Visualizing COVID19 data from different countries

Martin Beneš

Supervisor: Krzysztof Bartoszek

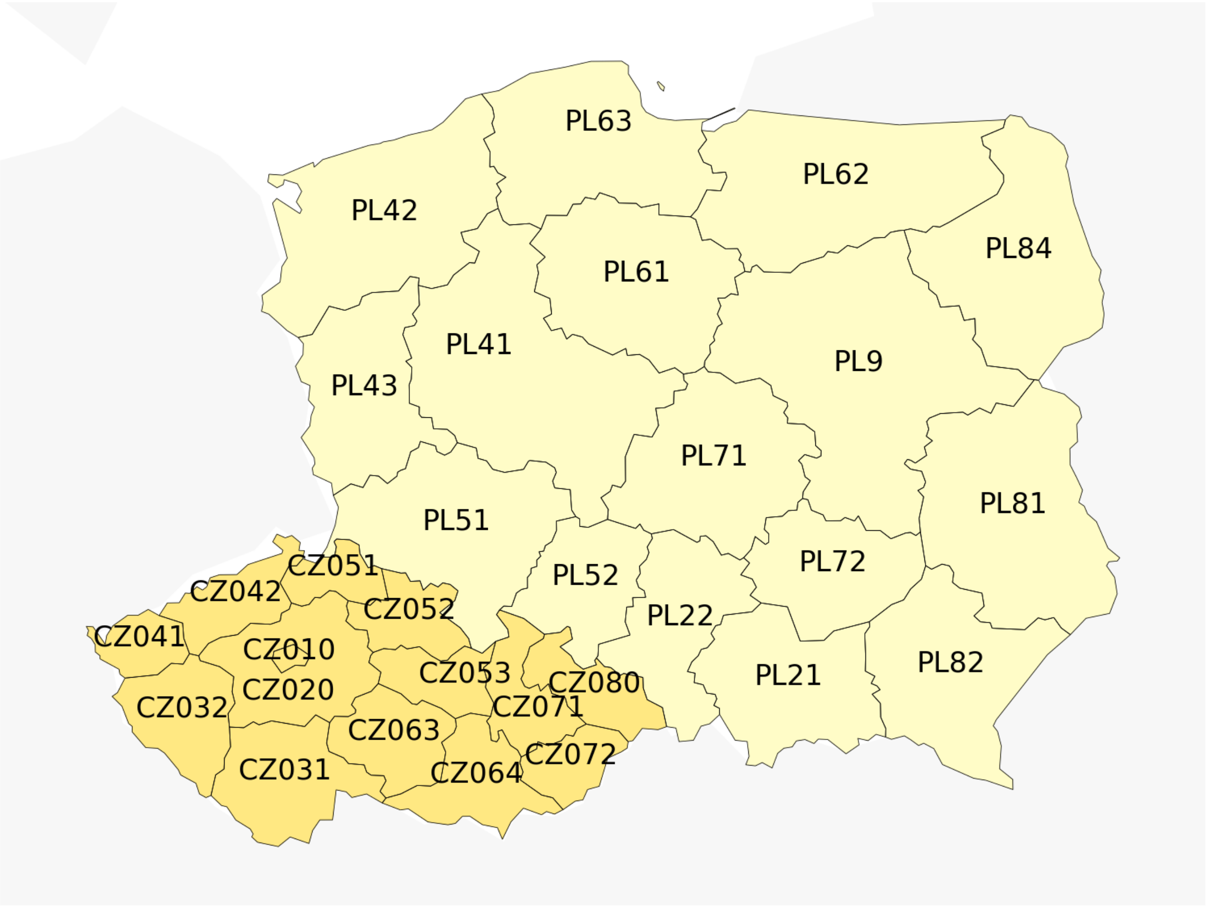
732A76 Research Project

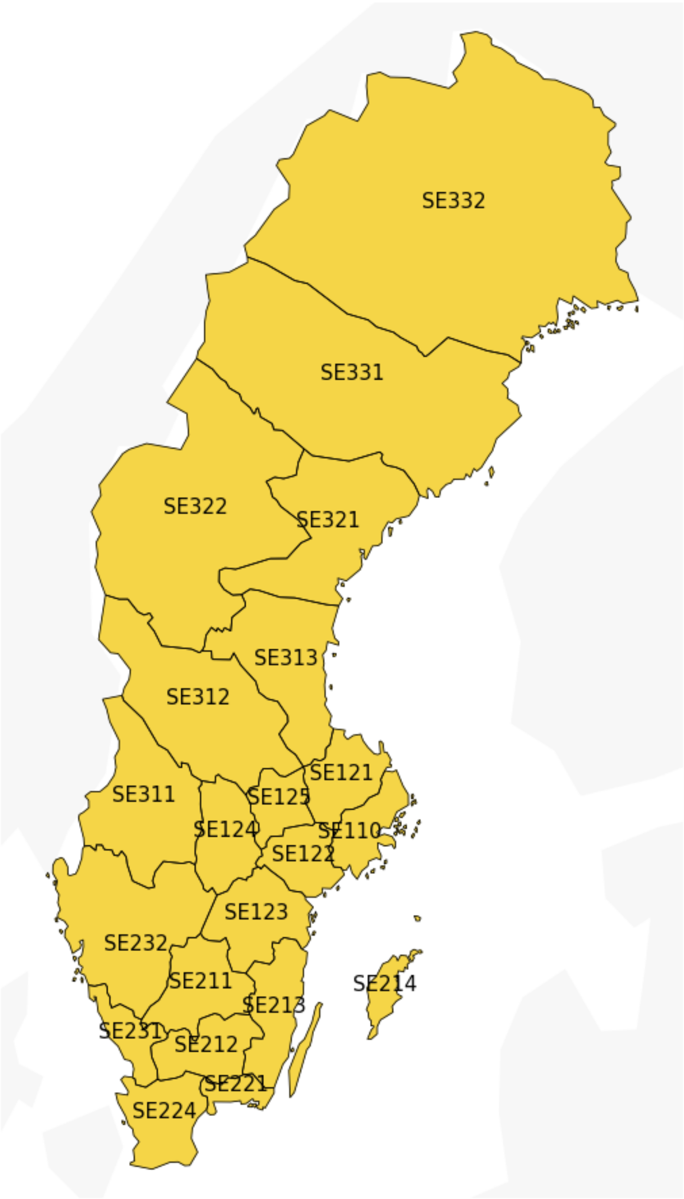
2020-01-12

Goals

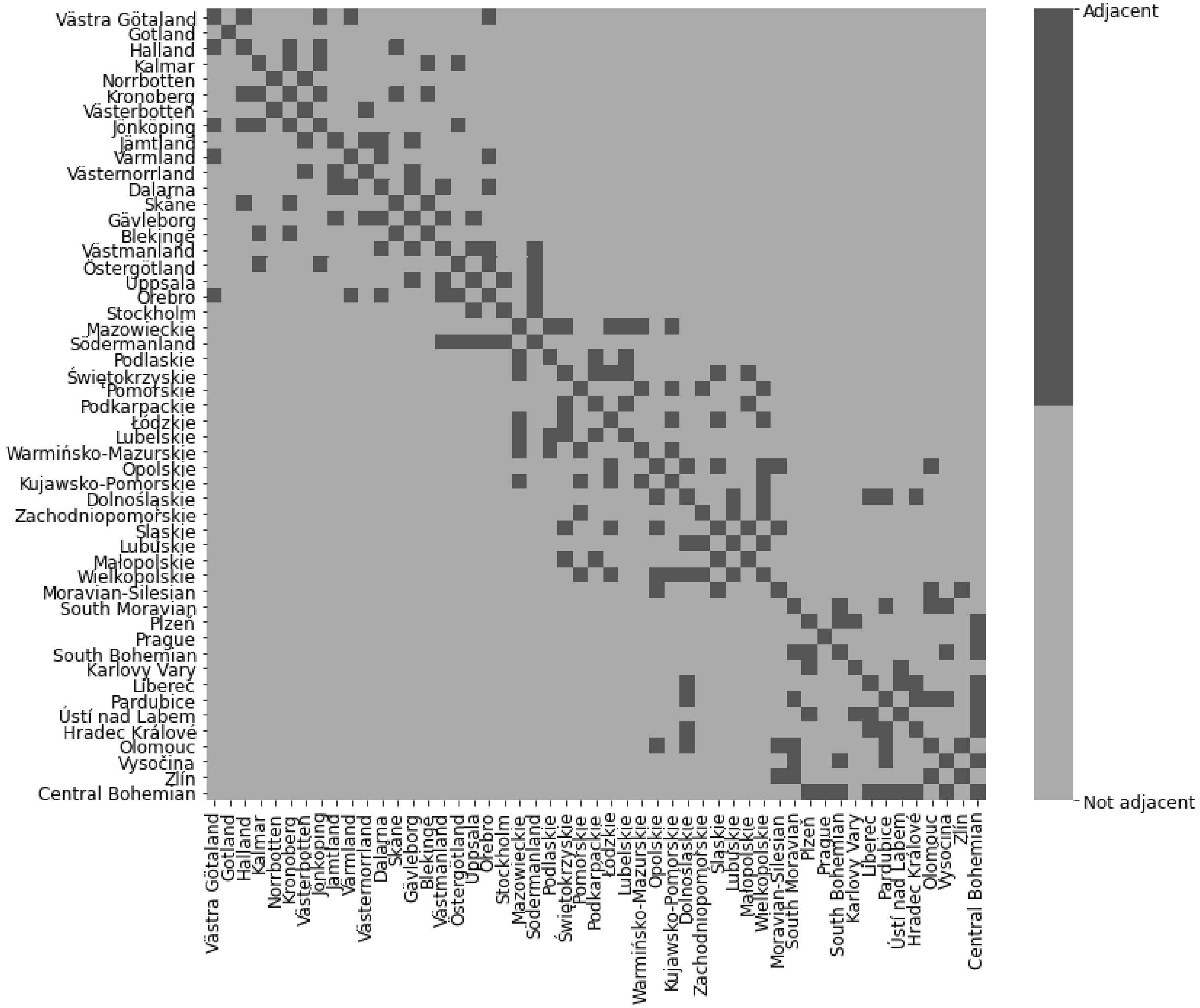
- Download mortality, population and COVID19 death data.
- Design similarity measure for administrative divisions.
  - What features are used?
  - Are regions of Czechia, Poland and Sweden comparable?
  - Clusters? Outliers?
- Design similarity measure for the COVID19 deaths data.
  - Use the metric to make regional comparison.
  - Clusters? Outliers?
  - Explain the observations

Administrative divisions





Administrative divisions



Data: Czechia

```python
