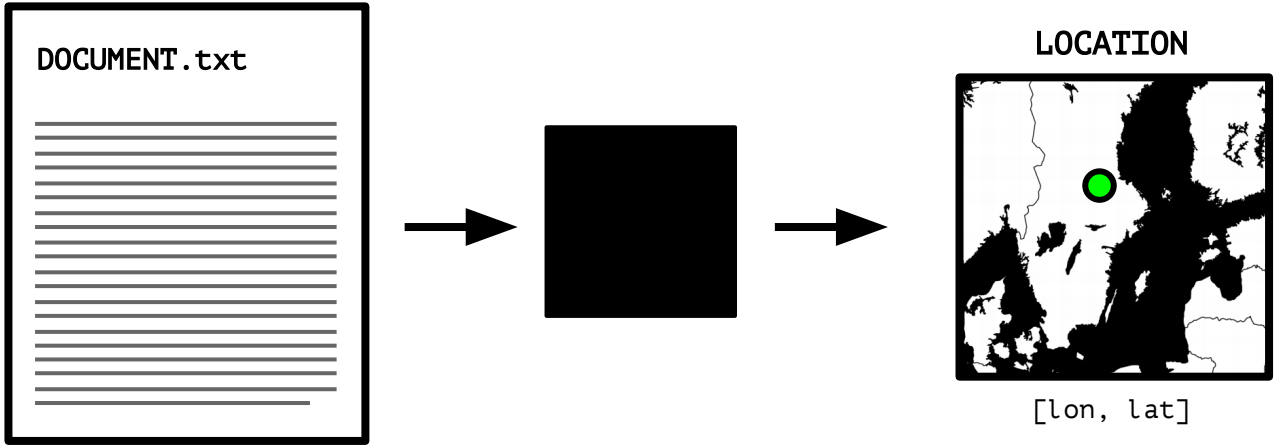
ME

- Max Berggren (@maxberggren)
- Undergraduate at KTH
  - Royal Institute of Technology
  - Stockholm, Sweden
- Working at Gavagai Labs
  - “Teaching computers to read”
  - SINUS project
    - Mapping the contemporary Swedish language

TASK

- Predict author location from text



SUMMARY

- Priedhorsky et al. 2014
- Improving on ↖
- Results
- Applications of the method:
  - Maps of Swedish dialects

SIGNAL OR NOISE

- Words carry information about position

SIGNAL OR NOISE

- Words carries information about position
- “I’m taking the tram now”
  - Tram in three Swedish cities

SIGNAL OR NOISE

- Words carries information about position
- “I’m taking the tram now”
- “God I hate Stockholm, people are so stressed”
  - Most Swedes have an opinion about the capital
  - I.e. speaking about Stockholm does not imply that you are there

SIGNAL OR NOISE

- Words carries information about position
- “I’m taking the tram now”
- “God I hate Stockholm, people are so stressed”
- “Oh lovely, lovely Falköping”
  - Mentioning a small town will make it likely that the author is from its proximity

TRAINING DATA

- Twitter gardenhose for tweets with geographic metadata
- ~2% of Swedish Twitter posts have latitude and longitude
- 4 429 516 tweets ≈ 630 MB
  - Gathered May to August 2014

TRAINING DATA

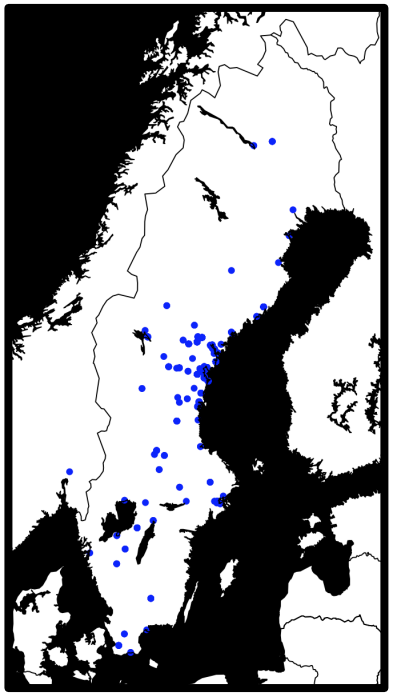
- 4 429 516 tweets with a coordinate

Lorem ipsum dolor tweet sit amet, twat consectetur adipiscing elit tweet tweet.	lat, lon

- Every uniqe n-gram can be mapped to a geographic distribution of coordinates

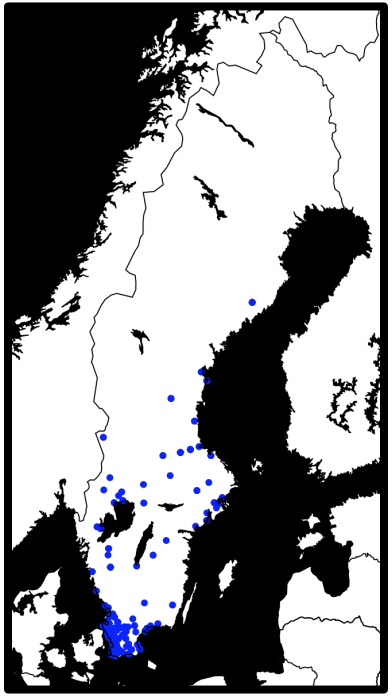
TRAINING DATA

Birsta



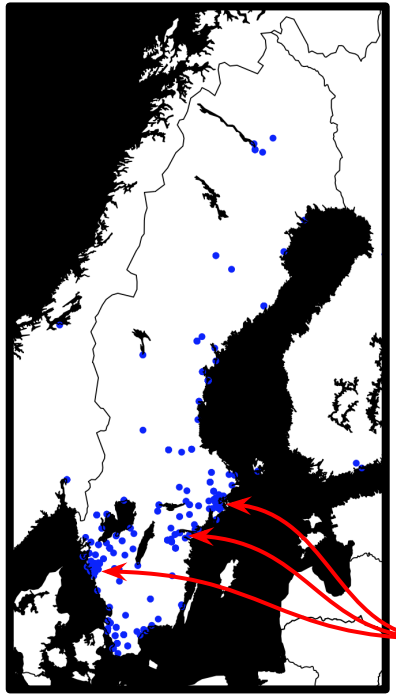
Big shopping centre in northern Sweden

litta



Southern slang for the swedish word for little

spårvagn



tram

. . .

