Improved Corona incidence maps via kernel density estimates

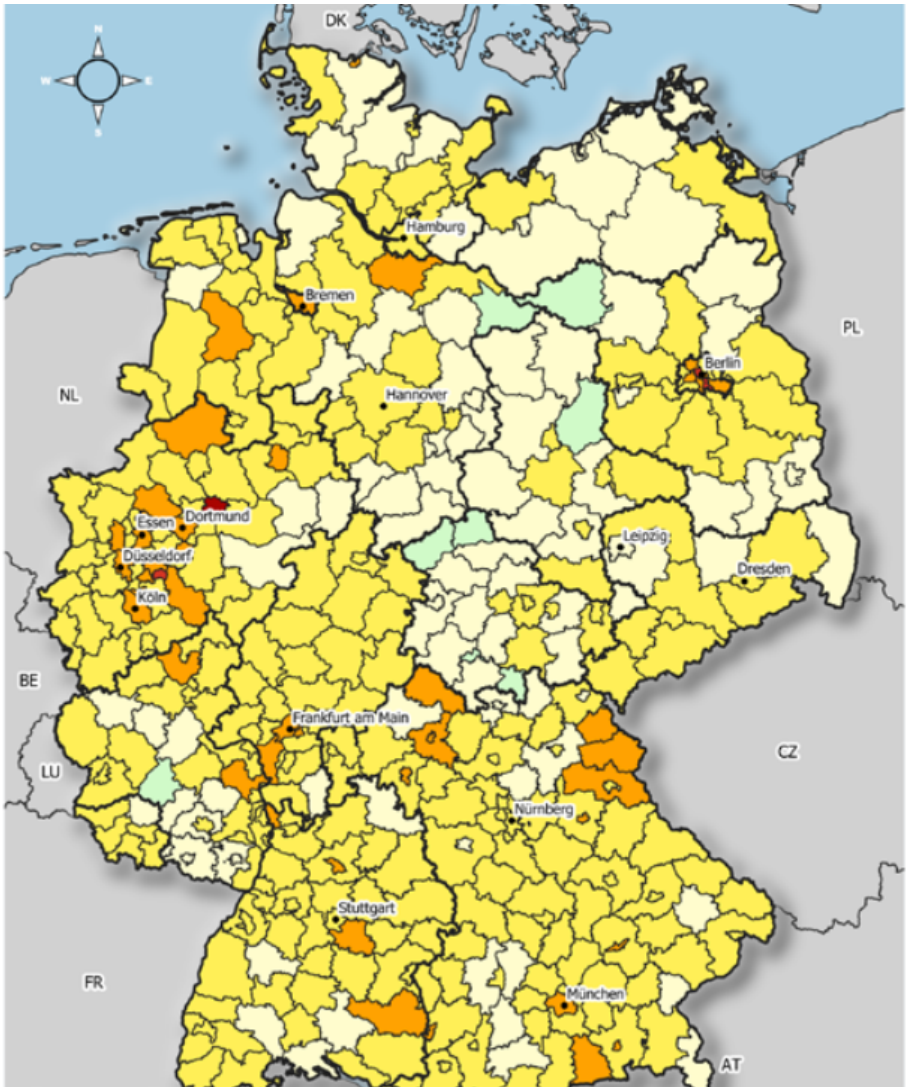
Ulrich Rendtel (FU Berlin)
Marcus Groß (INWT Statistics GmbH, Berlin)
Lukas Fuchs (Joint Berlin Master Program Statistics)
Jingying Shang (Joint Berlin Master Program Statistics)
Andreas Neudecker (INWT Statistics GmbH, Berlin)

16. November 2021
Kolloquium "Stat. Methoden in der empirischen Forschung"
Online

Freie Universität Berlin

A typical Corona incidence map

Incidence= (no. of infections in 7 days/ no. of inhabitants) x 100000



Übermittelte Fälle der letzten **7** Tage

COVID19-AKTIVITÄT
Stand: 03.10.2020

Fälle pro 100.000 Einwohner

- keine Fälle übermittelt [8]
- >0,0 - 5,0 [94]
- >5,0 - 25,0 [261]
- >25,0 - 50,0 [45]
- >50,0 - 100,0 [3]
- >100,0 - 500,0 [1]