import covid19czechia as CZ
x = CZ.covid_deaths()
```

Listing 1: covid19czechia usage example

- https://onemocneni-aktualne.mzcr.cz/
- Death cases with date, sex, age, region (LAU-1)
- CSV format
- Python package covid19czechia

Data: Sweden

```python
import covid19sweden as SE
x = SE.covid_deaths()
```

Listing 2: covid19sweden usage example

- https://scb.se/om-scb/nyheter-och-pressmeddelanden/overdodligheten-fortsatter-att-sjunka-efter-toppen-i-april/
- Weekly counts by region (NUTS-3)
- XLSX format
- Python package covid19sweden

$$w_i \sim \text{Multinomial}(n = w, \pi_i = \frac{1}{7}), i = 1, \ldots, 7 \quad (9)$$

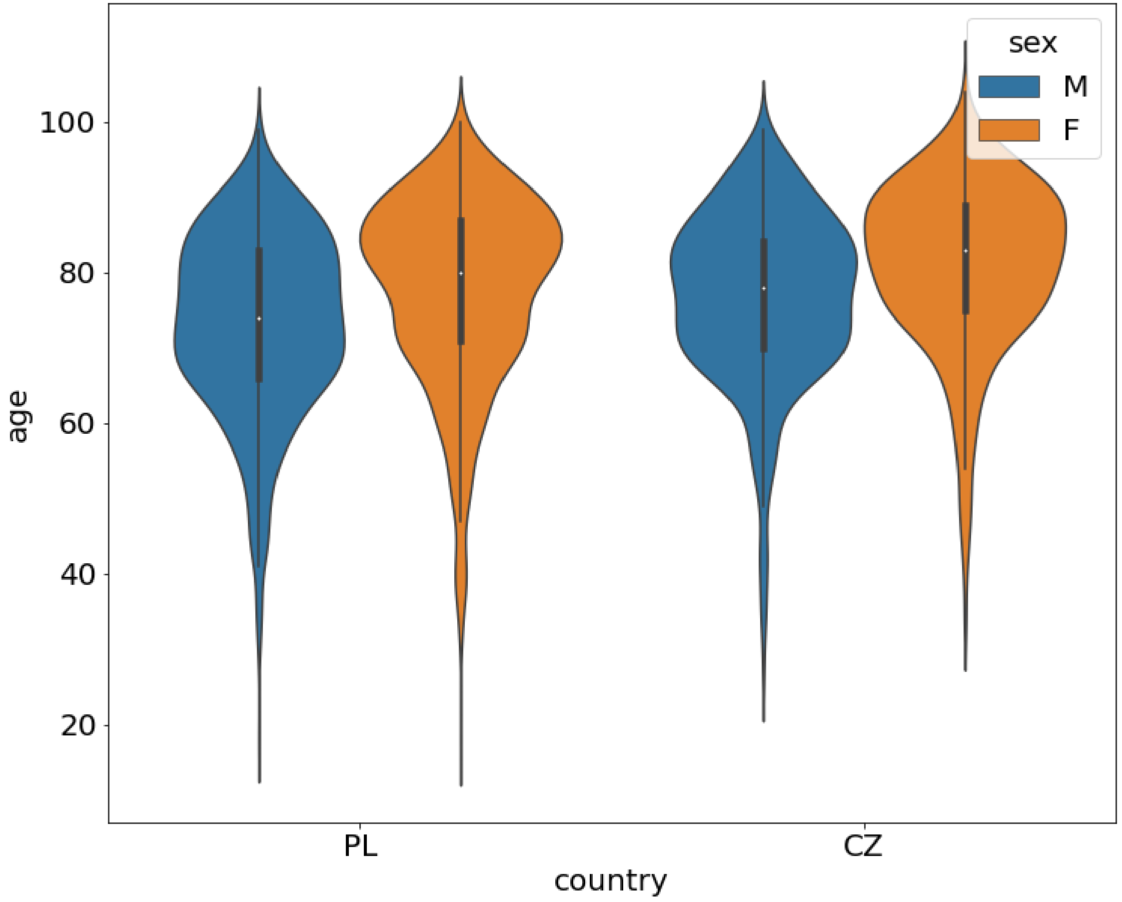
Data: Poland

```python
import covid19poland as PL
x = PL.covid_deaths()
```