(cities with tram)

MODEL

- Fit 2D Gaussian functions on the distrubitions (Priedhorsky, 2014)
- Gaussian Mixture Model
- Three Gaussians
- Python
  - Sci-kit learn

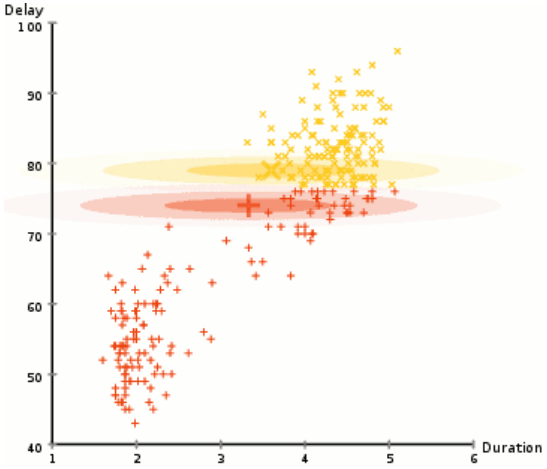


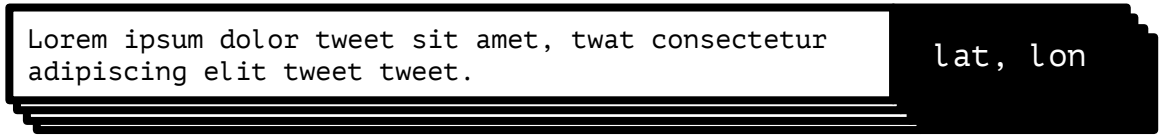
Image: sci-kit learn documentation

MODEL

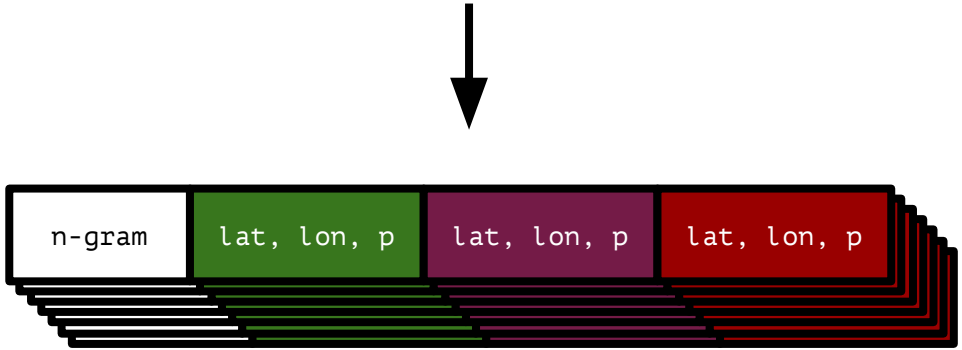
- Placeness: $p = e^{\frac{100}{-\rho}}$ ("peakiness")
- Where ρ is the log probability in the mean of the gaussian
- Log placeness of some words:

		Gaussian		
		1st	2nd	3d
Falköping		58	9	9
Stockholm		37	10	10
spårvagn	*"tram"*	36	18	15
och	*"and"*	16	15	9

MODEL



Tweets with metadata



“Bag-of-Gaussians”

PREDICTING

- Use n-gram Gaussians in centroid

PREDICTING

- Weighted arithmetic mean (centroid)

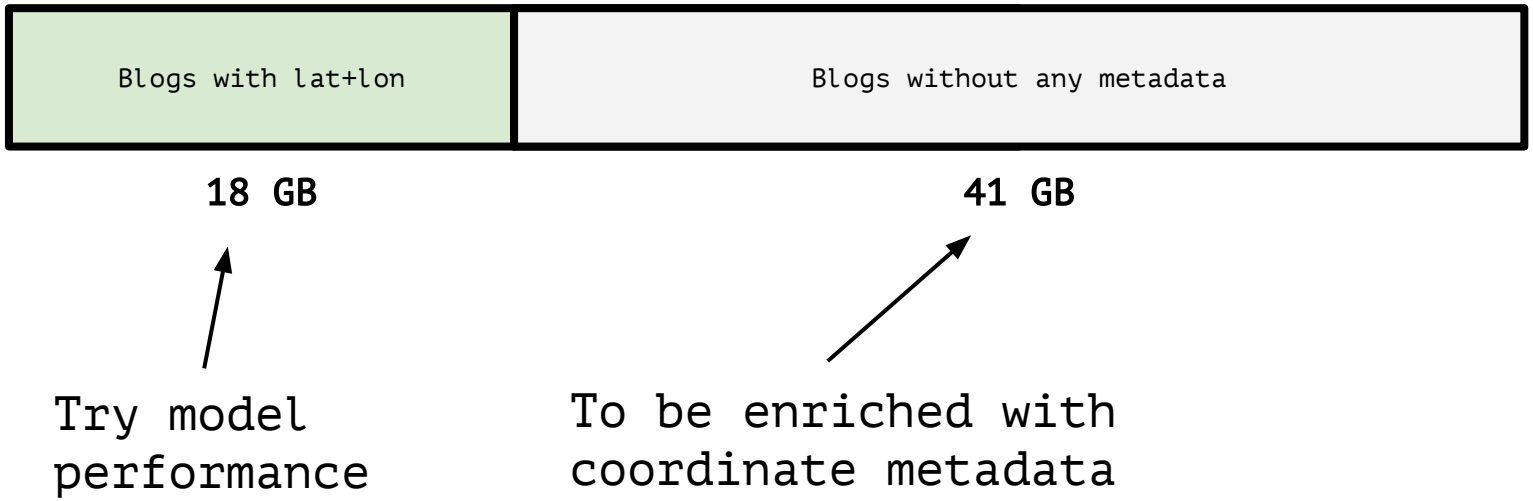
$$\overline{\mu}^i = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}^i \qquad \overline{p}^i = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}^i \qquad M = \frac{\sum_{i=1}^{n} \overline{\mu}^n \cdot \overline{p}^n}{\sum_{i=1}^{n} \sum_{j=1}^{3} p_j^n}$$

coordinates

- Where $\overline{\mu}^i$ is a vector of Gaussian means, $\overline{p}^i$ the Gaussians' placeness (weight), and n the number of Gaussians.

