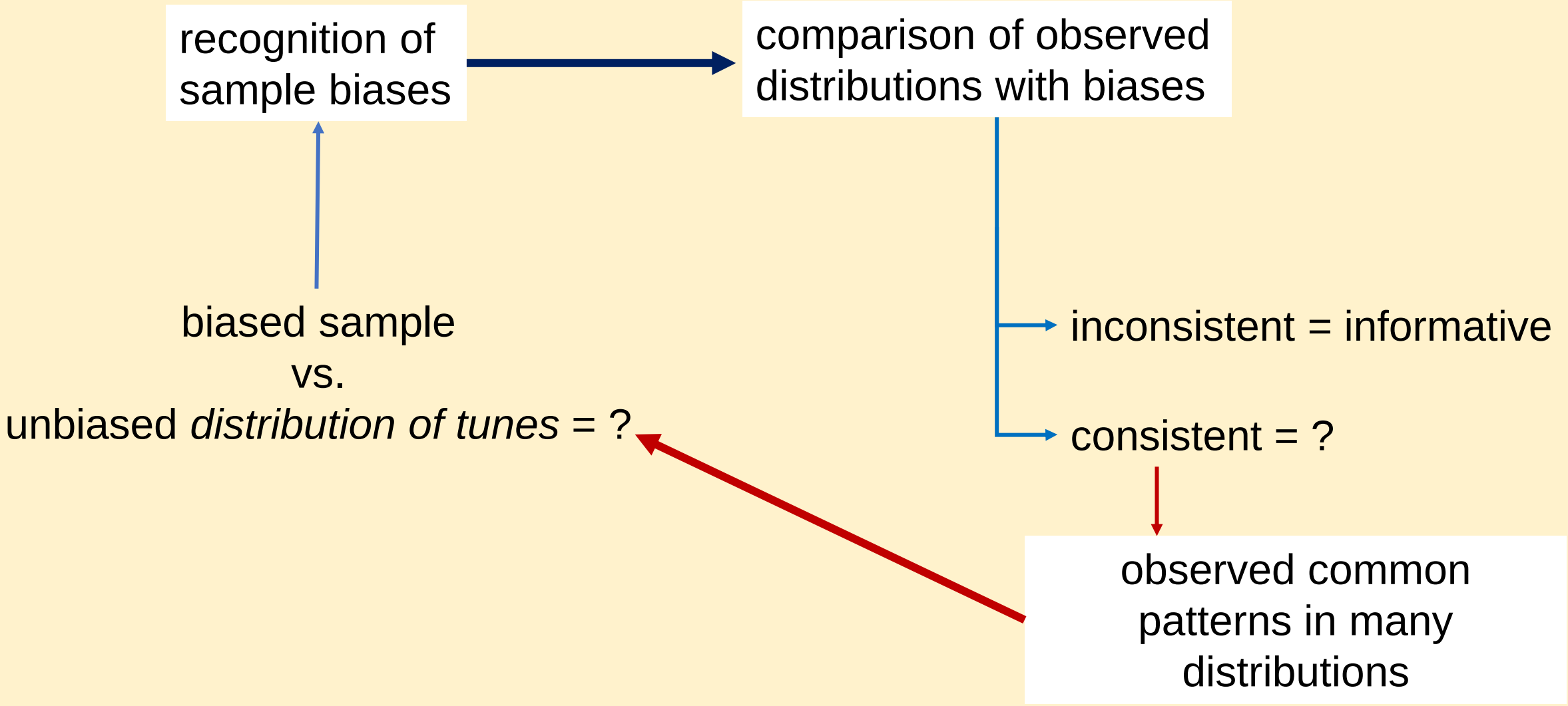


- 1% of locations is matched with more than 100 melodies
- over 30% is matched with a single melody

Problems with Musical Folklore Collection of Oskar Kolberg

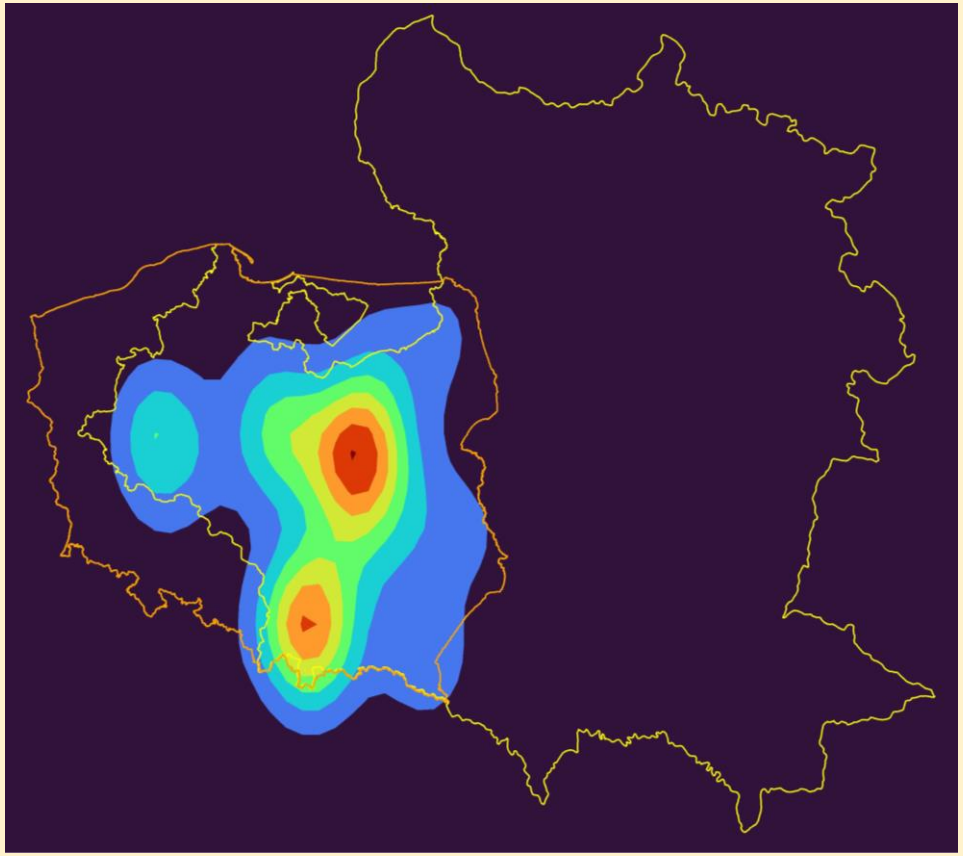
- omitted musical types (in a given location) - *„missing or absence”* problem
- unequal probability of collecting tunes in various locations
- accidental or faulty observations (*outliers*)

Internal validation of Oskar Kolberg collection

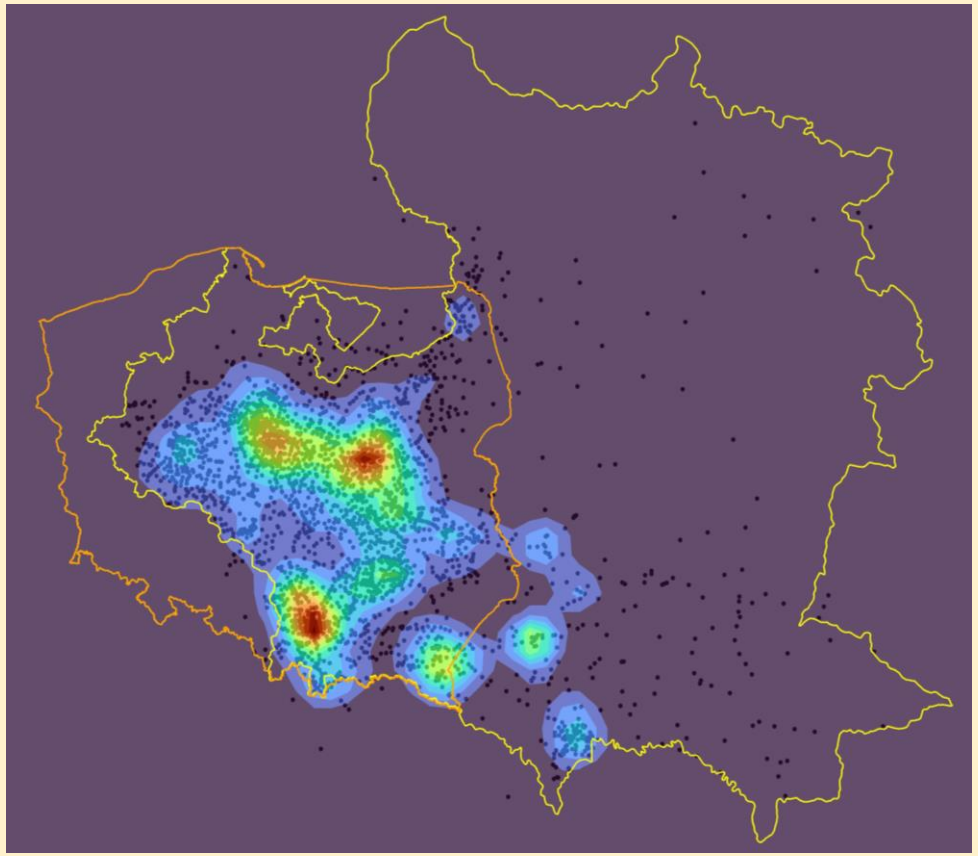


Recognition of sample biases

Kernel density based on main destinations of Kolbergs' travels



Kernel densisty based on locations of tunes in Kolberg



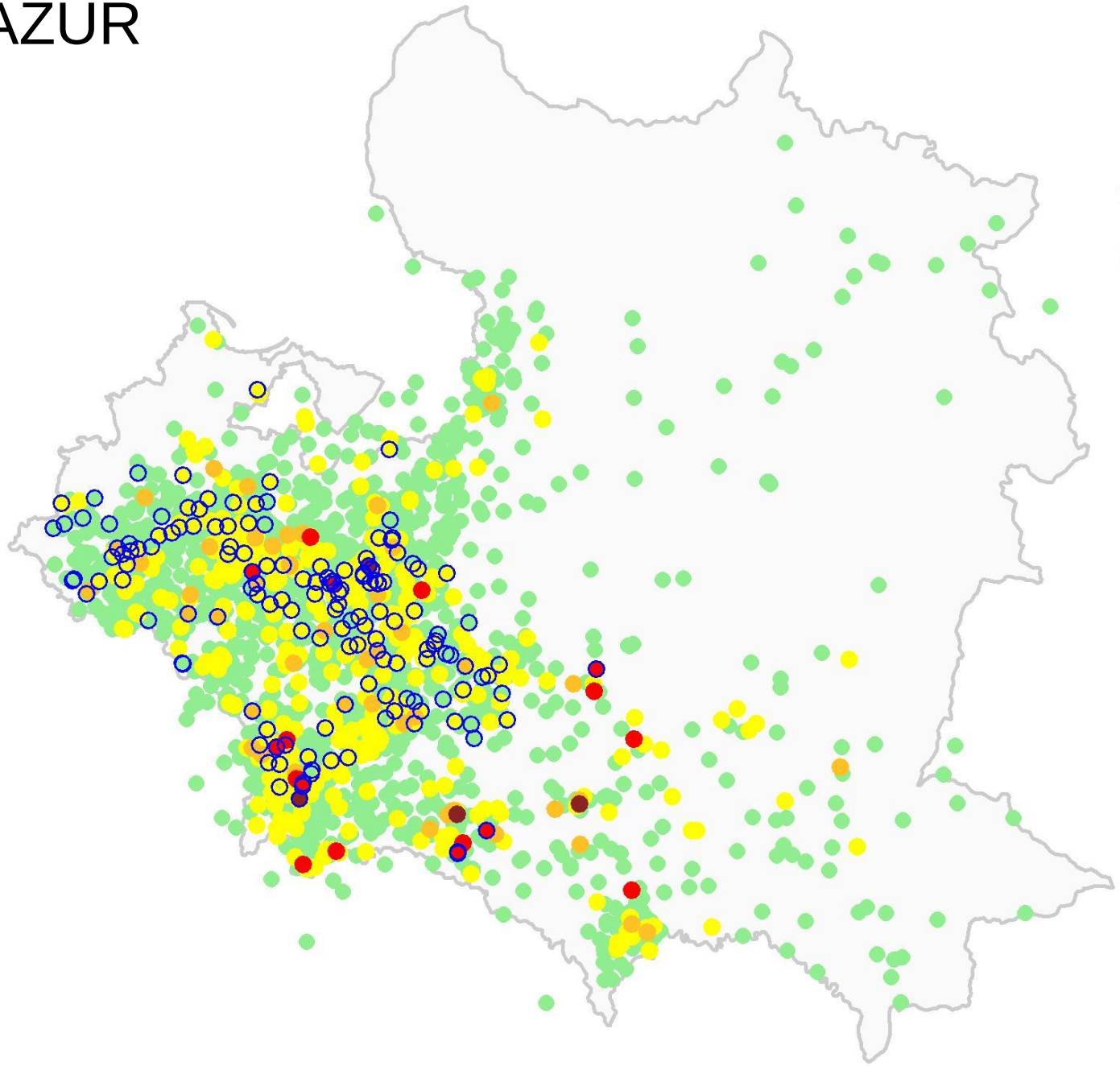
Measure of non-compliance with sample biases