Listing 3: covid19poland usage example

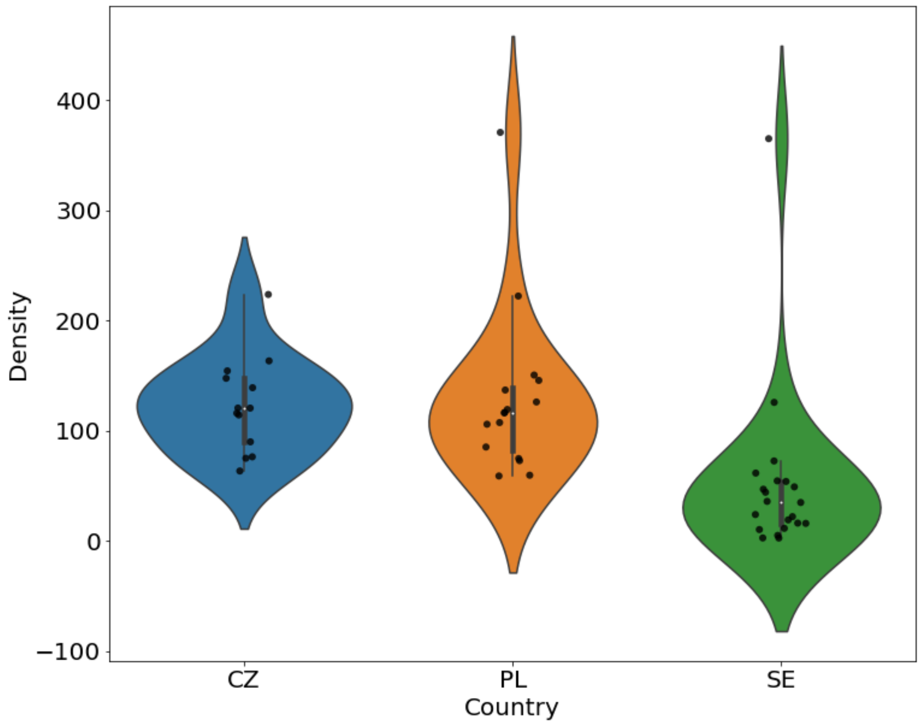
- https://twitter.com/MZ_GOV_PL
  - RegEx parsing the tweets
  - Data between 2020-03-12 and 2020-10-09 = by region, gender, age
  - Data between 2020-10-10 and 2020-11-23 = overall country counts
- https://www.gov.pl/web/koronawirus/pliki-archiwalne-wojewodztwa
  - Data after 2020-11-24 = overall regional counts
  - CSV
- Python package covid19poland

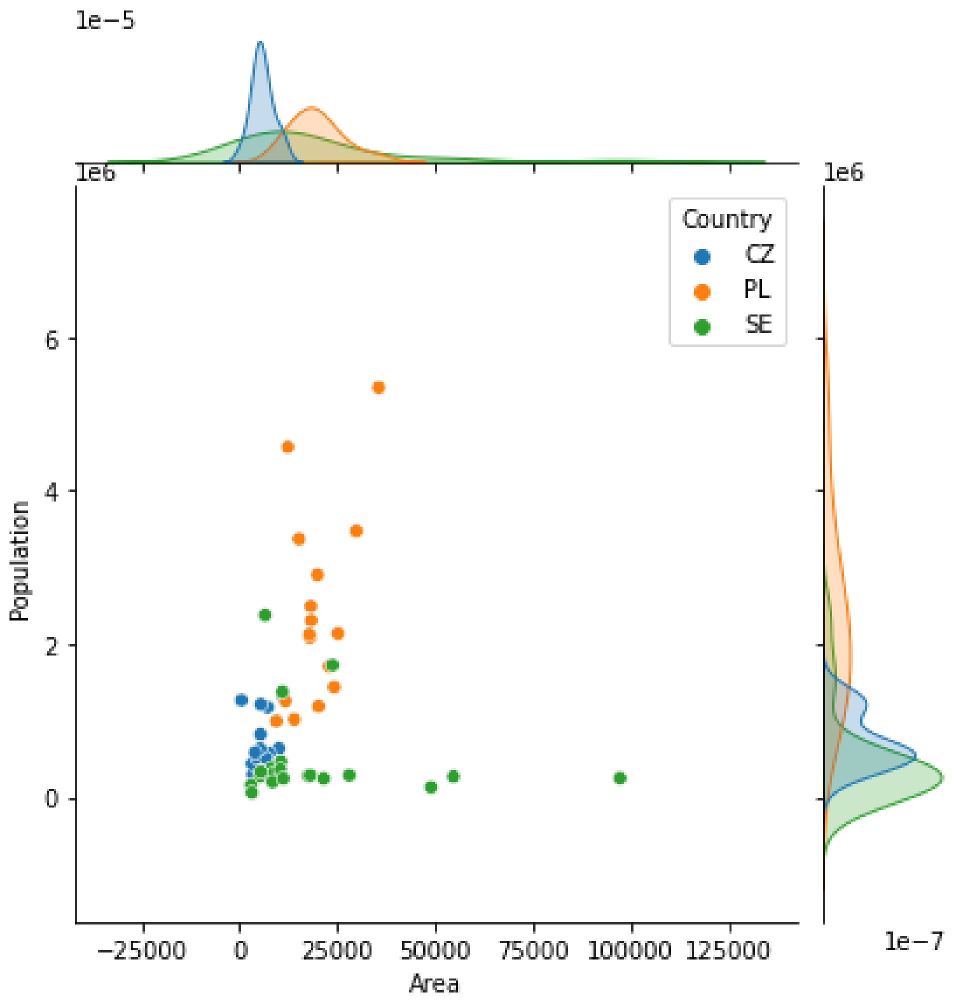
Age distribution



Region statistics

- Population, area
- Population density





Region statistics

Figure 7. Summary of the regional divisions.

Statistics		**Countries**		
		Czechia	*Sweden*	*Poland*
	N	14	21	16
Population	μ	753845	491790	2402327
	σ	343842	587474	1266901
Area	μ	5634	19394	19542
	σ	2759	22605	6836
Density	μ	297	51	129
	σ	651	78	76

Figure 12. IQR outliers.

Country	**Outliers**		
	Population	*Area*	*Density*
Sweden		SE322, SE331, SE332	SE110
Poland	PL9, PL22		PL22
Czechia			CZ010

Regional statistics

$$
\begin{aligned}
H_0 &: \text{Data} \sim t(\cdot) \\
H_A &: \text{Data} \nsim t(\cdot)
\end{aligned}
\quad (7)
$$

Figure 8. P-values for Kolmogorov-Smirnov test (eq. 7).