PREDICTING

- Priedhorsky et al. 2014 predicts tweets
- Enriching a Swedish dataset of blogs



RESULTS

Baseline

	Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
	$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
GAZETTEER	–	450	626	62	450	964	0.31	0.31
TOTAL	20	256	380	51	256	516	0.34	0.34

Centroid

- Gazetteer: baseline (most frequent city)
- $\tilde{e}$ median error, $\bar{e}$ mean error
- Total: Thresholded centroid
  - i.e. n-grams needs log placeness > 20
- error < 100 km (typical county)

FILTERING

- Use list of known places
- Find interesting distributional contexts
  - Window (6+0), (3+3) and (0+6)

1. Find most frequently occuring contexts
2. Rank contexts by ability to return words with high placeness (percentage of words with log(p) > 20)

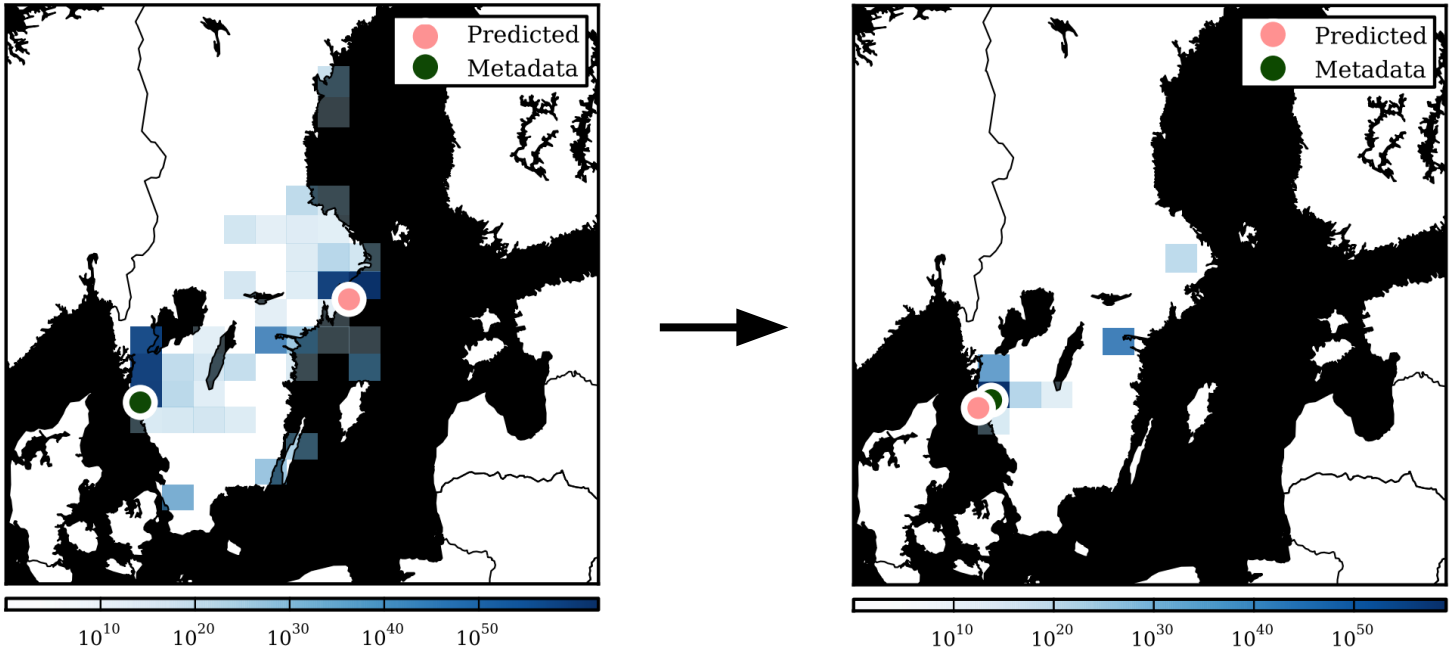
FILTERING

- Resulting regexpes
  - “go to <location>”
  - “off to <location>”
  - “live(s) in <location>”

- <location> filtered by $0.00008 \times N \leqslant f_{wd} \leqslant N/300$
  - N = length of text, f_{wd} = frq of word

FILTERING

- Preprocessing -> fewer Gaussians



RESULTS

	Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
	$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
GAZETTEER	–	450	626	62	450	964	0.31	0.31
TOTAL	20	256	380	51	256	516	0.34	0.34
FILTERED CENTROID	20	200	365	44	200	460	0.38	0.38

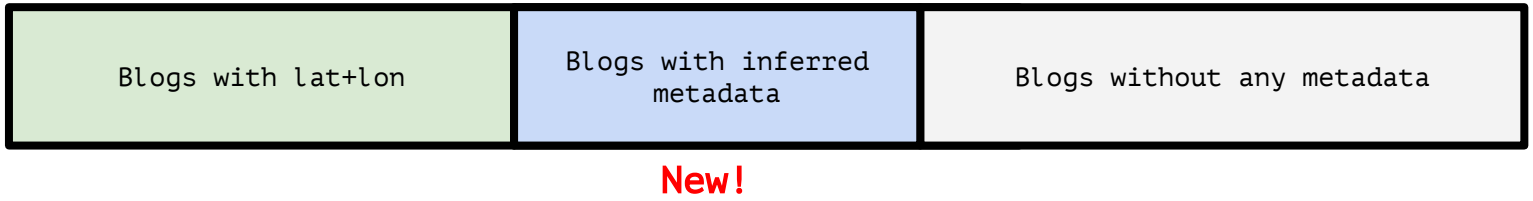
RESULTS

	Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
	$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
GAZETTEER	–	450	626	62	450	964	0.31	0.31
TOTAL	20	256	380	51	256	516	0.34	0.34
FILTERED CENTROID	20	200	365	44	200	460	0.38	0.38

	Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
	$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
FILTERED CENTROID	—	204	365	45	204	464	0.38	0.38
FILTERED CENTROID	10	204	365	45	204	464	0.38	0.38
FILTERED CENTROID	20	200	365	44	200	460	0.38	0.38
FILTERED CENTROID	40	145	333	32	145	396	0.44	0.32
FILTERED CENTROID	50	90	286	22	90	321	0.52	0.23
FILTERED CENTROID	60	70	271	13	70	330	0.53	0.04