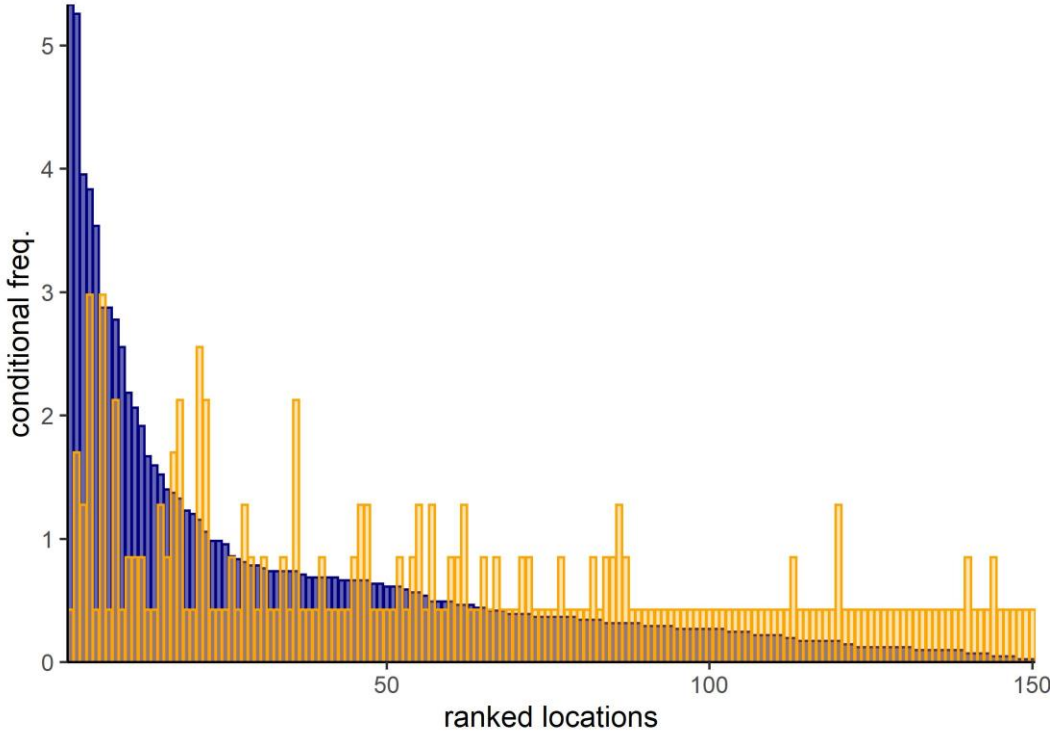
- Cross-entropy - a measure how close two distributions are

$$H(p,c) = -\sum_{l=1}^{N_l} p_l log_2(c_l)$$

→1180 „raw” categories, 56 categories with N >= 10

MAZUR



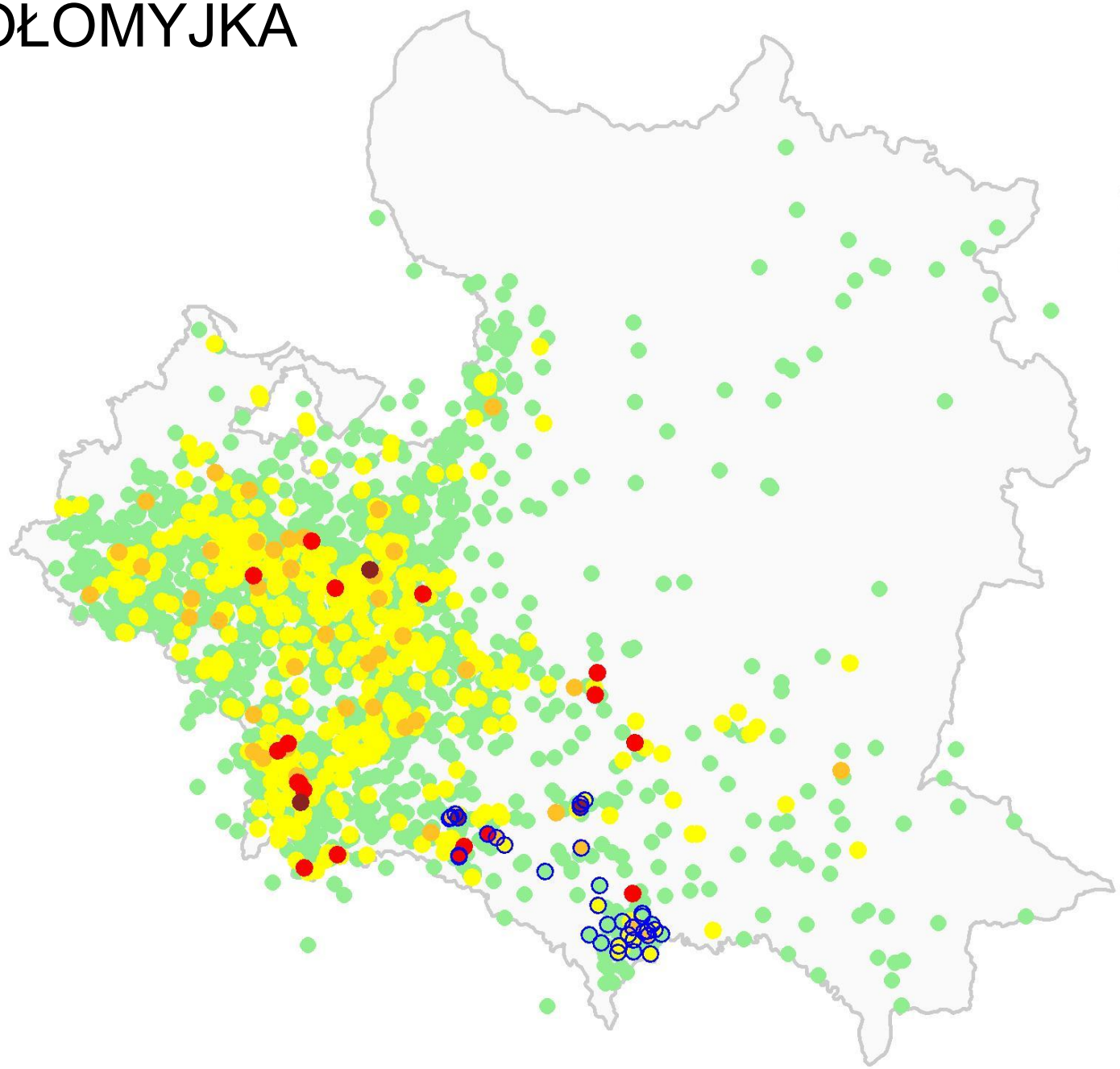


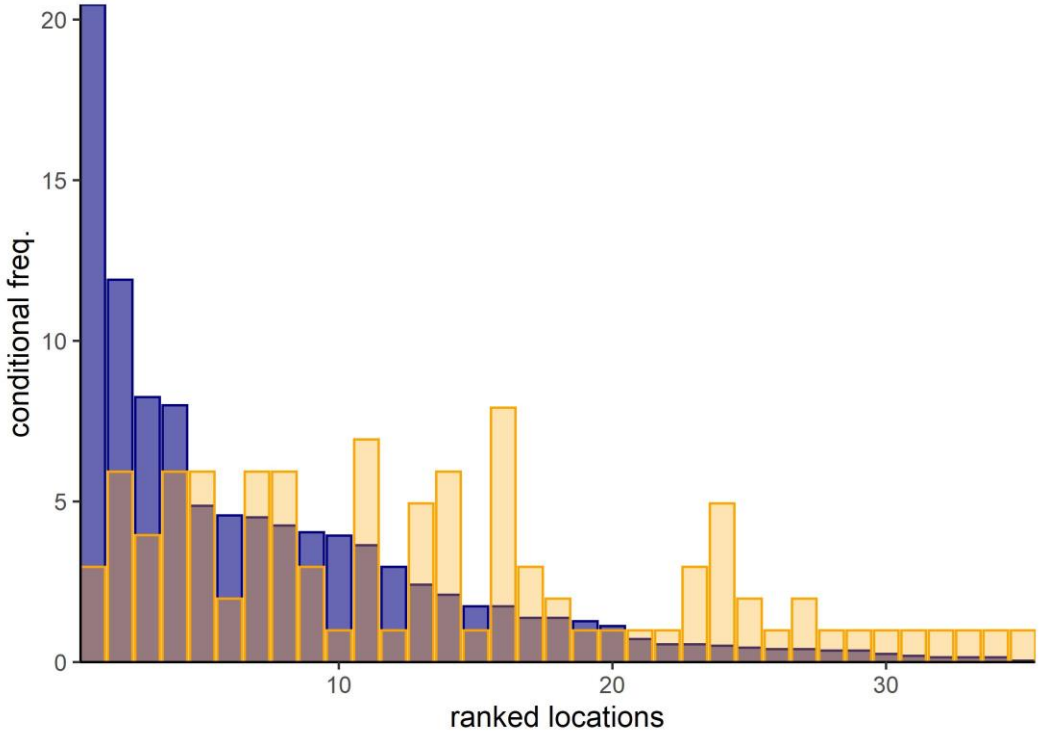
N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