Country	Attributes: *Population*	*Area*	*Density*
Czechia	0.141	0.001	0.097
Sweden	0.116	0.009	0.083
Poland	0.001	0.001	0.129

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 \\
H_A &: \mu_1 \neq \mu_2
\end{aligned}
\quad (6)
$$

Figure 9. P-values for t-test test (eq. 6).

Country		Attributes: *Population*	*Area*	*Density*
Sweden	Poland	$1.82 \cdot 10^{-5}$	0.98	$4.2 \cdot 10^{-3}$
Sweden	Czechia	0.143	0.031	0.094
Poland	Czechia	$1.01 \cdot 10^{-4}$	$3 \cdot 10^{-7}$	0.314

Regional comparison

- Hypothesis: Close regions might form outbreak clusters.
- What are close regions?
  - Close by distance

$$kd(x_1, x_2) = \exp\left(-\frac{d_{GC}(x_1, x_2)^2}{2h^2}\right) \tag{1}$$

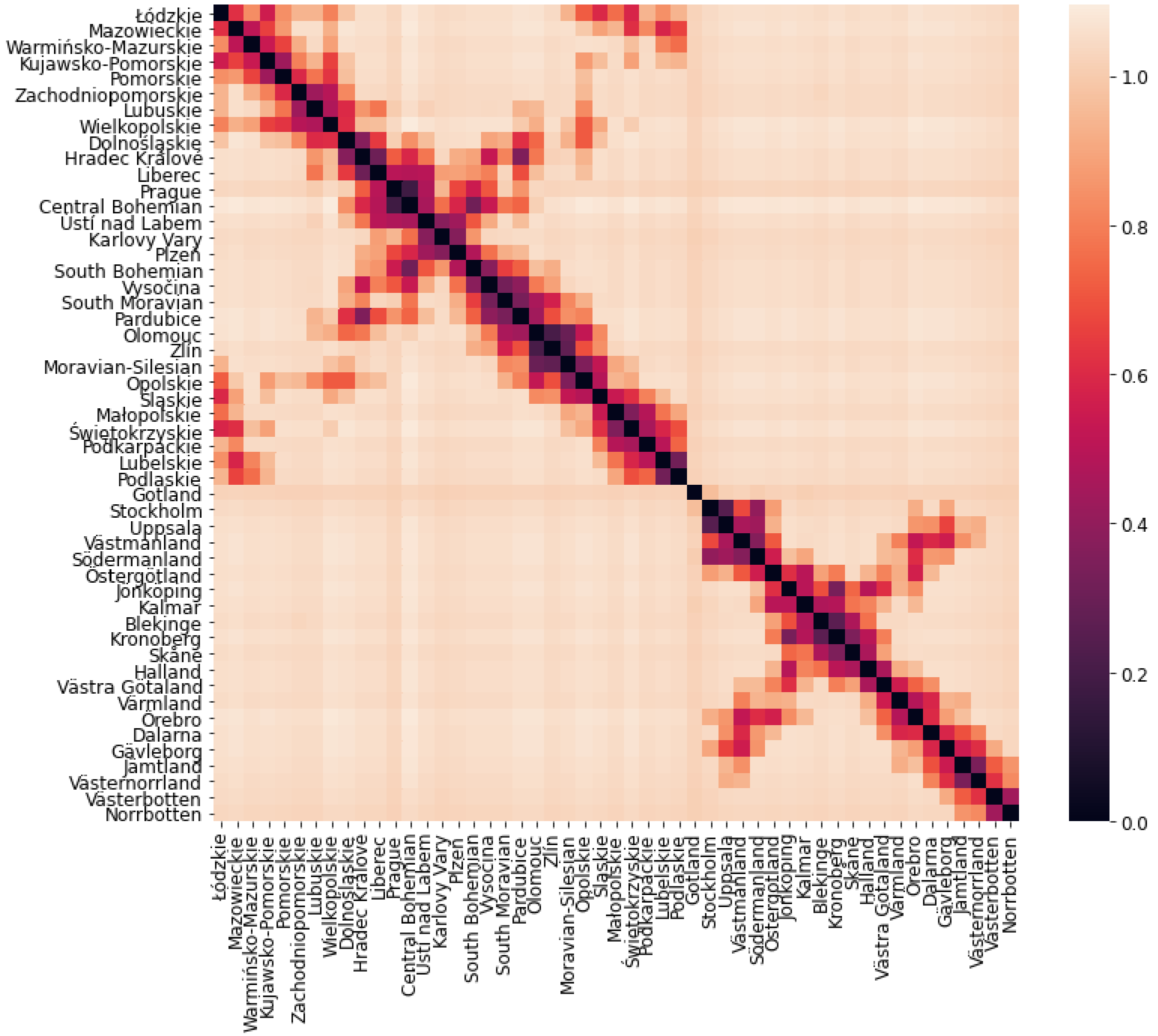
  - Close by number of common neighbors

$$d(x, y) = 1 - \frac{|neighbors(x) \cap neighbors(y)|}{|neighbors(x) \cup neighbors(y)|} \tag{2}$$

  - Both

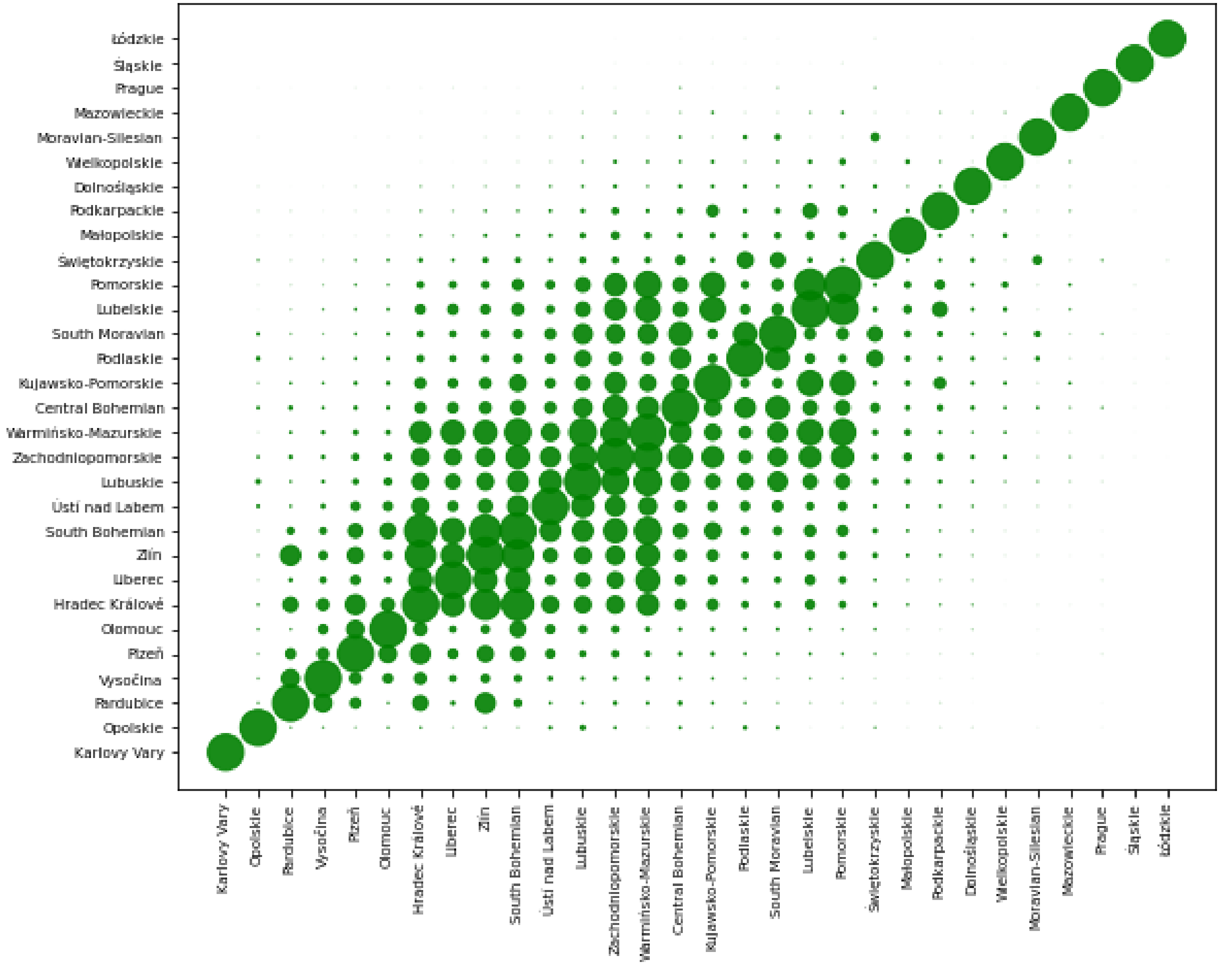
$$d(x, y) = \sqrt{kd(\overline{x}, \overline{y}) \cdot d_{adj}(x, y)} \tag{3}$$