	Kreis	Fälle	Inzidenz
1	SK Hamm	180	100,5
2	SK Berlin Neukölln	187	56,5
3	SK Remscheid	62	55,9
4	SK Berlin Mitte	213	55,5
5	SK Berlin Tempelhof-Schöneberg	166	47,2
6	SK Berlin Friedrichshain-Kreuzberg	136	47,0
7	SK Frankfurt am Main	329	43,7
8	SK Duisburg	212	42,5
9	SK Berlin Charlottenburg-Wilmersdorf	141	41,3
10	SK Bremen	231	40,6
11	LK Rhön-Grabfeld	31	38,9
12	SK Schweinfurt	21	38,9
13	LK Oberbergischer Kreis	103	37,8
14	SK Leverkusen	61	37,2
15	SK Gelsenkirchen	94	36,1

Aims of incidence maps

- Make infection numbers comparable across regions with different population sizes
- Show regions with low and high incidence figures (Range)
- Identify local clusters with high incidences (Spatial distribution)
- Temporal development of spatial distribution (Animated maps)

Fields of applications

- Disease maps
- Regional voting results
- Poverty maps, social atlases

⋮

-

Standard incidence maps via Choropleths

- Based on reference areas. Usually Counties, federal states, etc.
- Area value: (no. of new infections within a week/ no. inhabitants) x 100000 in area
- Temporal development of spatial distribution (Animated maps)
- **But:**
- For different area systems quite different maps are generated
- Unrealistic assumption of uniform local incidence within area
- Information reduction by discrete display of levels (6 colours)
- Discontinuities at borders avoid identification of clusters

A density framework for the display of maps

- Density is independent from reference areas
- Kernel density estimates are smooth and flexible
- **However:** Individual geo-coordinates are needed!
- Only local aggregates are known!
- Approach: Use statistical missing data techniques, like EM algorithm.

Definition of local incidences based on densities

- Density of infections $f_I(x_1, x_2)$ for geo-coordinate (x_1, x_2)
- Density of population $f_P(x_1, x_2)$ for geo-coordinate (x_1, x_2)
- Total number of infections N_I
- Total number of population N_P
- Definition of local incidence at coordinate (x_1, x_2):

$$f_{I|P}(x_1, x_2) = \frac{N_I}{N_P} f_I(x_1, x_2) / f_P(x_1, x_2) \times 100000$$

Estimation of local incidence in case of known geo-coordinates

- Replace $f_I(x_1, x_2)$ and $f_P(x_1, x_2)$ by kernel density estimates:

$$\hat{f}_{NW}(x_1, x_2) = \frac{N_I \hat{f}_I(x_1, x_2)}{N_P \hat{f}_P(x_1, x_2)} \times 100000 \qquad (1)$$

  This is the nonparmetric Nadaraya/Watson estimator, see for example, Härdle (1991).
- Use a **joint smoothing factor** from the distribution of infections!

The estimation of incidences with aggregates for areas

- Compute $\hat{f}_{NW}(x_1, x_2)$ by an iterative procedure: Simulated Expectation Maximation (SEM) algorithm
- S-Step: Sample from the actual distributions $\hat{f}_P^{(n)}$ for the population and $\hat{f}_I^{(n)}$ for the infected persons.
- M-Step: Estimate new densities $\hat{f}_P^{(n+1)}$ from the population sample and $\hat{f}_I^{(n+1)}$ from the infection sample.
- Calculate incidence for each iteration $\hat{f}_{NW}^{(n)}(x_1, x_2)$
- Repeat the iterations B-times for a burn-in phase and then R-times for a replication phase.
- Compute the final incidence $\hat{f}_{NW}(x_1, x_2)$ by:

$$\hat{f}_{NW}(x_1, x_2) = \frac{1}{R} \sum_{r=1}^{R} \hat{f}_{NW}^{(B+r)}(x_1, x_2) \qquad (2)$$