CE = 3.06
N = 235

KOŁOMYJKA



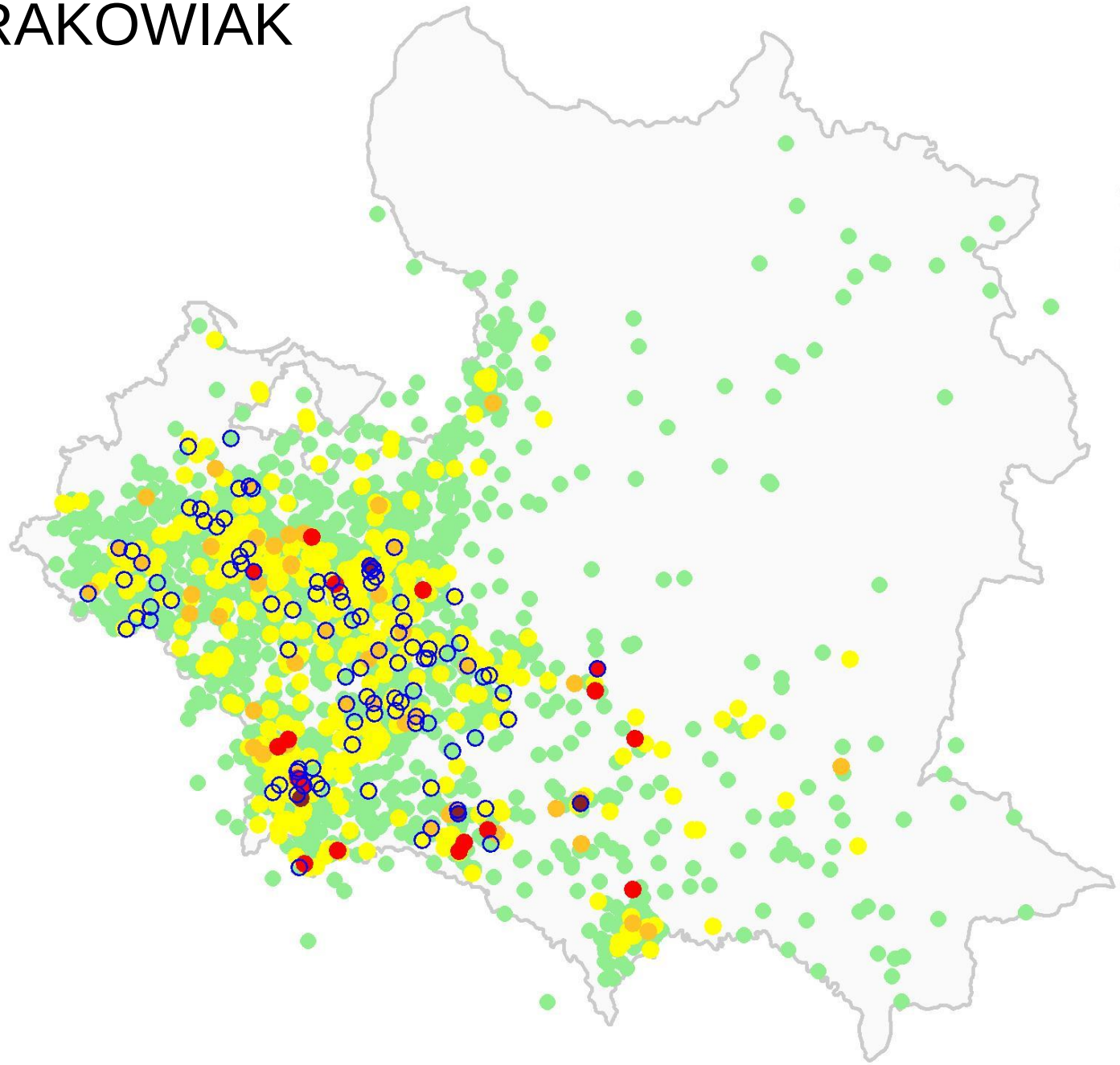


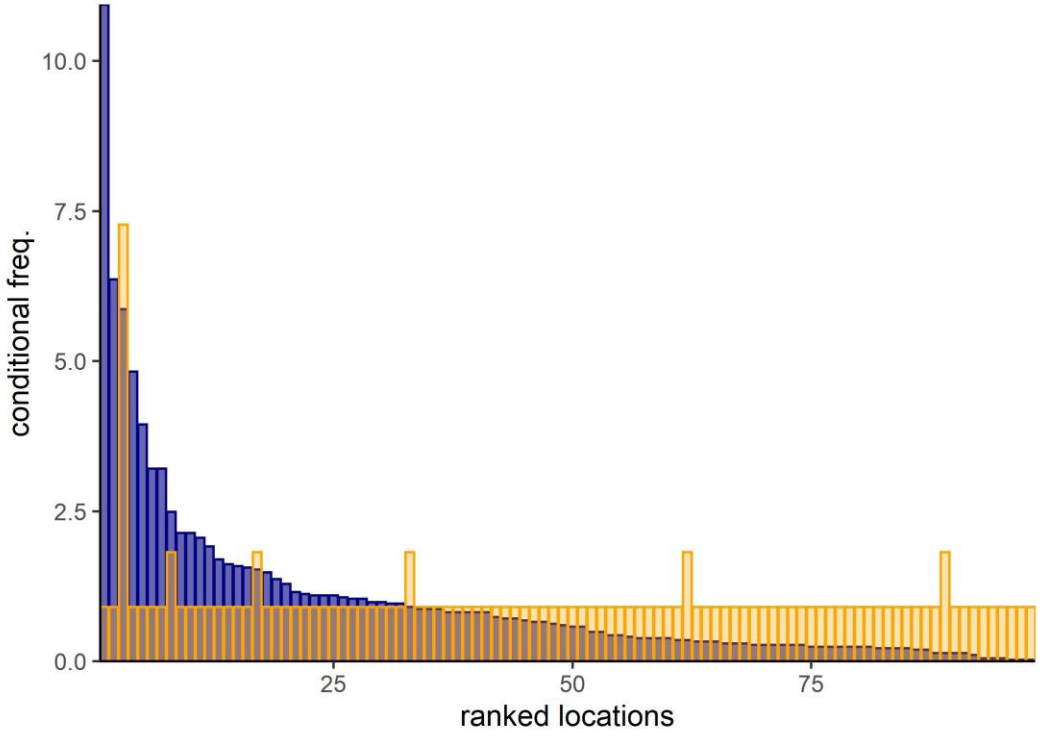
N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

CE = 2.23
N = 101

KRAKOWIAK



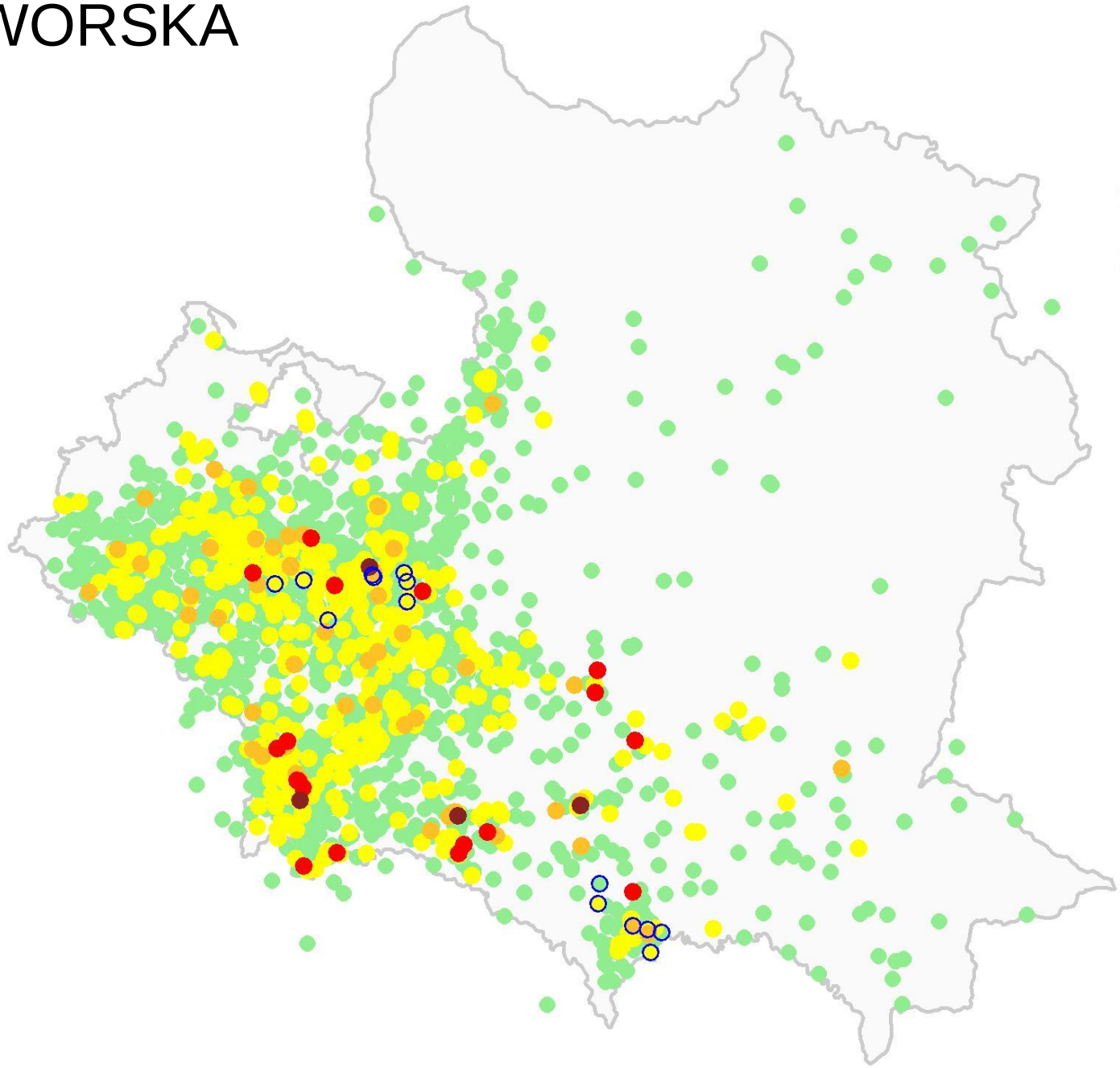


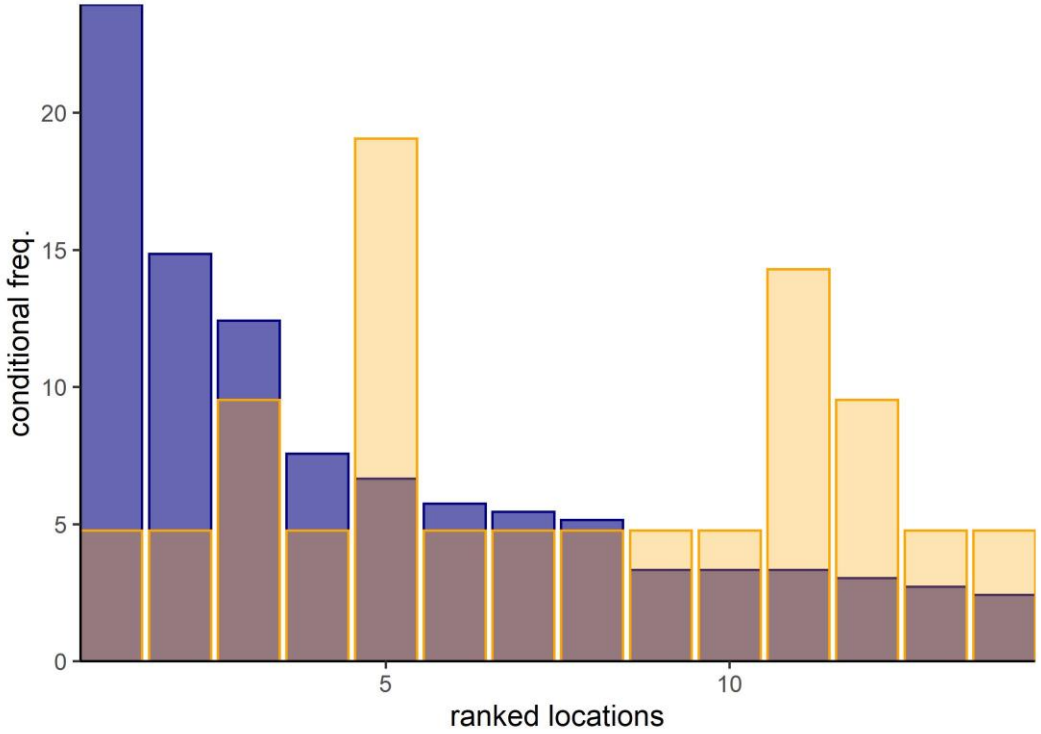
N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