Regional comparison: location



- Östergötland, Jonköping, Kalmar, Blekinge, Kronoberg, Skåne, Halland (*Southern Sweden*)
- Örebro, Södermanland, Stockholm, Uppsala, Västmanland (*Stockholm*)
- Jämtland, Västernorrland, Västerbotten, Norrbotten (*Northern Sweden*)
- Prague, Central Bohemian, Liberec, Hradec Králové, Dolnośląskie (*Bohemia*)
- Dolnośląskie, Lubuskie, Wielkopolskie, Zachodniopomorskie (*Western Poland*)
- Łódzkie, Mazowieckie, Podlaskie (*Northern Poland*)
- Podkarpackie, Świetokrzyskie, Ślaskie (*Eastern Poland*)
- Ślaskie, Opolskie, Moravian-Silesian, Zlín, Olomouc (*Silesia*)
- Pardubice, South Moravian, Vysočina (*Moravia + Bohemia*)
- Plzeň, Ústí nad Labem, Karlovy Vary (*Bohemia*)

Czekanowski diagram



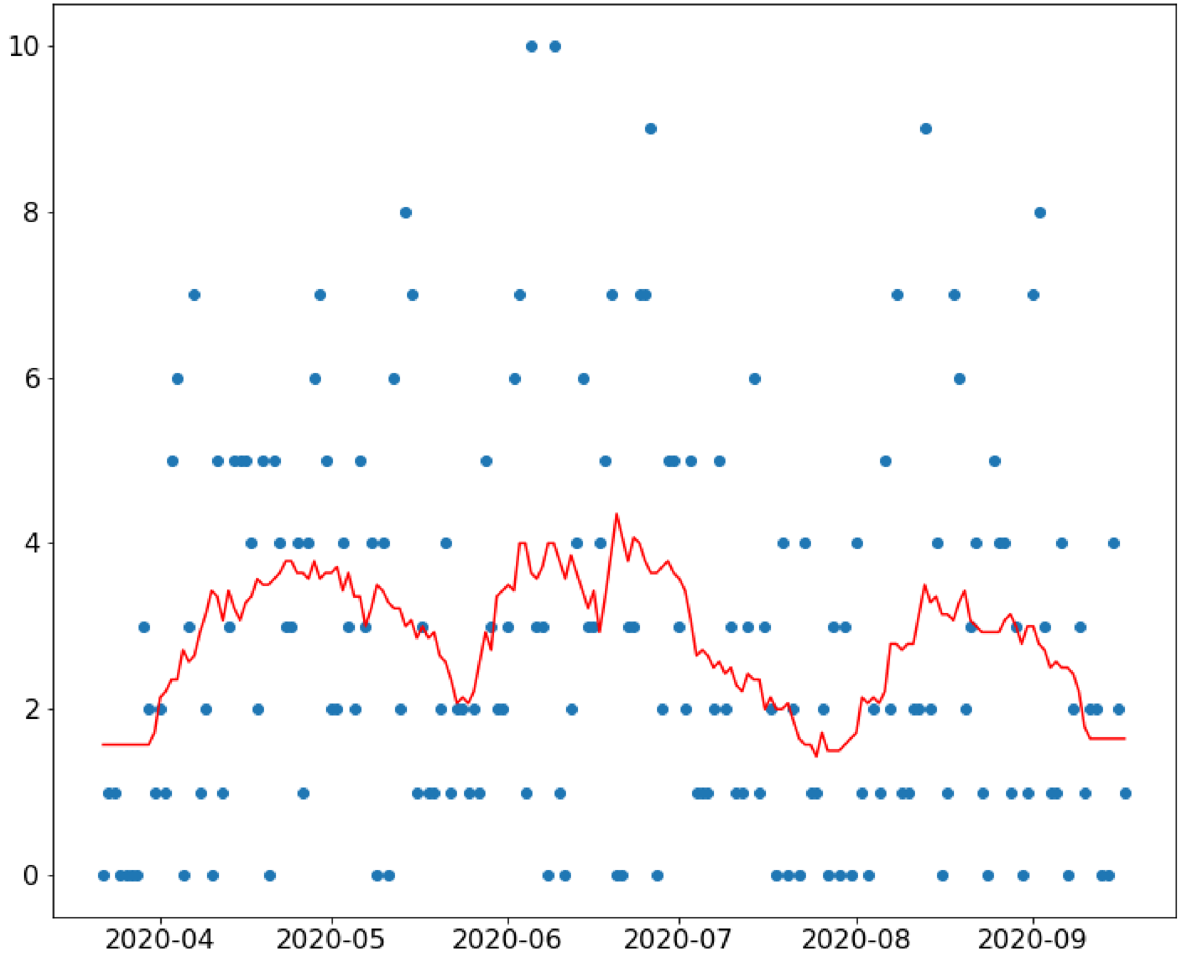
Czekanowski diagram

```
# fitness of each chromosome
fitness = _ga.population_score(pop,obj)

# crossover
parents,pscore  = _ga.select_parents(pop,fitness,
                                      Nparents)
children,cscore = _ga.crossover(parents,obj)
# create mutants
mutants,mscore = _ga.mutate(children,obj,mutprob)

# war
pop = _ga.war(popsize,(parents,pscore),
              (children,cscore),(mutants,mscore))
```