MAPS

- Enriching a Swedish dataset of blogs

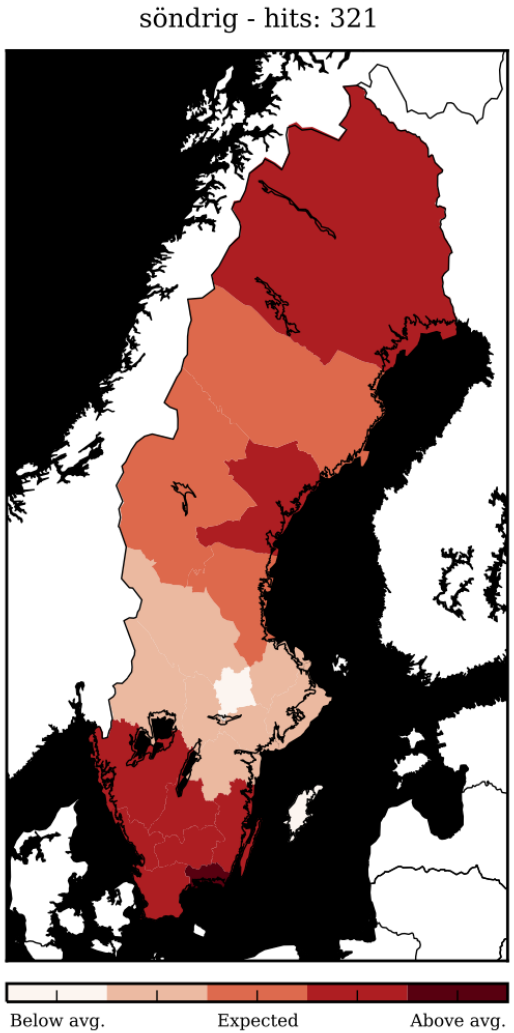


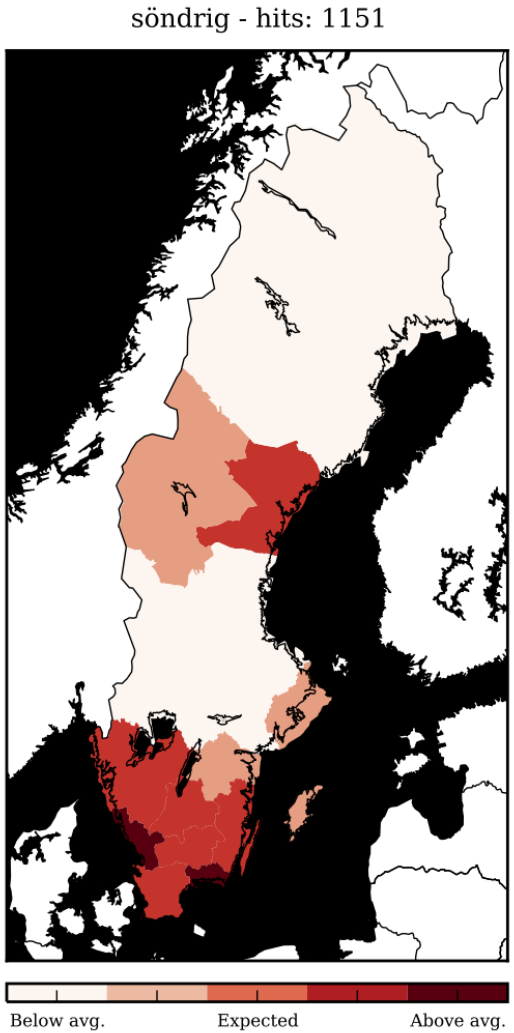
- Let’s query the dataset for words and see where people use them!

“BROKEN”

query on blog dataset **without** inferred lat lon

321 hits





query on blog dataset **with** inferred lat lon

1151 hits

goal:

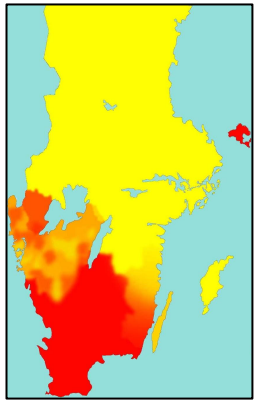


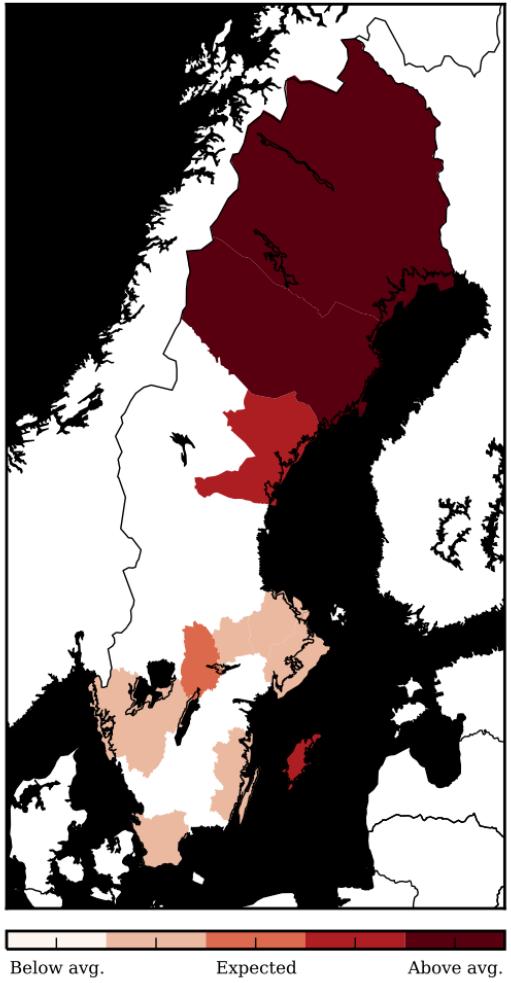
image: Mikael Parkvall

“NOT”