CE = 1.45
N = 110

DWORSKA



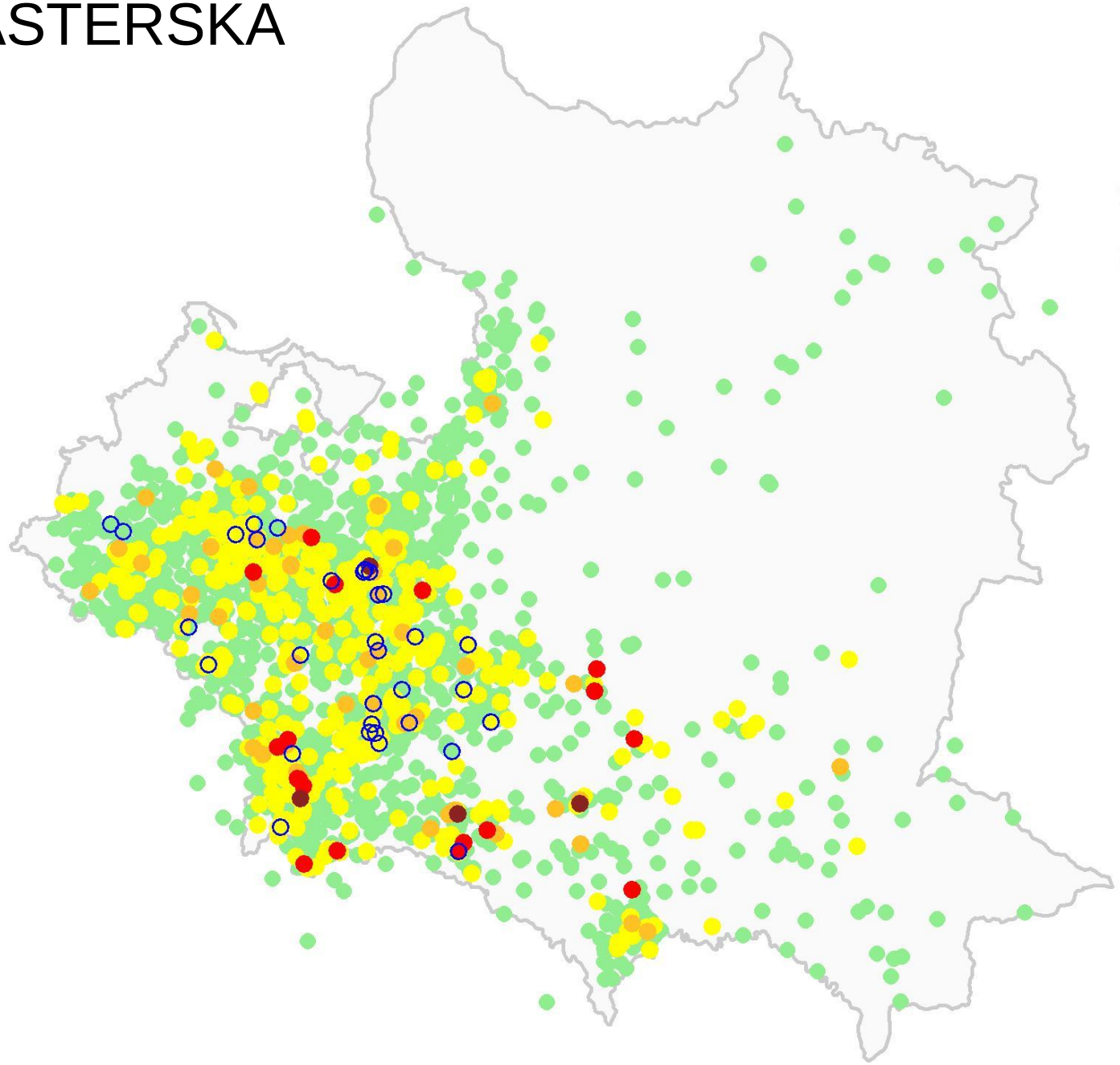


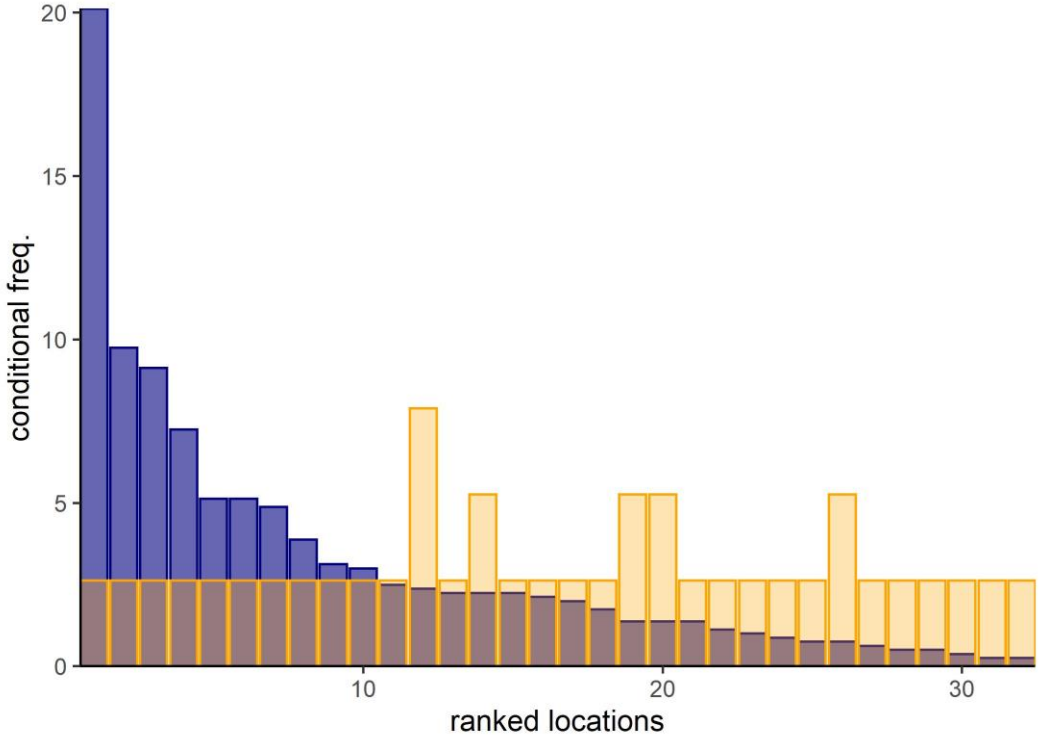
N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