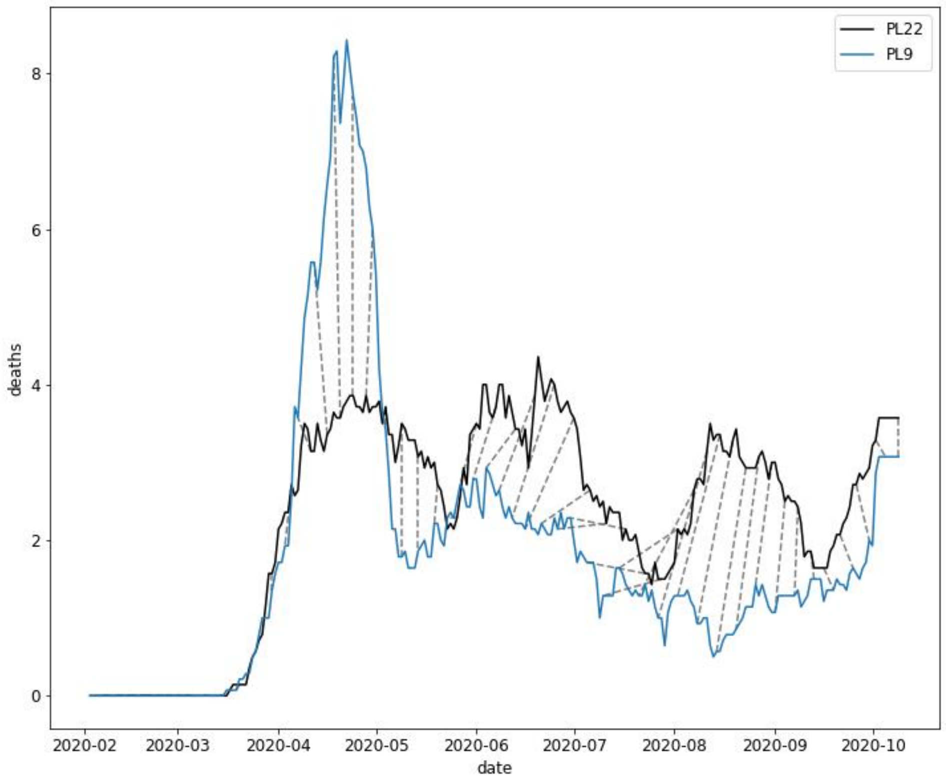
Listing 4: Genetic algorithm

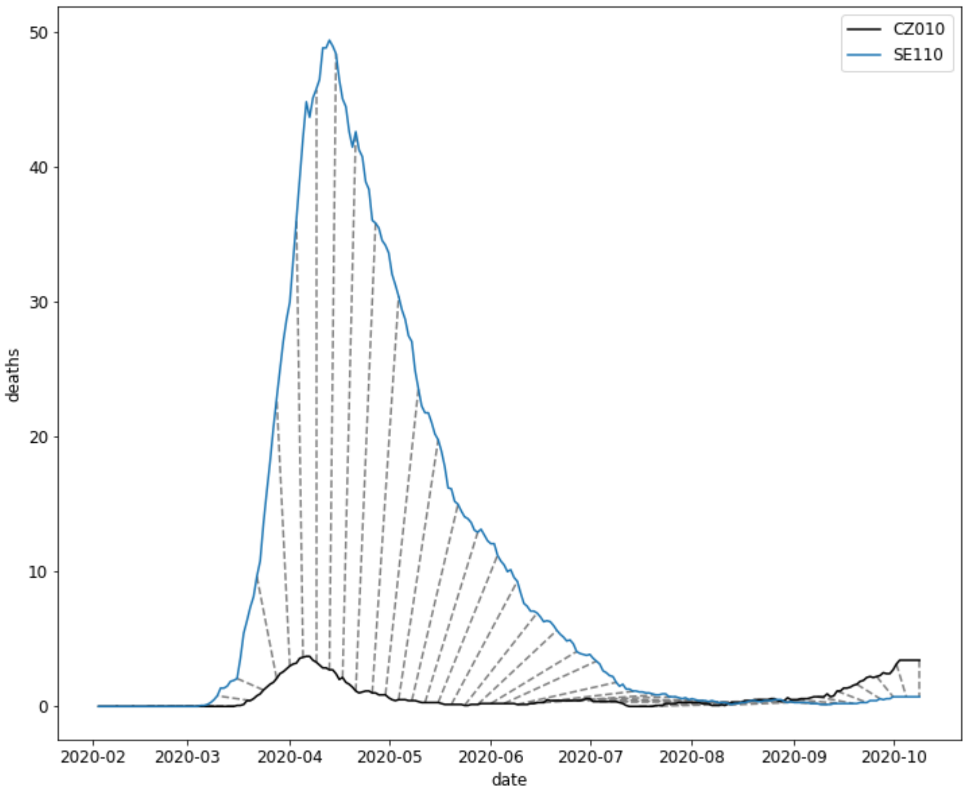
$$U_m = \frac{2}{n^2} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \frac{(i-j)^2}{W_{ij}+1} \tag{5}$$

Smoothing



Dynamic Time Warping





Epidemic comparison

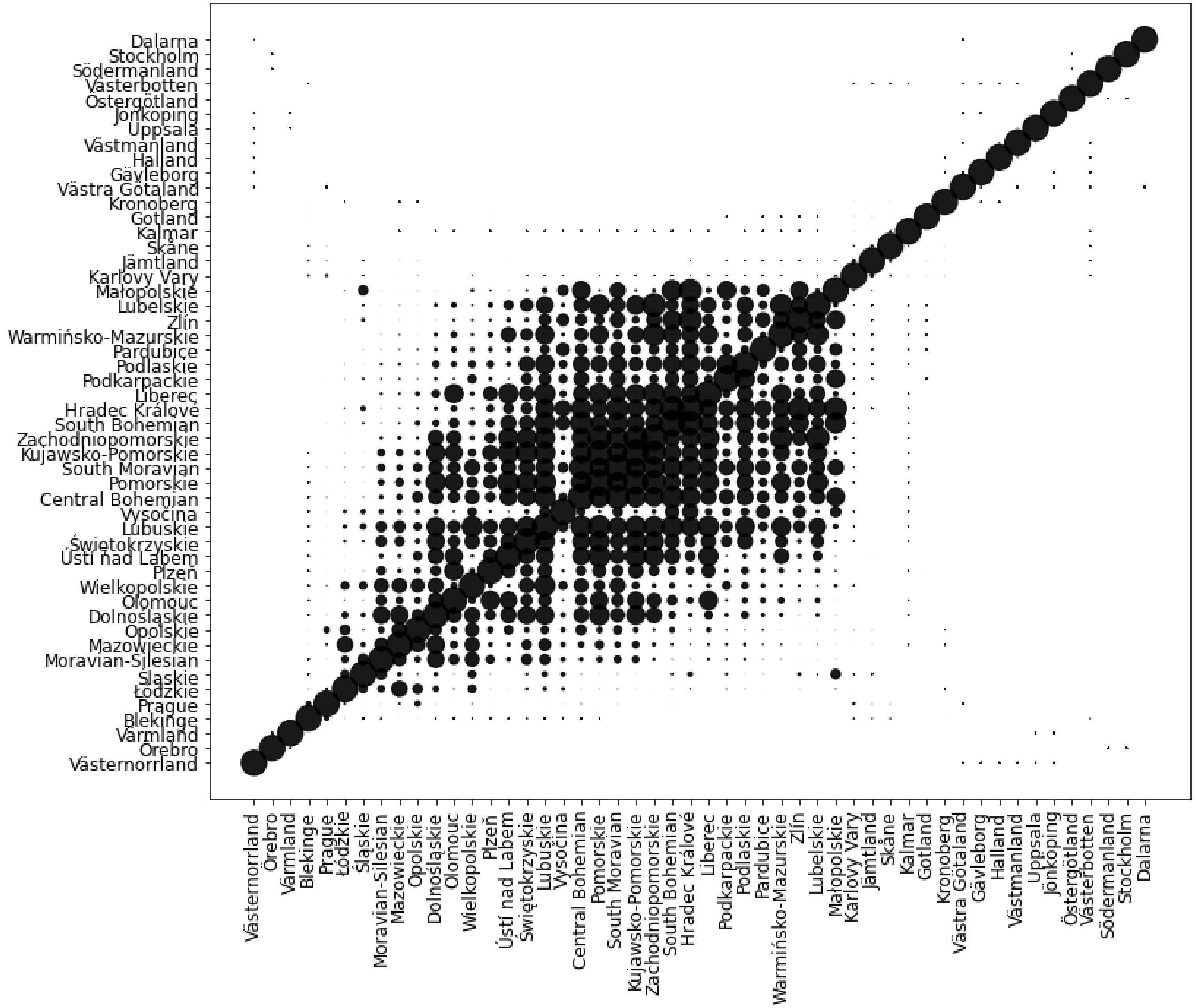
- Method
  - Dynamic Time Warping (DTW)
  - RBF kernel
  - Czekanowski diagram

```python
# distance matrix (metric dtw)
D = _covid.dtw_distance(data = data)
# rbf kernel
D = _czekanowski.distance_rbf(D)
# column permutation
P = _czekanowski.plot(D, cols = columns)

# Czekanowski diagram
import matplotlib.pyplot as plt
plt.scatter(P.x, P.y, s=P.Distance); plt.show()
```