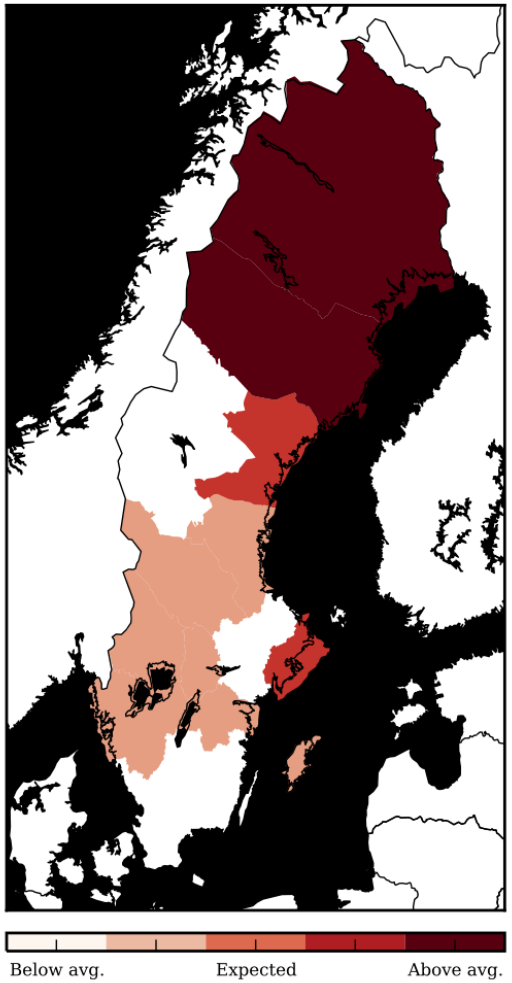
query on blog dataset **without** inferred lat lon

1646 hits

icket - hits: 1646



icket - hits: 4491



query on blog dataset **with** inferred lat lon

4491 hits

goal:

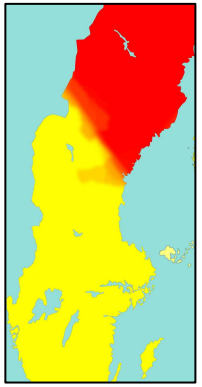
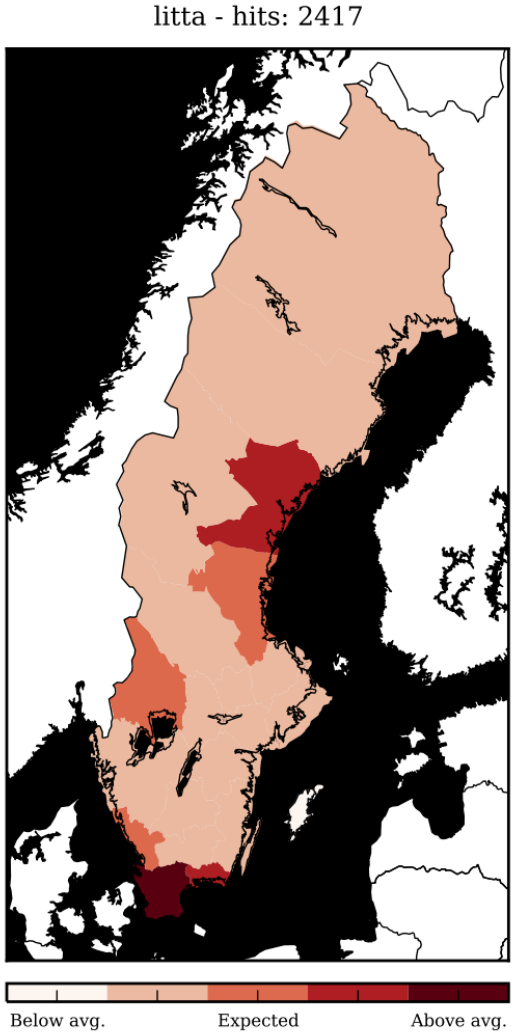


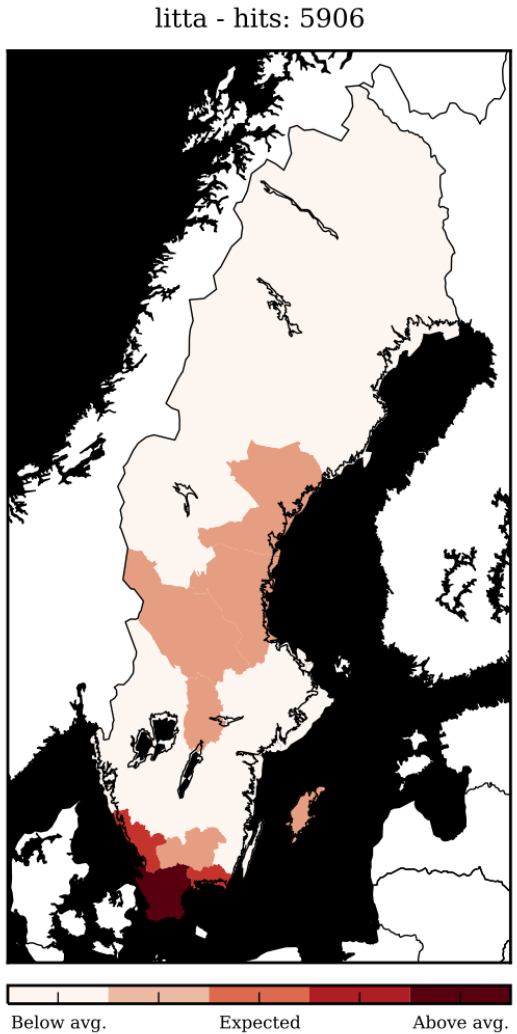
image: Mikael Parkvall

“LITTLE”