CE = 0.09
N = 21

PASTERSKA



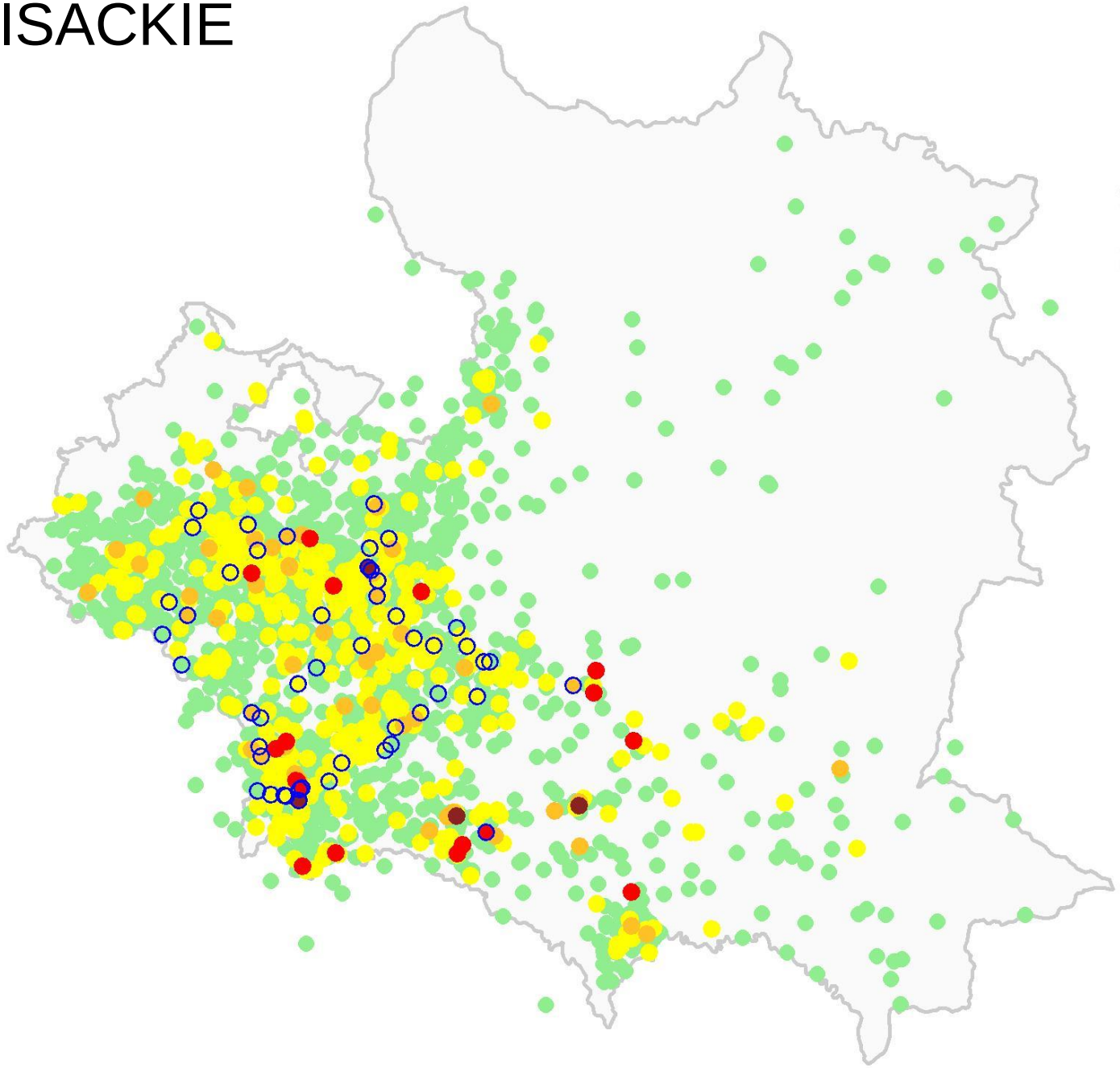


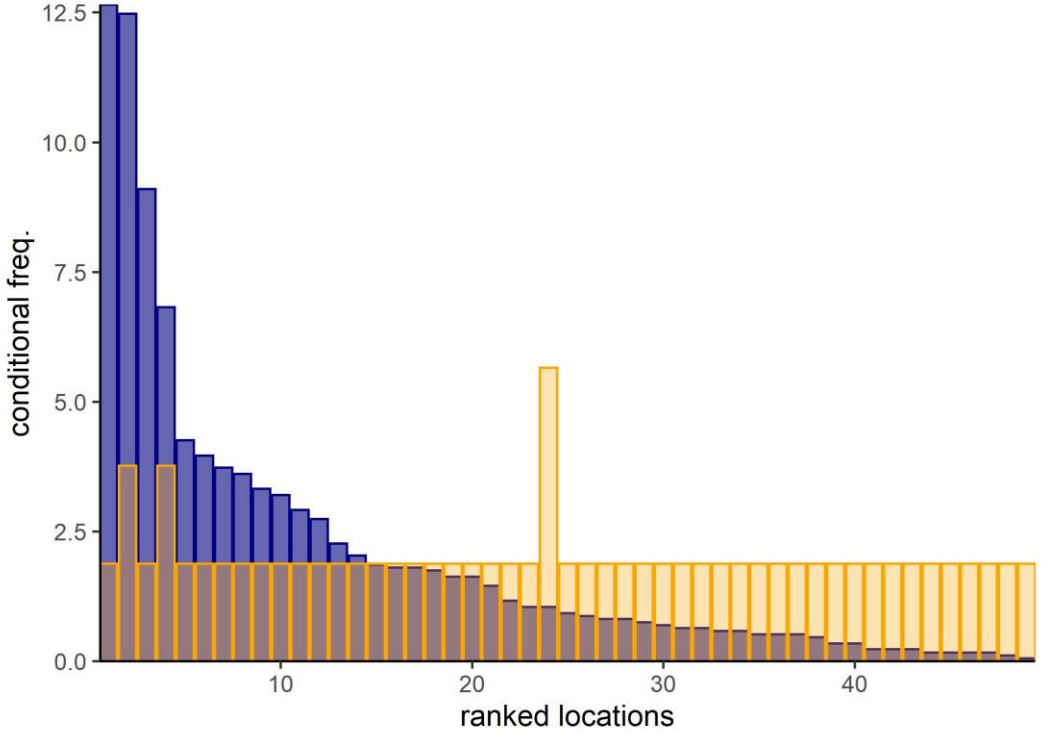
N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

CE = 0.19
N = 38

FLISACKIE

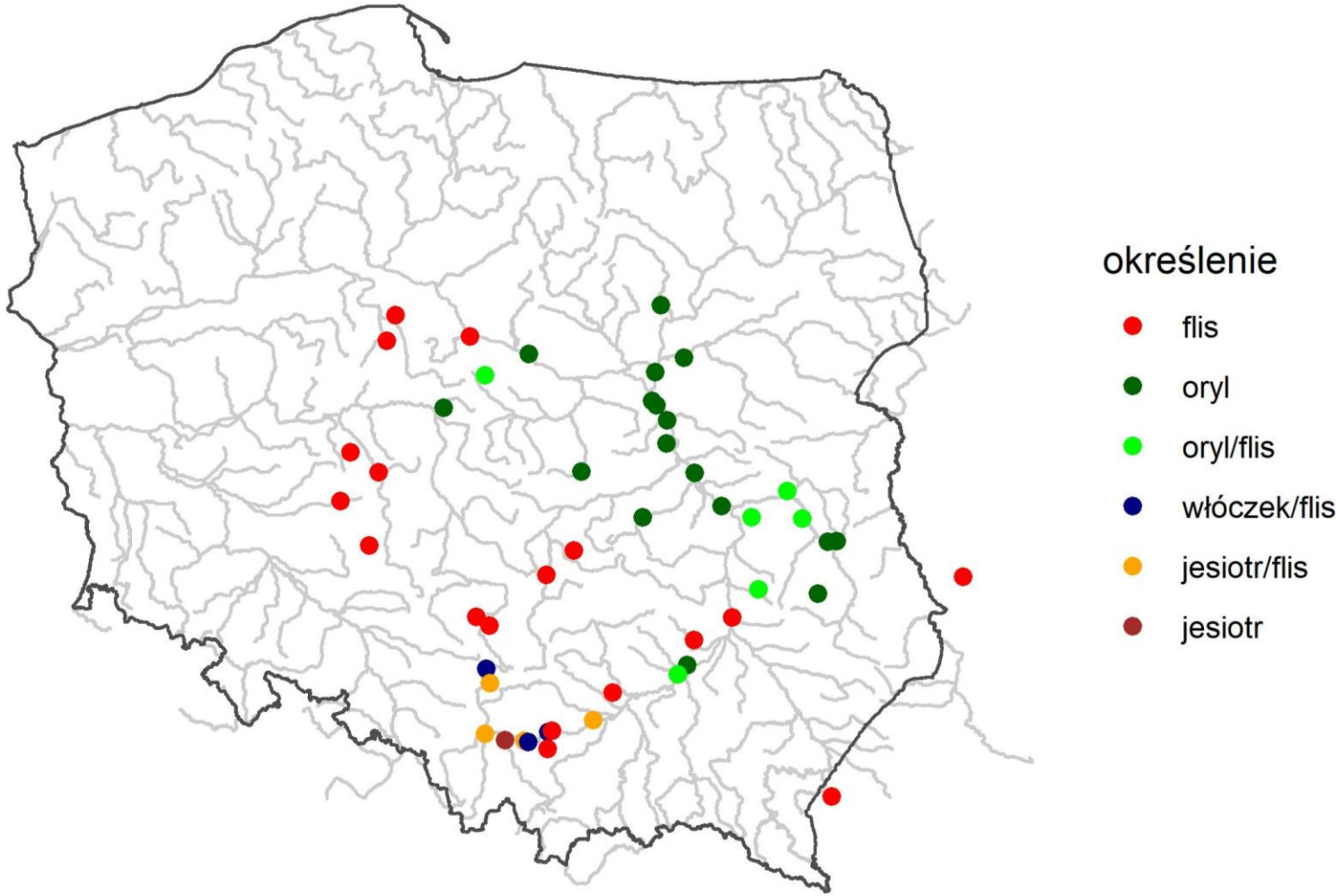




N of melodies

- 1-10
- 11-49
- 50-99
- 100-199
- 200+

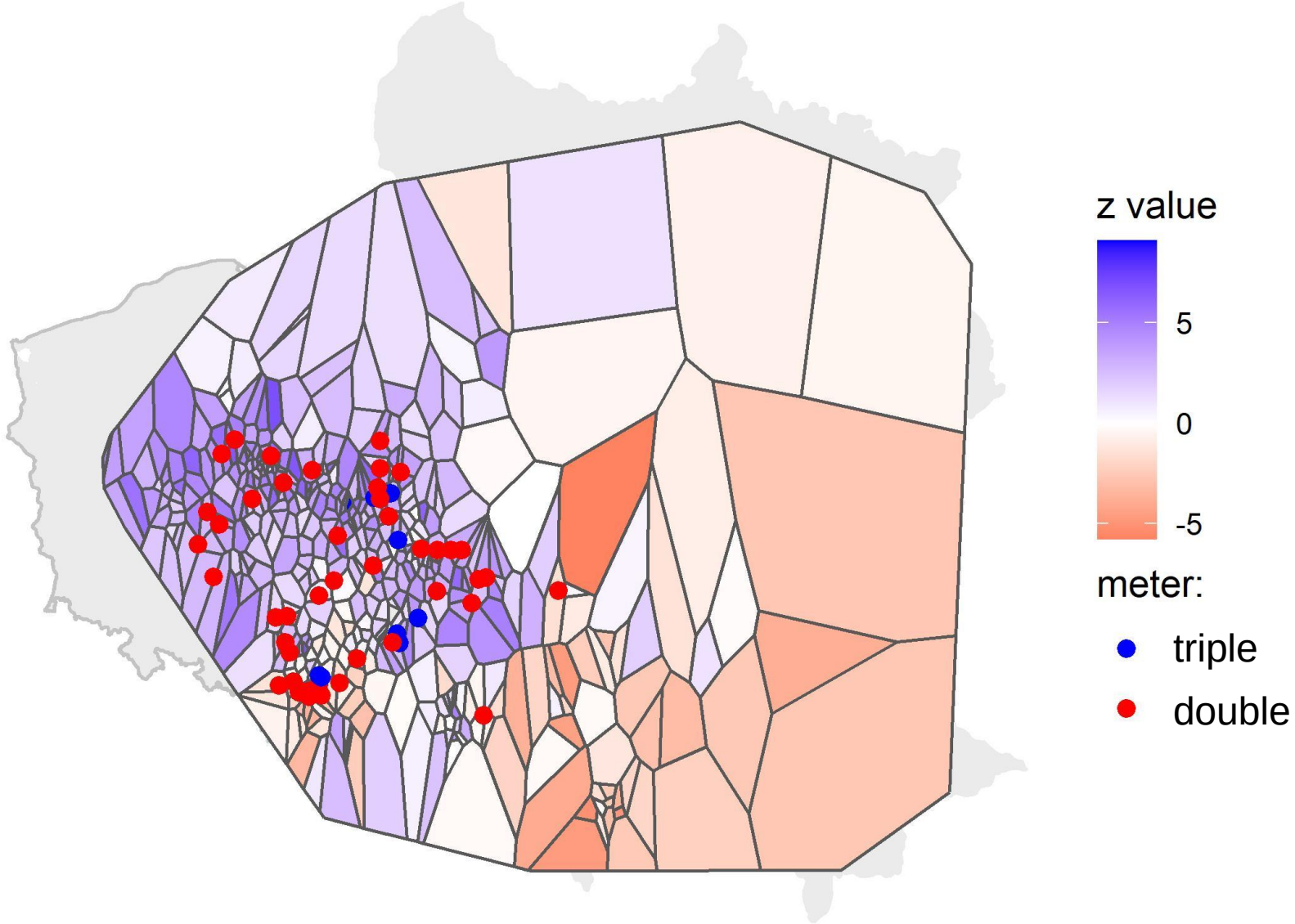
CE = 5.52
N = 49


określenie
flis
oryl
oryl/flis
włóczek/flis
jesiotr/flis
jesiotr

What if biases are not random?