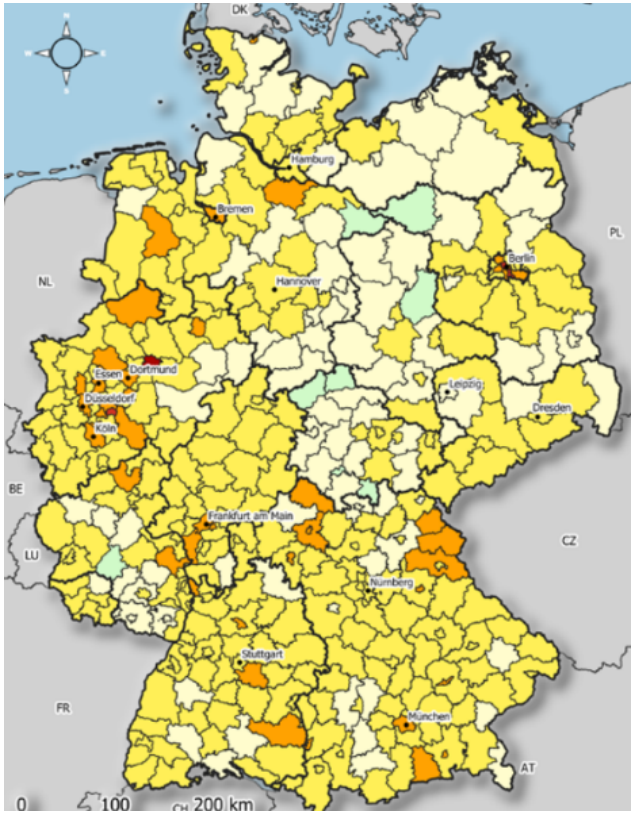
Freie Universität Berlin

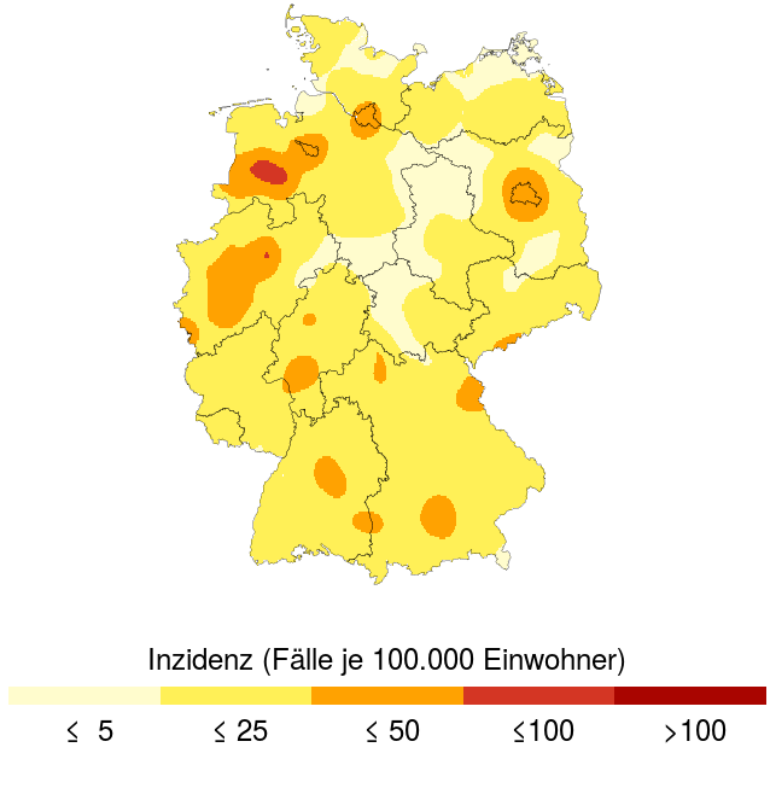
Realisation with the R-Package *Kernelheaping*

- Drawing a population sample in each area of size $N_{P,a}$ on a fine grid $\mathcal{G}$ in each area. Sampling is proportional to size with $\hat{f}_P^{(n)}$ as size variable. This generates the population sample $s_P^{(n)}$.
- Sampling of infected persons is with replacement from $s_P^{(n)}$ with sample size $N_{I,a}$ in area a. Sampling is proportional to size with $\hat{f}_I^{(n)}$ as size variable. This generates $s_I^{(n)}$.
- The smoothing factor h for the kernel estimation is calculated by a data driven procedure of Wand/Jones (1994) on the basis of the infection sample $s_I^{(n)}$.
- The estimated densities for the population and the infected persons are consistent with the area totals for the population and the number of infected persons.
- The R-Package *kernelheaping* for density estimation with "heaped" data is freely available.