query on blog dataset **without** inferred lat lon

2417 hits

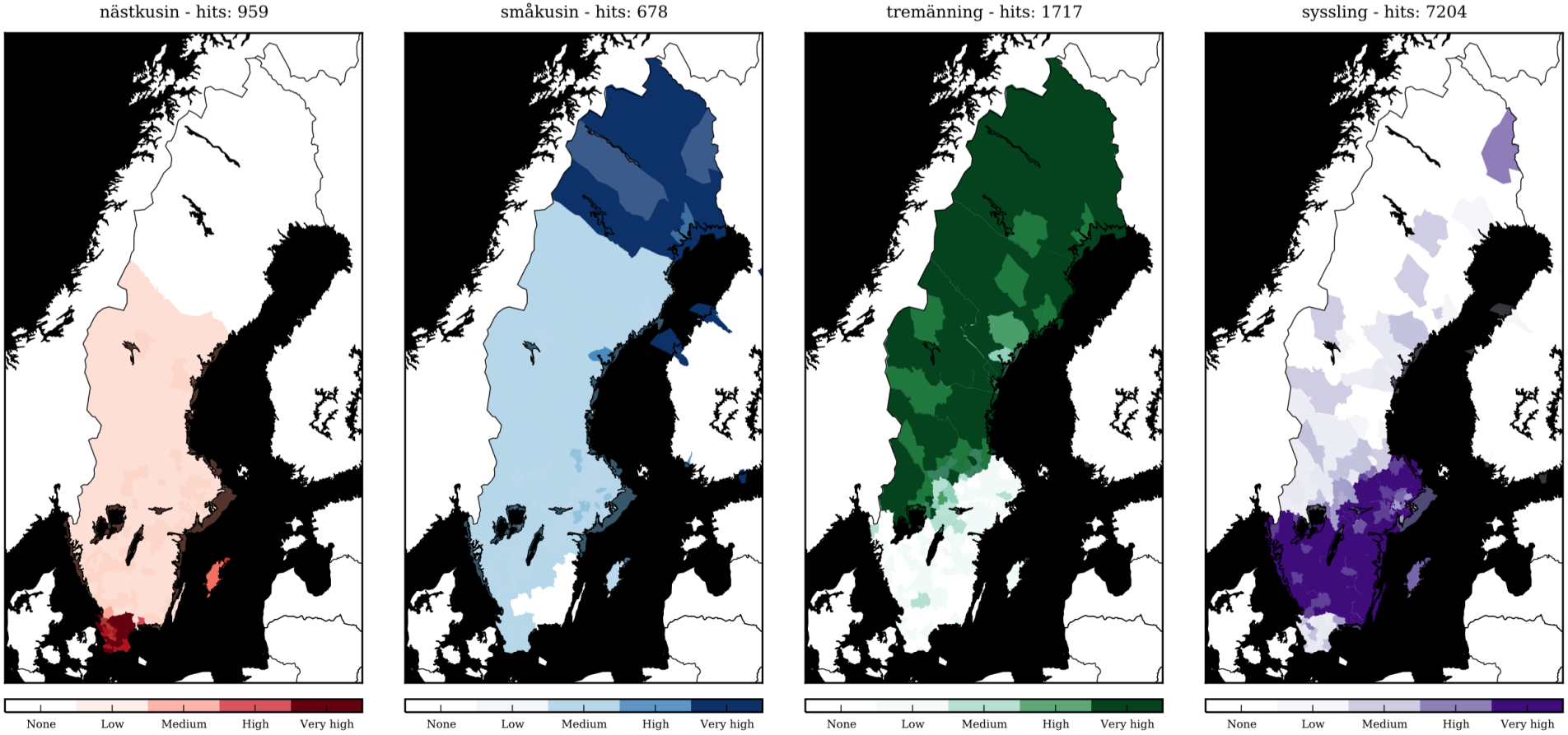




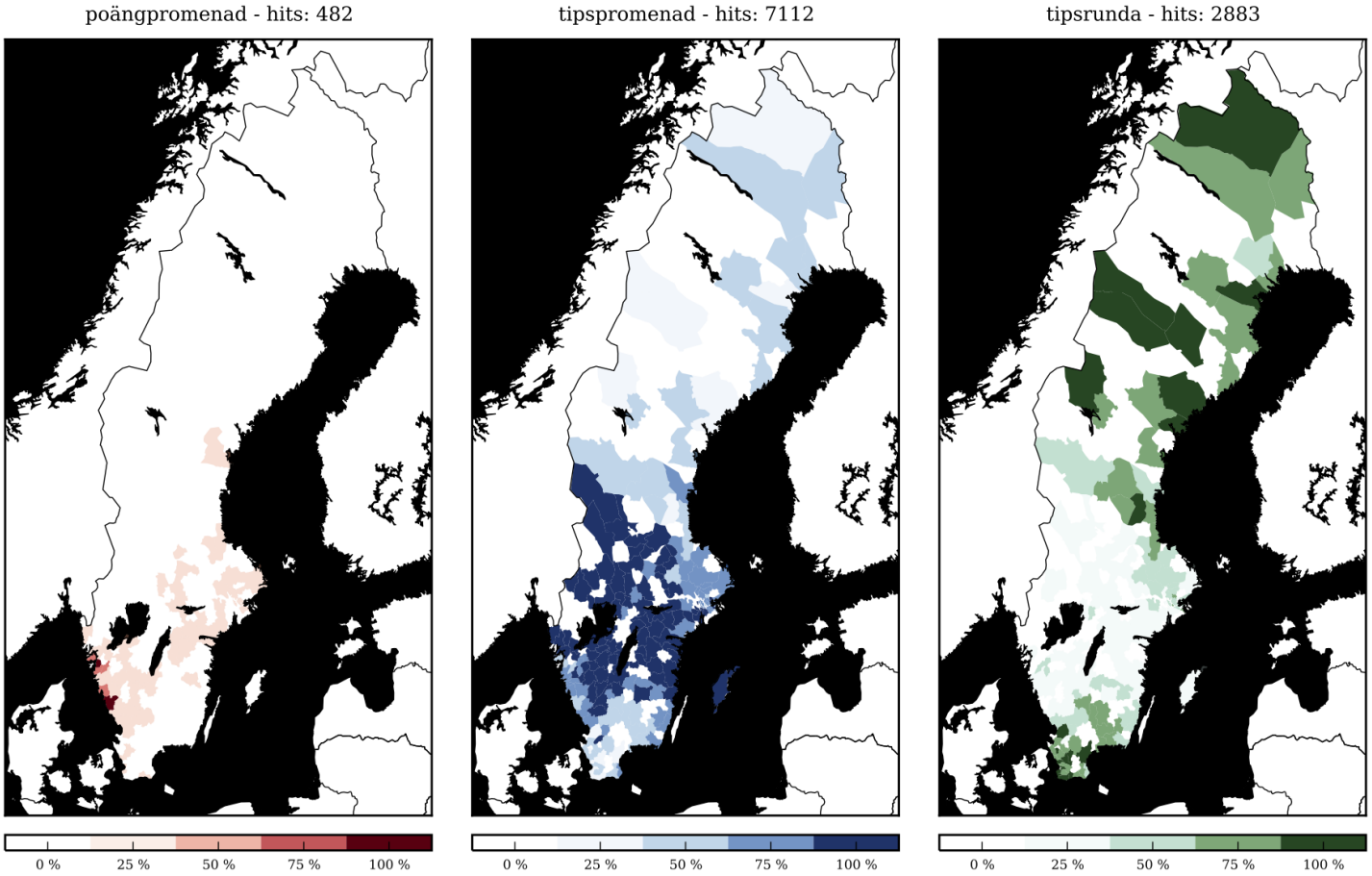
query on blog dataset **with** inferred lat lon