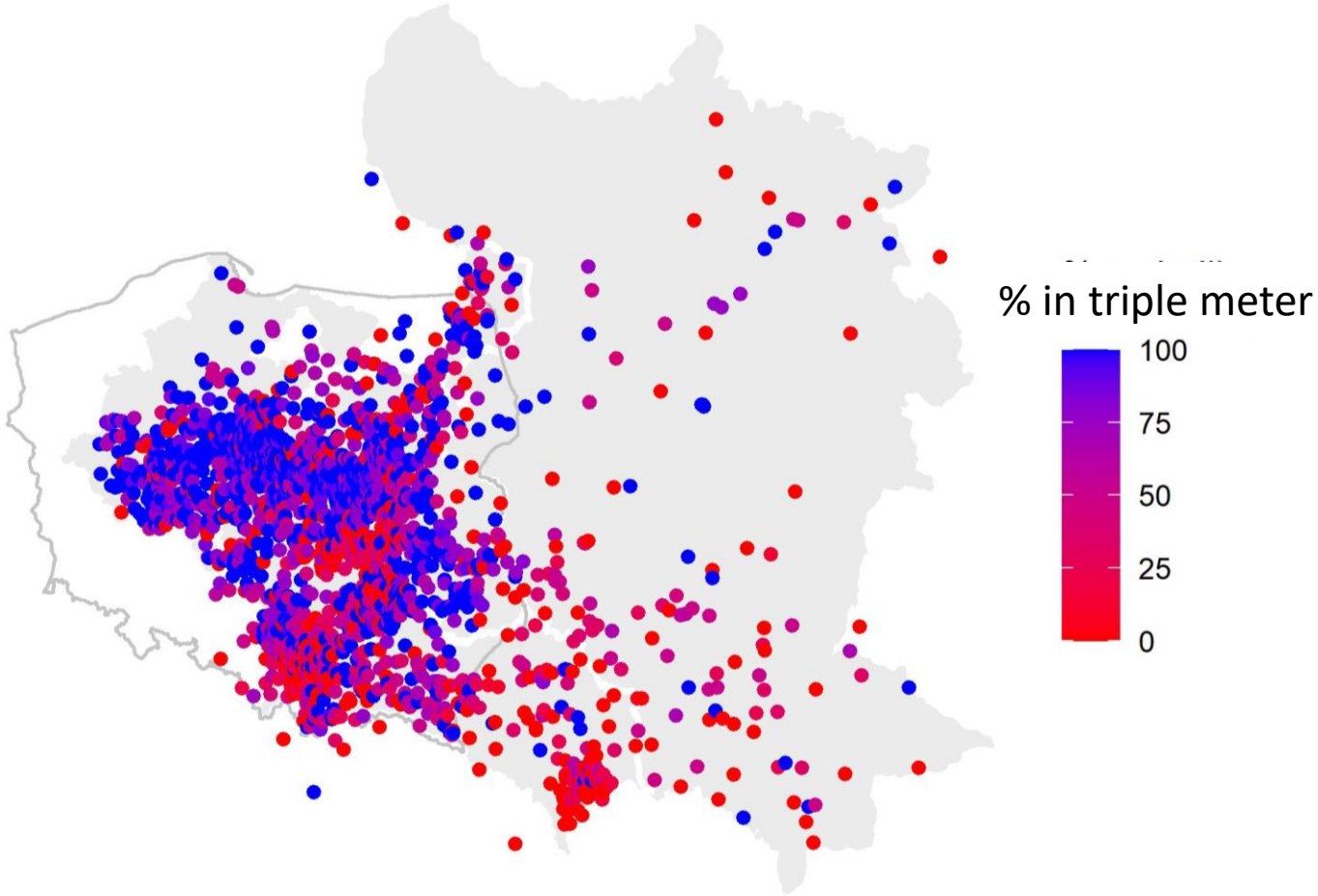
Musical features not related to sample biases

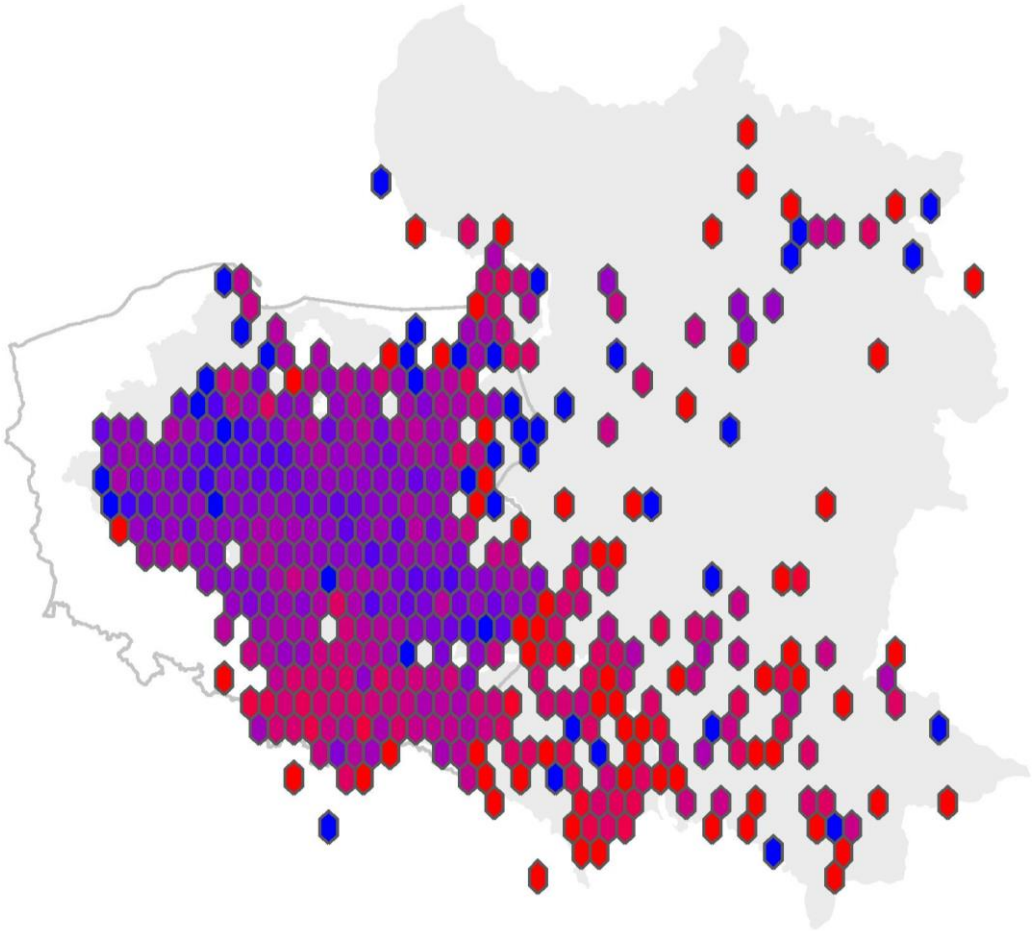
- meter (*every tune has some meter*)
- musical types (rhythm, structure, etc.)



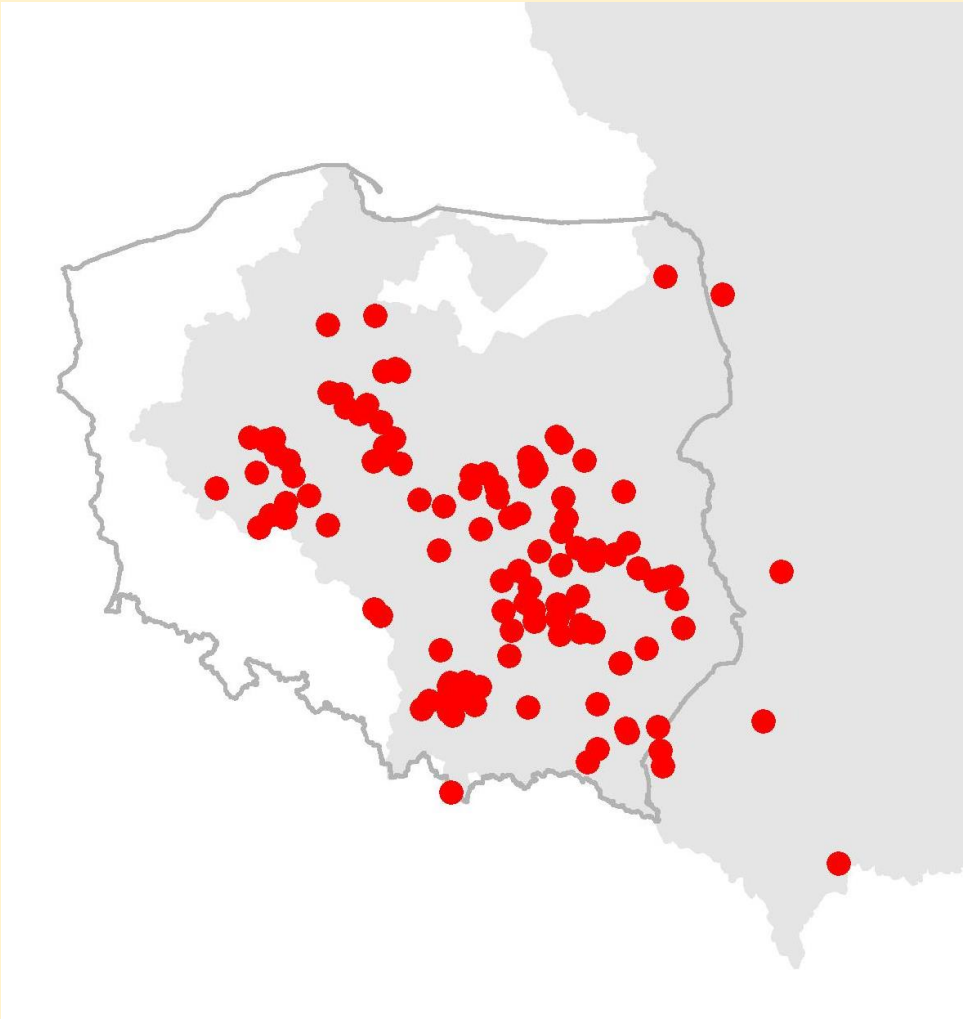
$$z = \frac{freq - 50\%}{\sqrt{\frac{freq(100\% - freq)}{N}}}$$