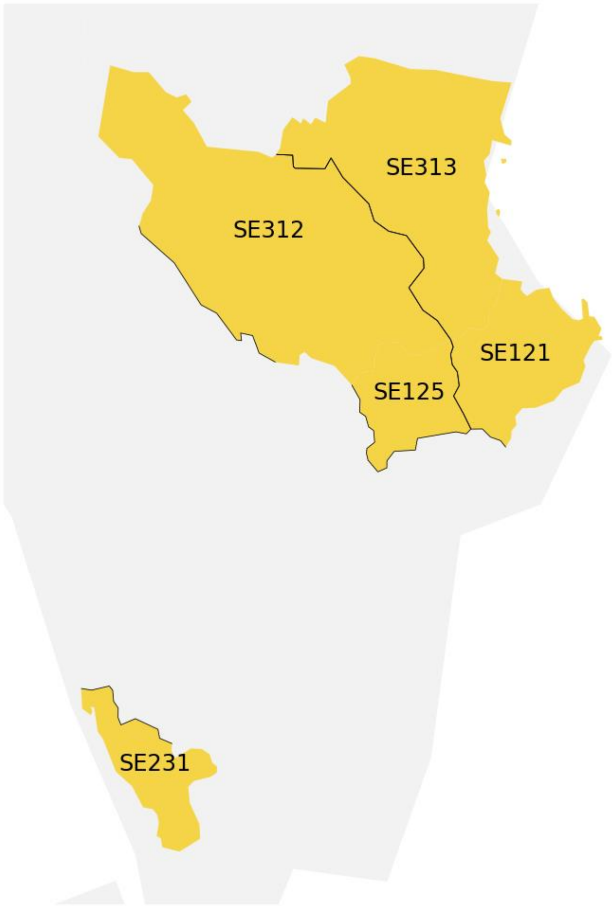
Listing 5: Czekanowski DTW method

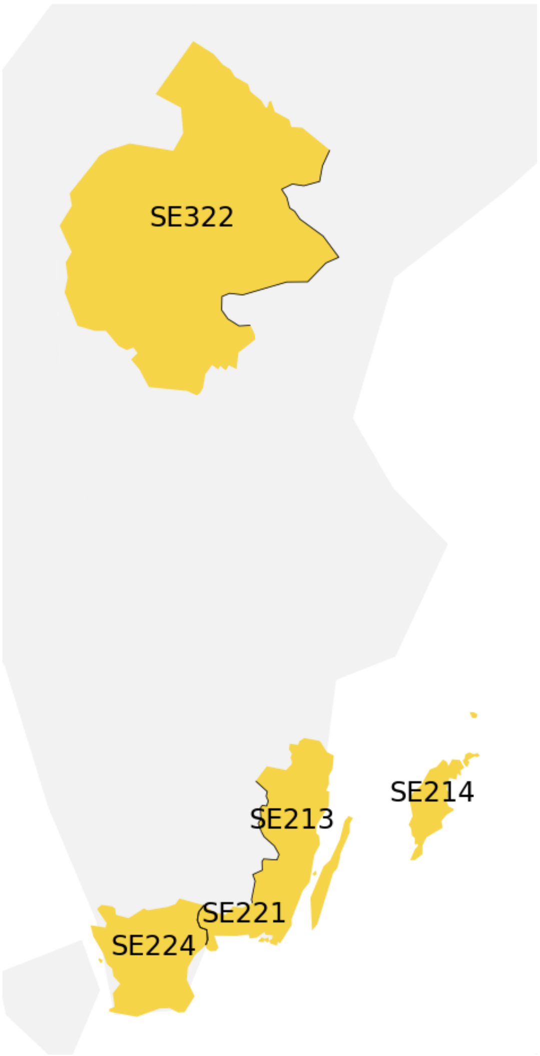
Results

– Central Bohemian, South Bohemian, Hradec Králové, Liberec, South Moravian, Zachodniopomorskie, Pomorskie, Kujawsko-Pomorskie
– Warmińsko-Mazurskie, Zlín, Lubelskie
– Ústí nad Labem, Świetokrzyskie, Lubuskie, Plzeň
– Podlaskie, Podkarpackie
– Mazowieckie, Opolskie, Dolnoślaskie
– Uppsala, Dalarna, Gävleborg, Västmanland, Halland
– Skåne, Blekinge, Kalmar, Karlovy Vary, Jämtland, Gotland, Prague