Freie Universität Berlin

A comparison of maps of Corona-incidences at 3. October 2020. Left: Official Choropleth map of German counties by the RKI. Right: Map computed via Kernelheaping algorithm based on the county figures of the RKI





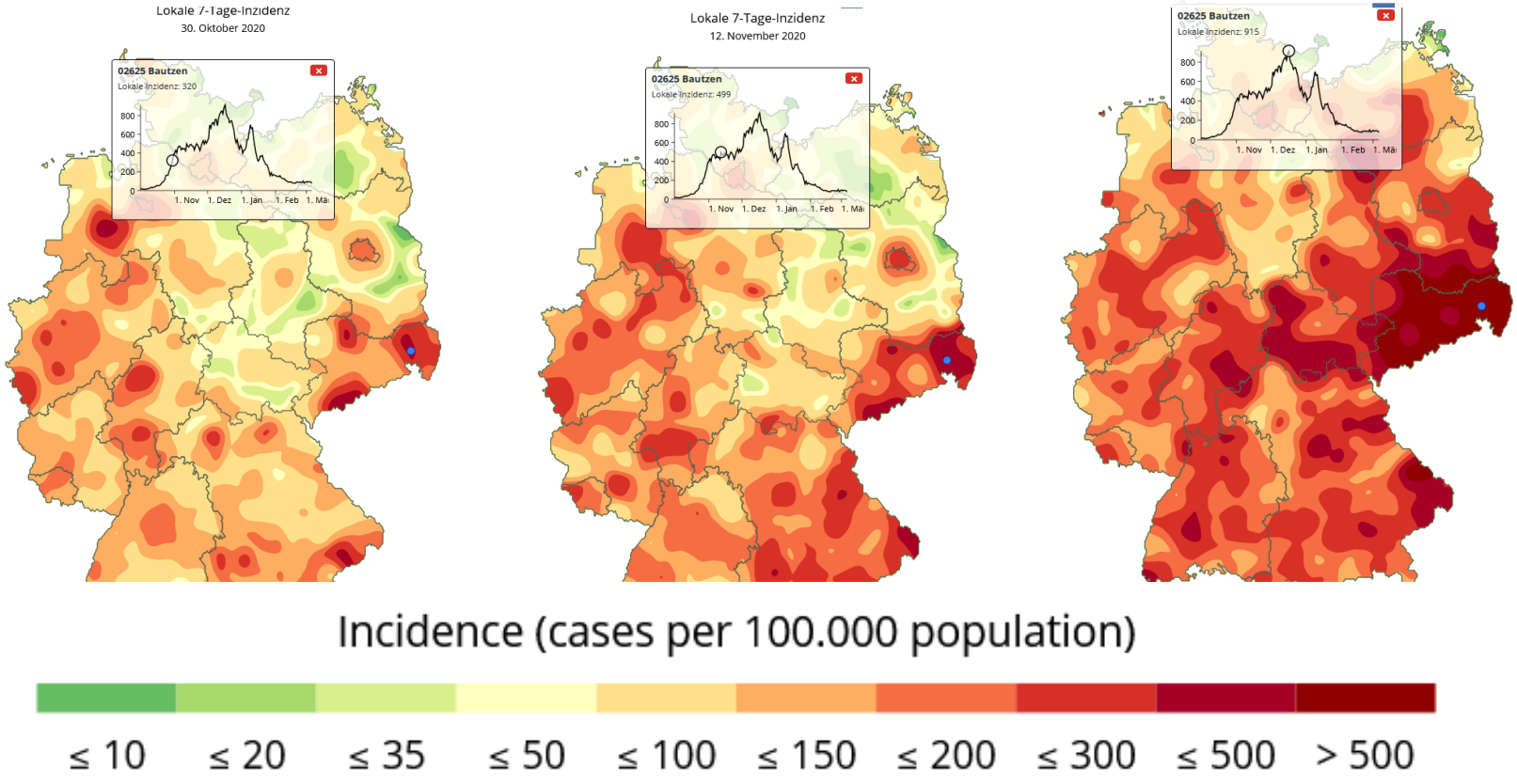
A comparison of Choropleth and Kernelheaping maps