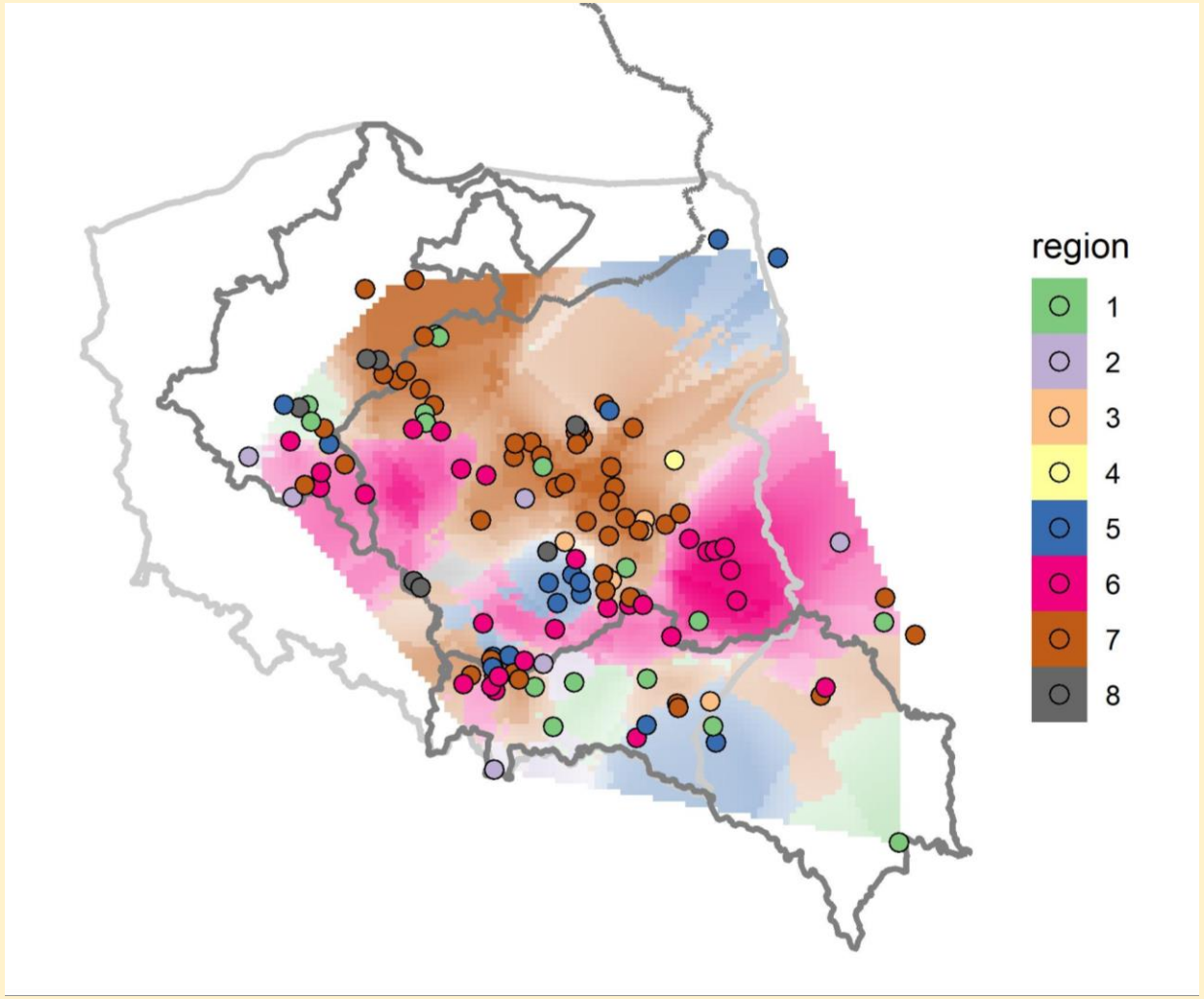
Tunes in triple-meter

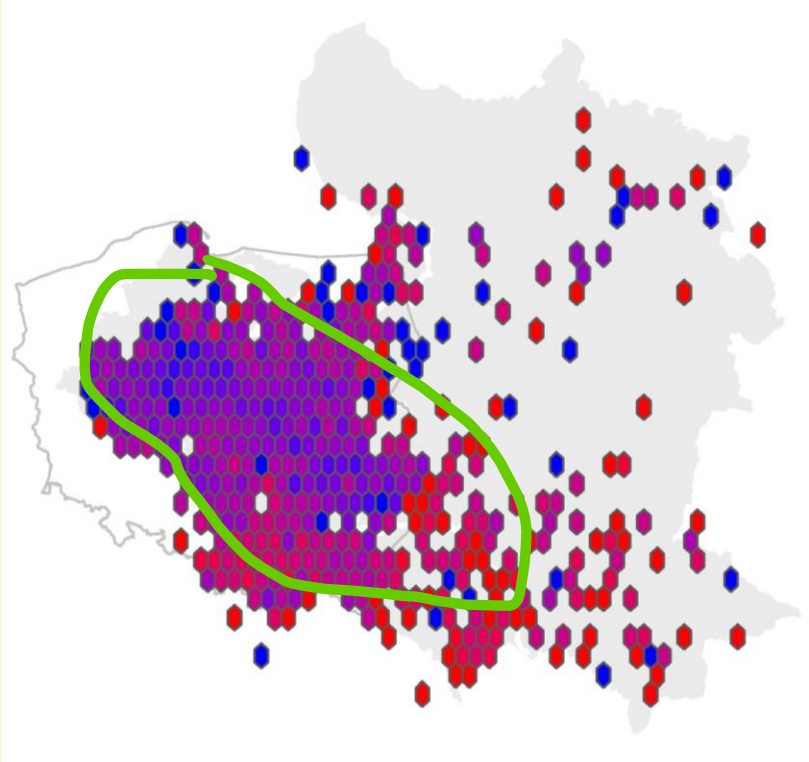


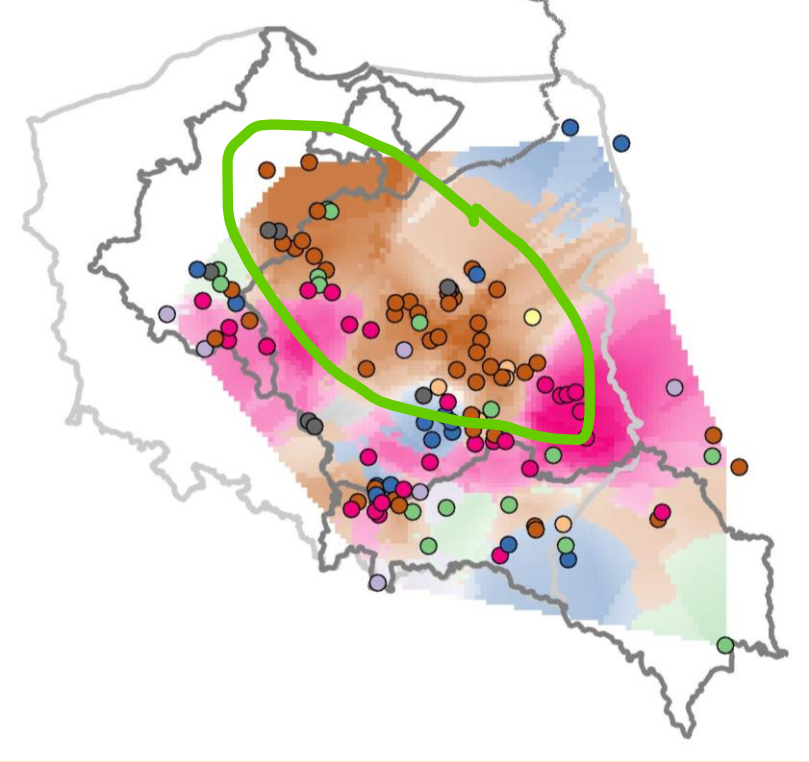


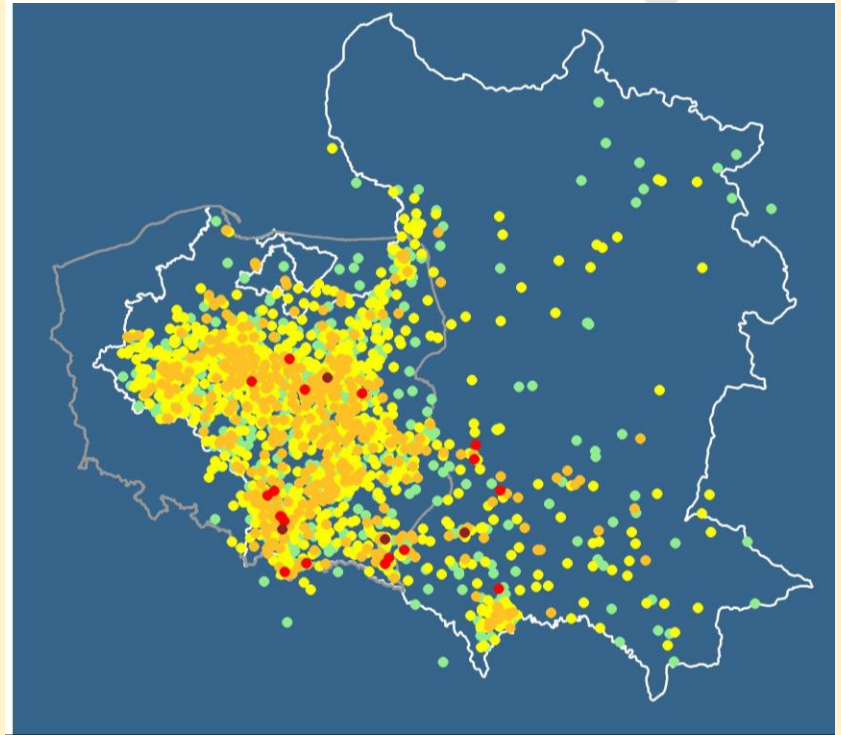
Krakowiaki

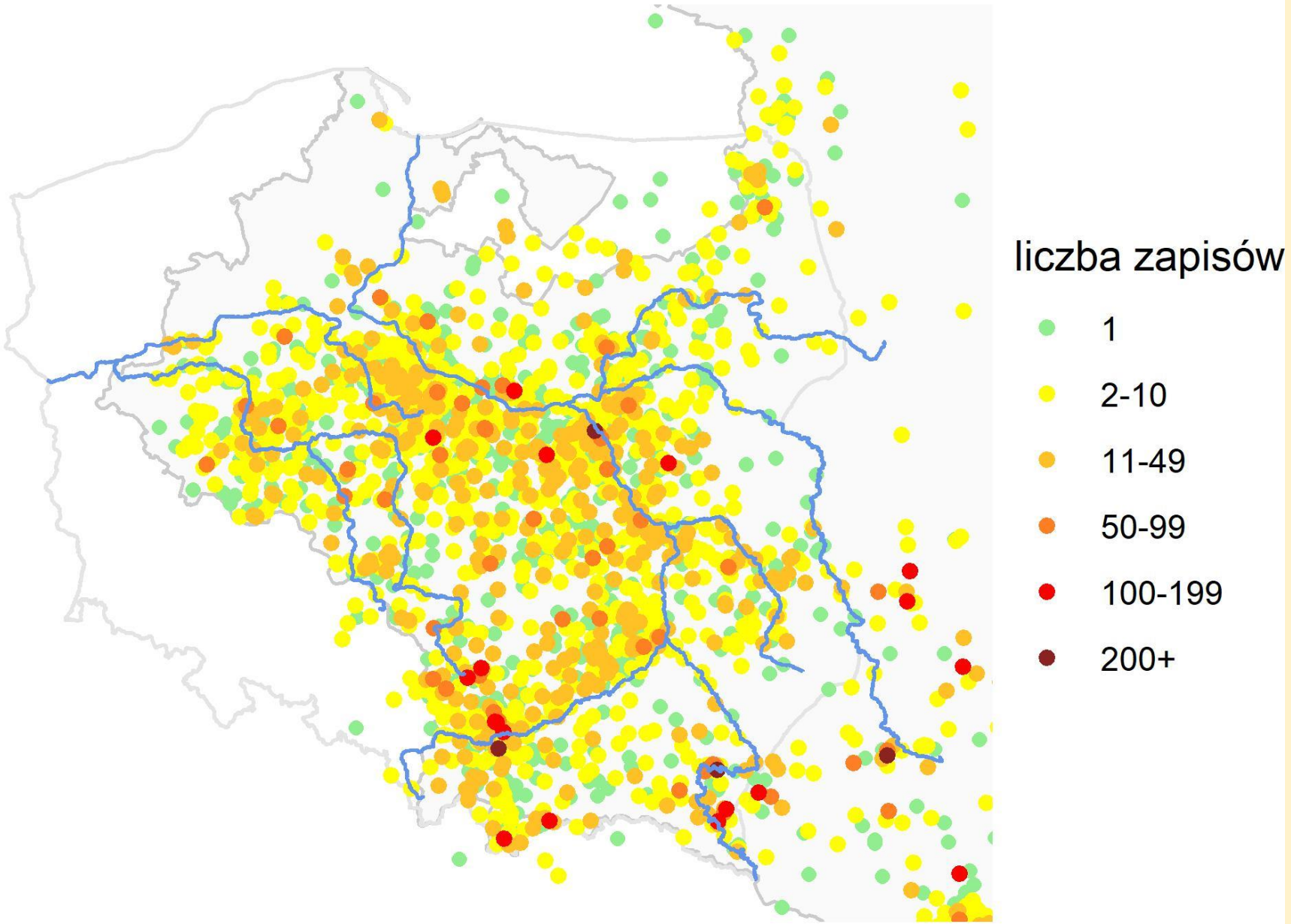


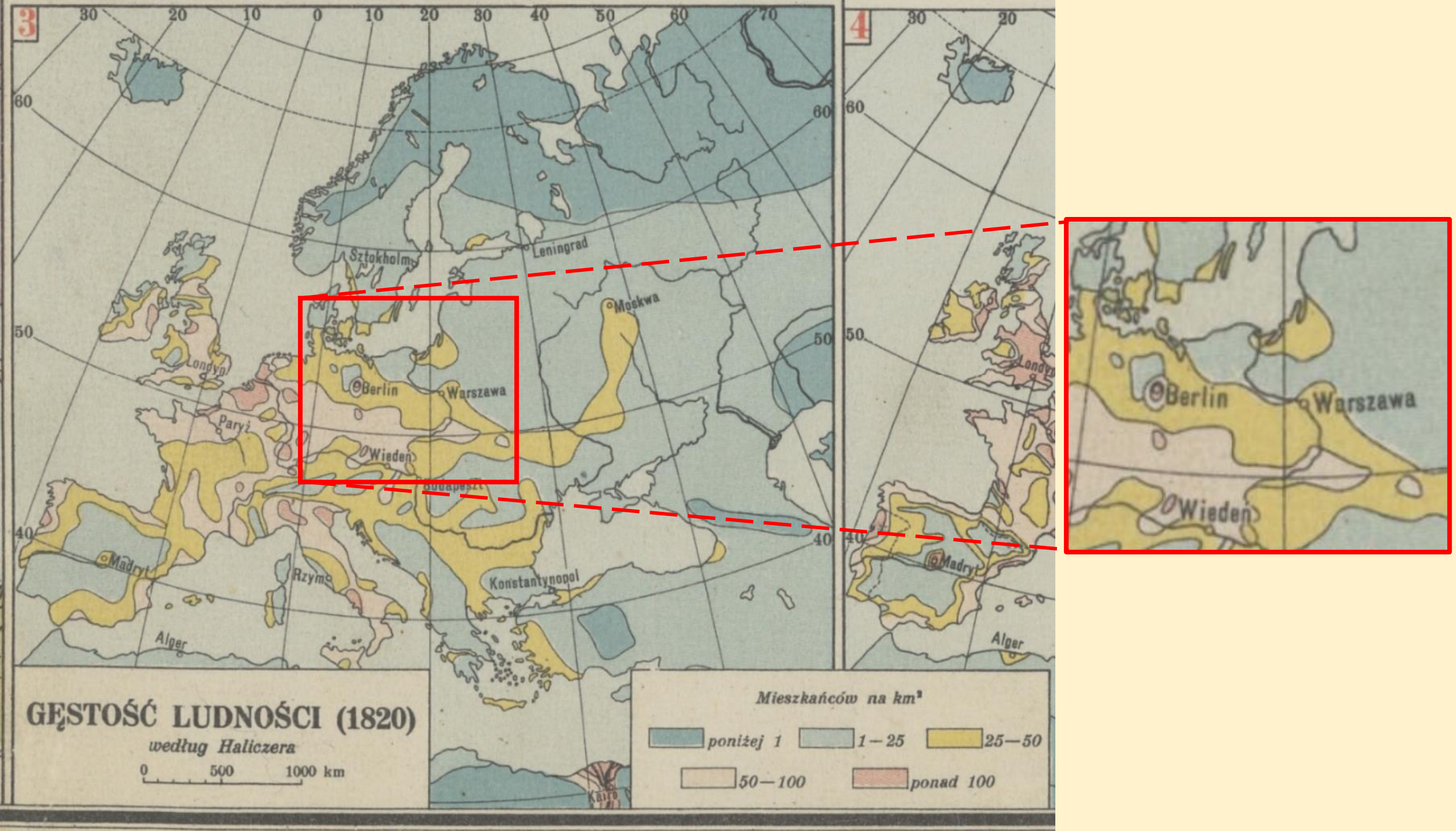




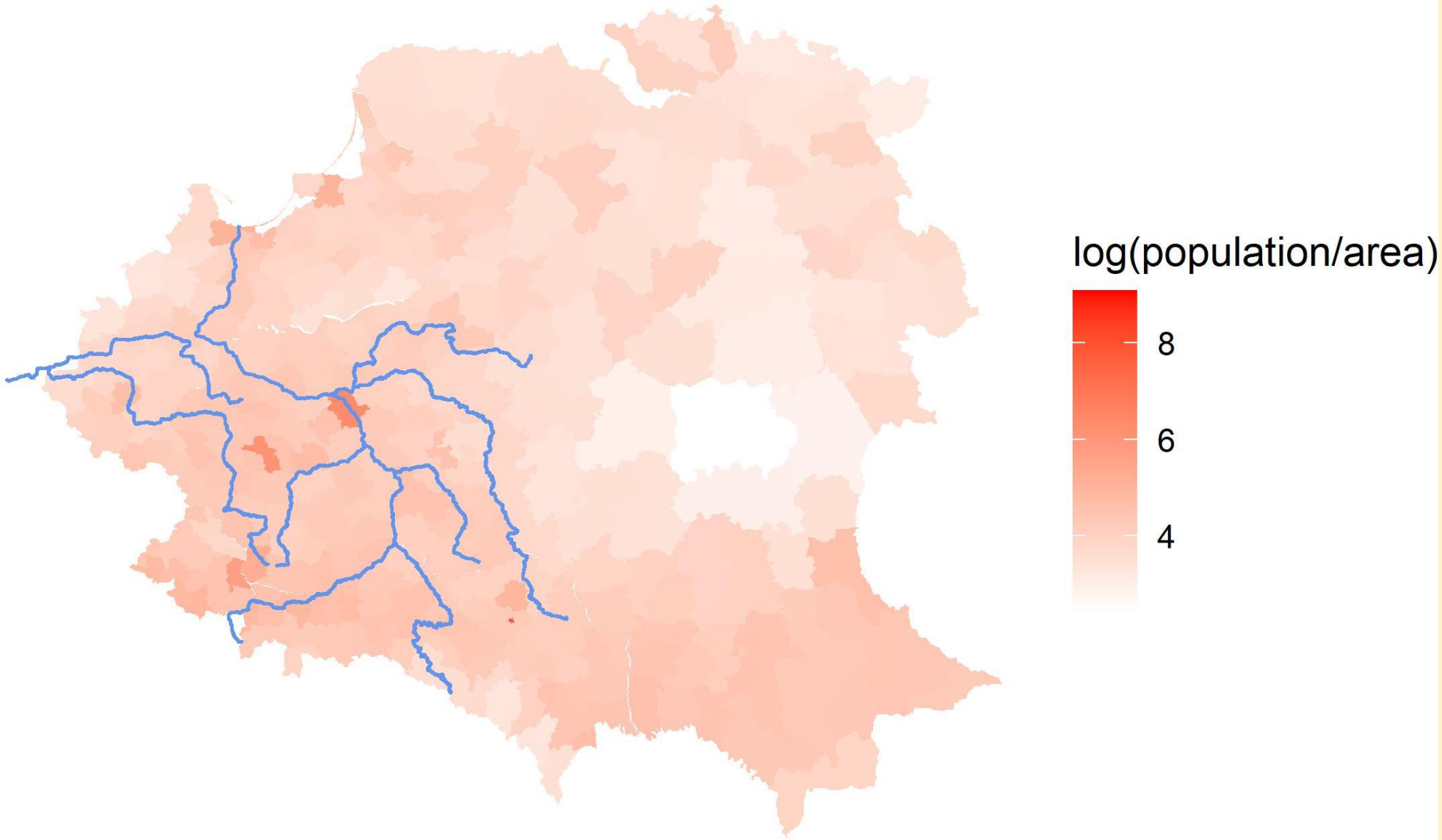


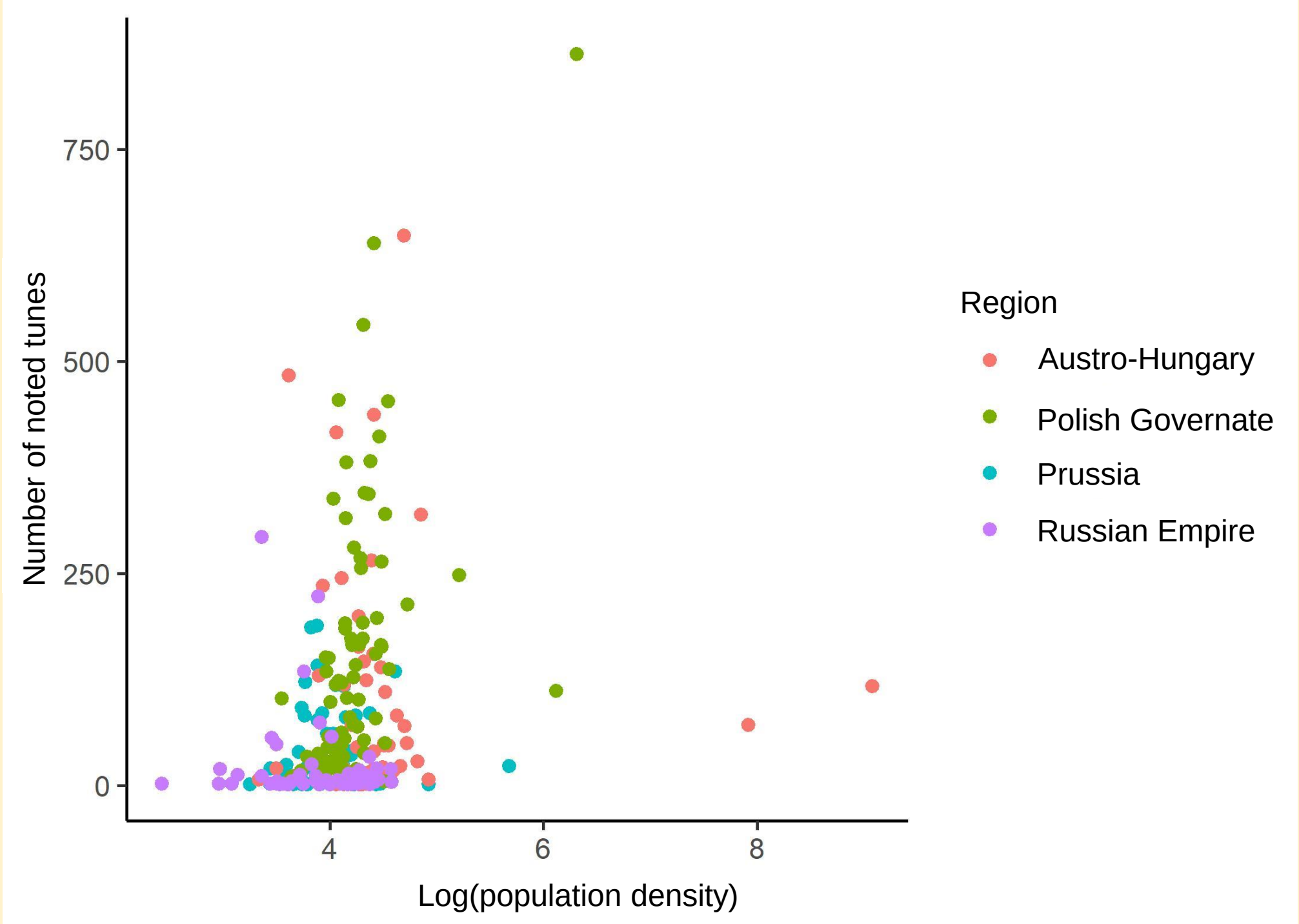



liczba zapisów
1
2-10
11-49
50-99
100-199
200+


3
4
Sztokholm
Leningrad
Moskwa
Londyn
Berlin
Warszawa
Paryż
Wiedeń
Budapeszt
Madryt
Rzym
Konstantynopol
Alger
Kairo
GĘSTOŚĆ LUDNOŚCI (1820)
według Haliczera
0 500 1000 km
Mieszkańców na km²
poniżej 1
1—25
25—50
50—100
ponad 100

POPULATION DENSITY 1871/1869/1897




Number of noted tunes
750
500
250
0
4
6
8
Log(population density)
Region
Austro-Hungary
Polish Governate
Prussia
Russian Empire

In summary

Seeking for:

- surprising distributions, inconsistent with sample biases
- common spatial patterns, not resulting from sample biases

Further steps:

- automatic evaluation of „surprisingness” and selection most informative categories
- automatic evaluation of common cultural spatial patterns

I would be grateful for any feedback!

r.mista@uw.edu.pl