5906 hits

VARIATIONS OF “SECOND COUSIN”

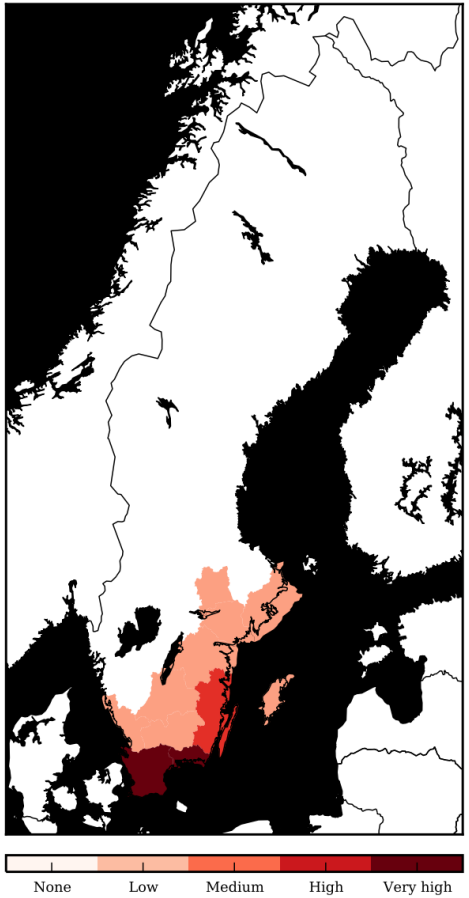


“TIPSPROMENAD”

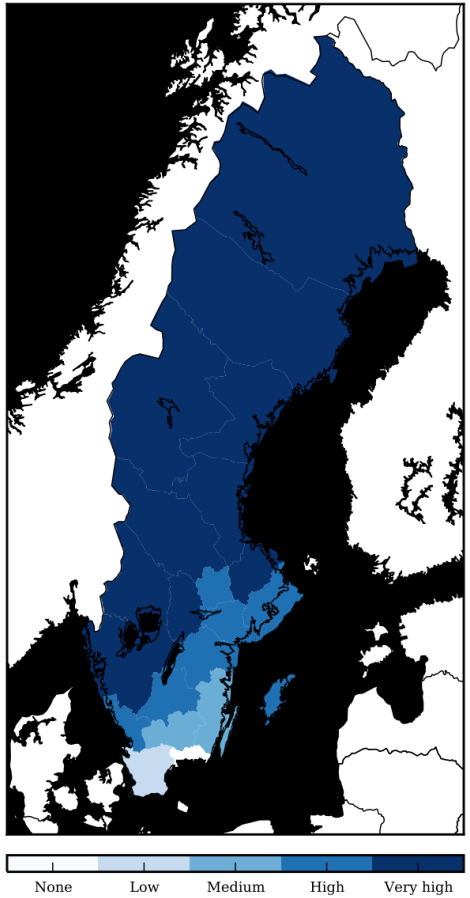


“CRY”