- Though based on the same data (RKI county aggregates) differences due to the joint analysis of neighbouring counties for Kernelheaping map.
- Scattered incidence regions (Choropleth) are united to a smooth region in the west of Germany (Northrhein-Westfalia).
- A high incidence region in the north-west of Germany is detected, which remains a stable high incidence region during the entire second pandemic wave in Germany.

Spatial and temporal comparisons of incidence clusters

- The daily infection numbers exhibit a strong seasonal pattern over the workdays: use of 7-day moving averages!
- For each day a new 7-day moving average is used to display the spatial and temporal development of incidence clusters.
- An internet application uses the RKI county data to display the Corona pandemic in Germany since the start of the second wave in October 2020.
  Link: `https://www.inwt-statistics.com/read-blog/covid-19_heat-map_of-local_7-day_incidences_over_time.html`
- Similar presentations with Choropleths fail: Despite the temporal smoothing by the moving averages we get only an erratic and spurious impression of the temporal development of incidences!
  Link:`https://interaktiv.tagesspiegel.de/lab/corona-analyse-in-welchen-regionen-die-zahlen-wieder-s`

The temporal development of infection rates in Saxony and in the city of Bautzen (Blue mark).



Left: 30. Oct. (Separate Clusters) Middle: 12. Nov (Merge of clusters) Right: 17. Dec. (Entire region infected)