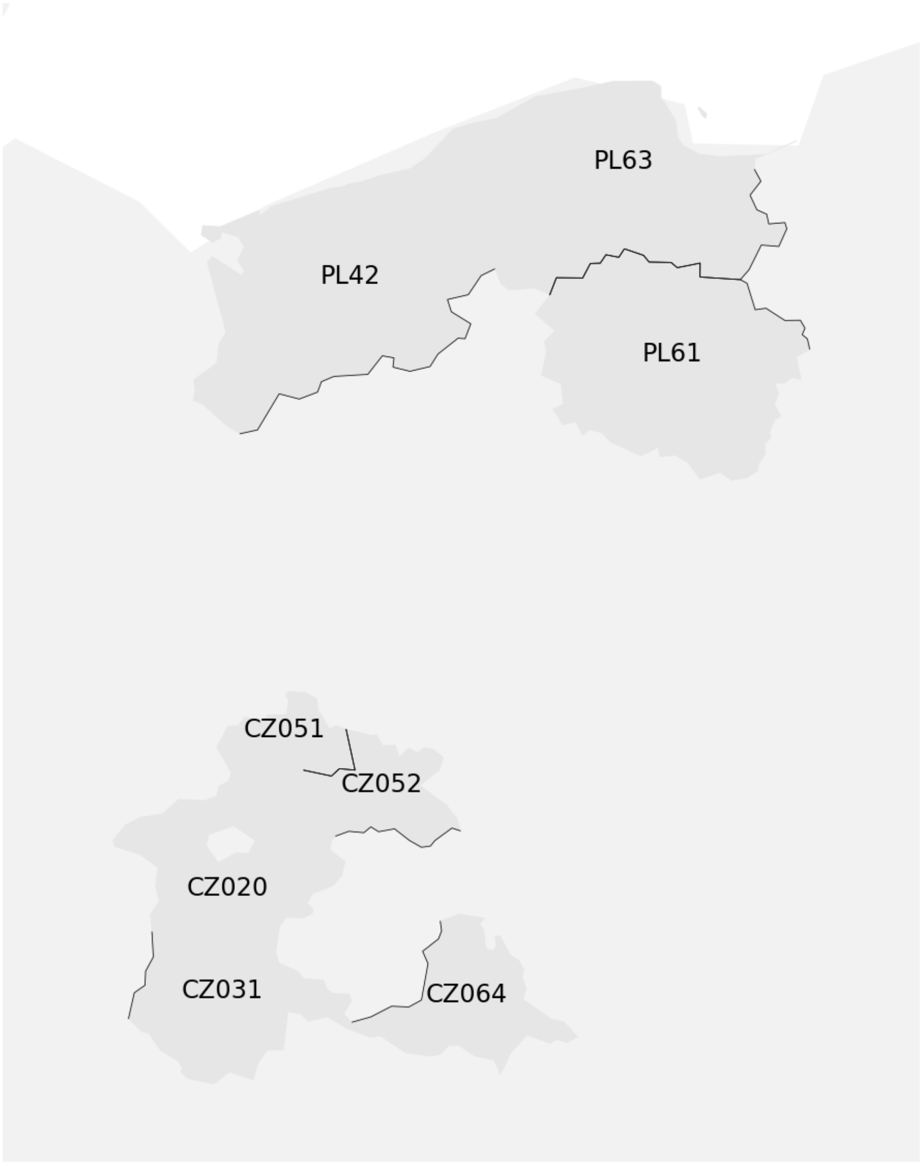


Results


SE313
SE312
SE121
SE125
SE231


SE322
SE214
SE213
SE221
SE224

Results



Conclusions

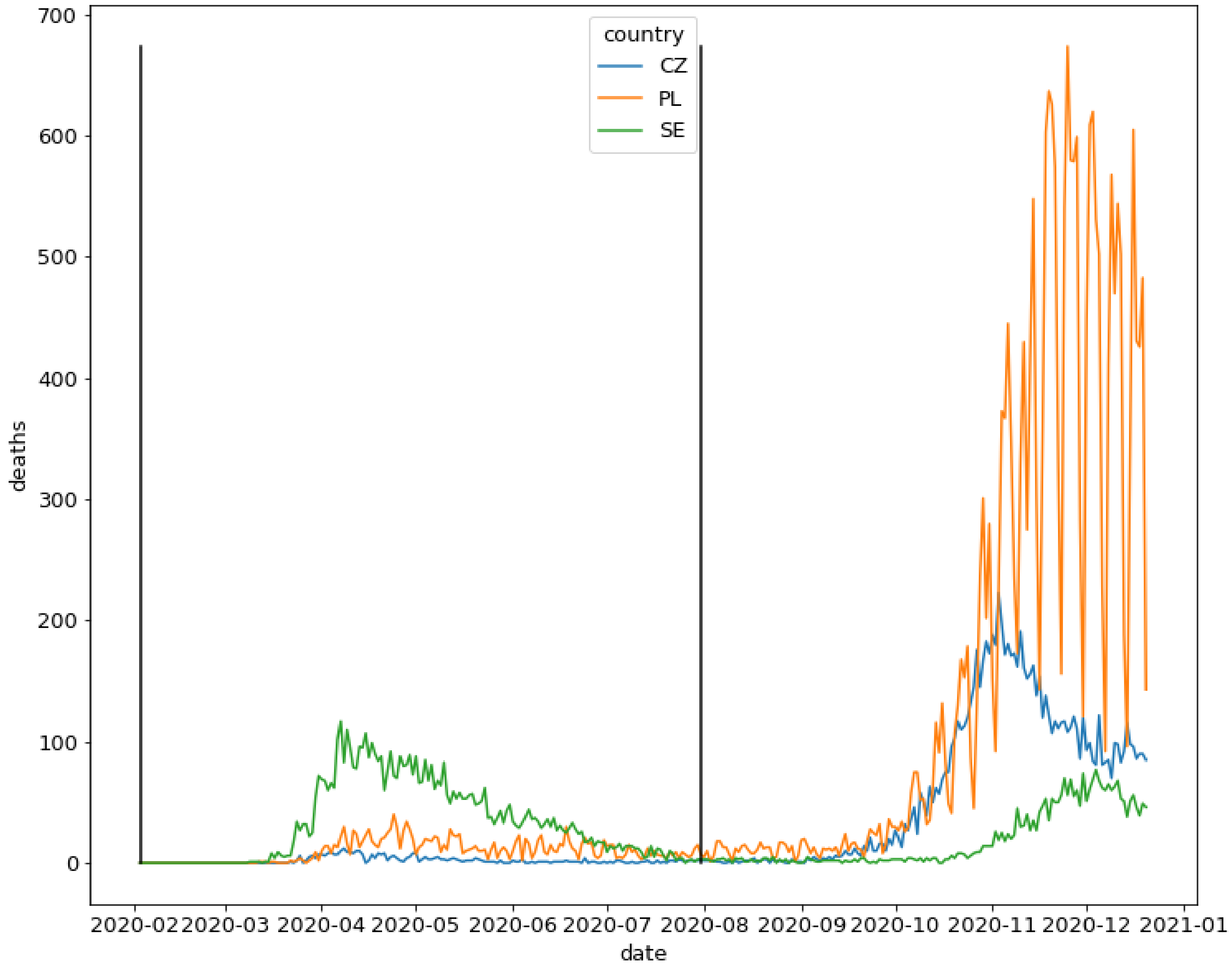
- Method yields reasonable results.
- Parameter values (kernel width) are crucial.
- Close regions sometimes do form infection clusters.

Weekday-independent deaths

$$
\begin{aligned}
H_0 &: \mu_i = \frac{1}{7} \\
H_A &: \mu_i \neq \frac{1}{7}
\end{aligned} \tag{10}
$$

Figure 21. P-values for equal ratio t-test (eq. 10).

Day	Country		
	Czechia	*Poland*	*Sweden*
Monday	0.581	0.001	0.429
Tuesday	0.496	0.06	0.088
Wednesday	0.784	0.112	0.731
Thursday	0.375	0.181	0.924
Friday	0.298	0.764	0.507
Saturday	0.112	0.737	0.394
Sunday	0.294	0.044	0.947



country
CZ
PL
SE
deaths
700
600
500
400
300
200
100
0
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
date

Thank you for attention!

Děkuji za pozornost!

Dziękuję za uwagę!