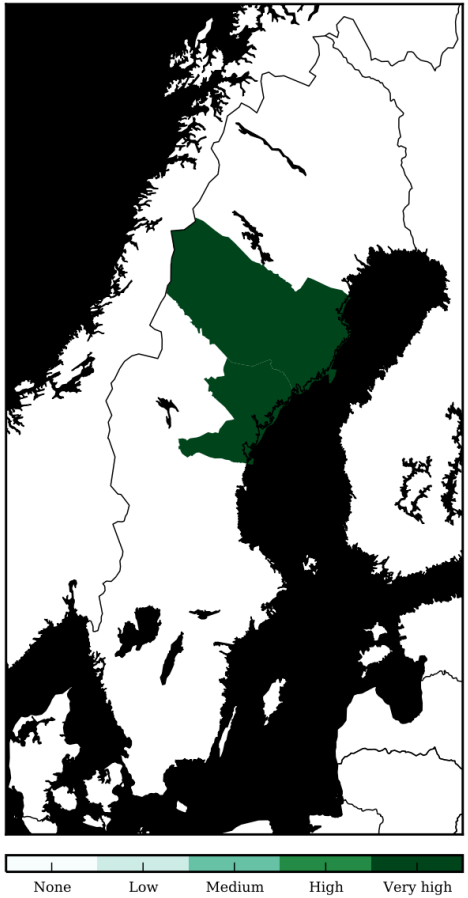
böla - hits: 2213



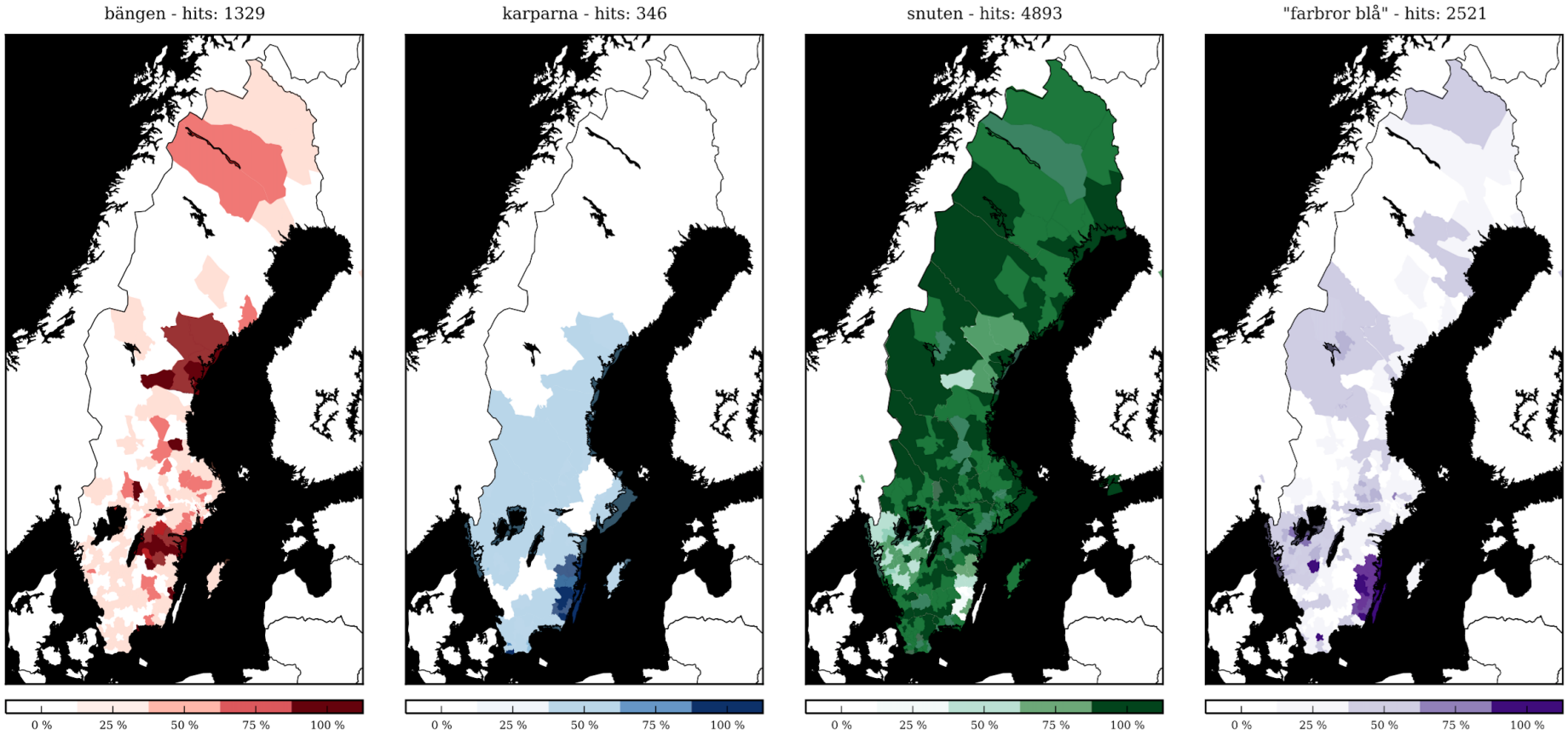
grina - hits: 5559



flänna/flännig - hits: 50



“THE POLICE”


bängen - hits: 1329
karparna - hits: 346
snuten - hits: 4893
"farbror blå" - hits: 2521
0 %
25 %
50 %
75 %
100 %

REFERENCES

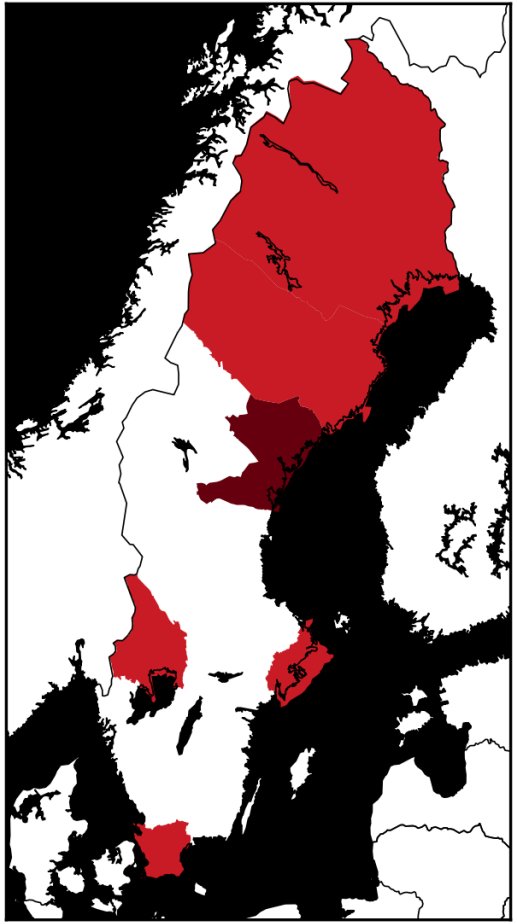
- Priedhorsky et al. 2014
  - http://arxiv.org/pdf/1305.3932.pdf

“DOLLAR-BILL”

query on blog dataset **without** inferred lat lon

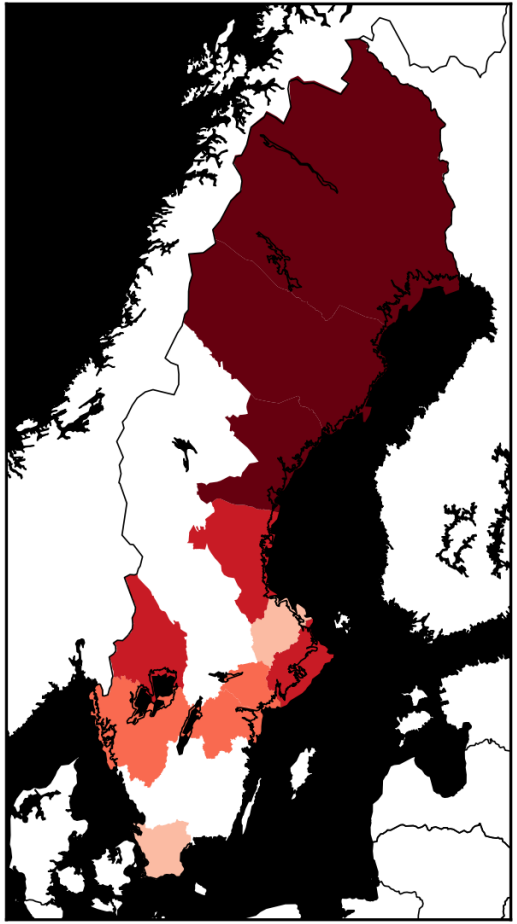
104 hits

hunka - hits: 104



Below avg. Expected Above avg.

hunka - hits: 340



Below avg. Expected Above avg.

query on blog dataset **with** inferred lat lon

340 hits

goal:

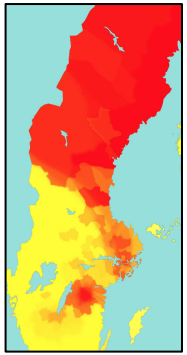


image: Mikael Parkvall