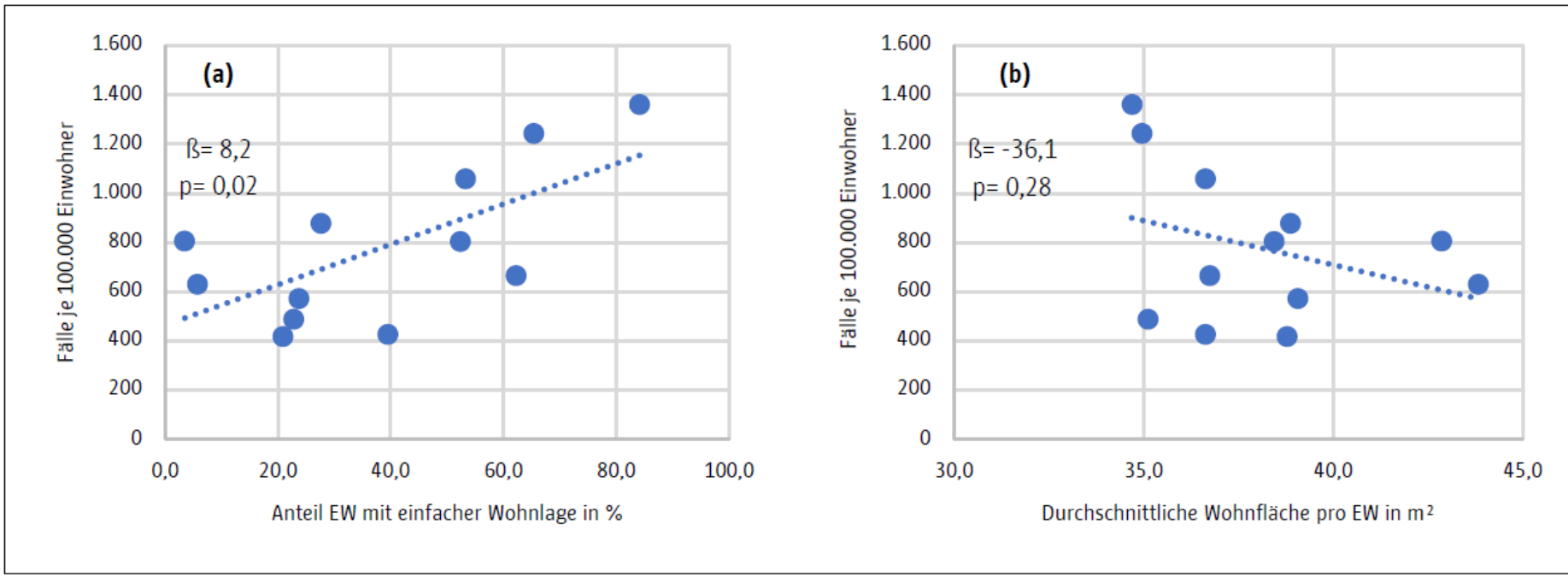
The level of aggregation

- The level of aggregation often depends on the available data.
  - Corona data: RKI data at county level
  - Berlin: 12 Bezirke (BZK City districts), 97 Ortsteile (ORT), 447 Neighbourhoods (LOR)
    Only incidences at BZK-level are delivered from RKI, despite local government (Senator für Gesundheit) asked for data at LOR-Level.
- Level of incidences? The lower, the better?
- Impact of regional data on incidence risks?

Regional Corona analysis for Berlin

Analysis at the level of 12 city districts:

Abbildung 7:
Anzahl der COVID-19-Fälle je 100.000 Einwohnerinnen und Einwohner in den Berliner Bezirken im Zusammenhang mit dem Anteil der Einwohnerinnen und Einwohner (EW) mit einfacher Wohnlage (a) und der durchschnittlichen Wohnfläche je Einwohnerin und Einwohner (b). (Stand: 29.10.2020)



(Datenquelle: RKI-COVID-19-Dashboard, SenStadtWohn, AfS / Berechnung und Darstellung: SenGPG – I A -)

Conclusion: Densely populated districts with low quality housing increase Corona incidences!

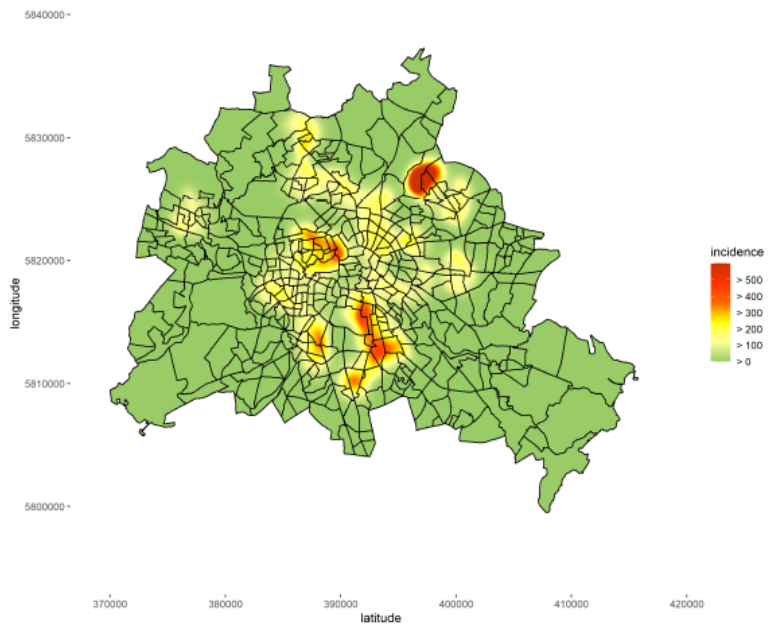
Freie Universität Berlin

A simulation study for Berlin

- How are local true incidences estimated?
- Scope: Comparison of maps (Choropleth vs Kernel density procedure) in case of known infection cases. Need for simulation data!
- Comparison of RMSE over city area.
- Simulation design
  - Start with infection clusters proportional to population density. Total infection numbers according to official records.
  - Generate infections according to official R-values via Poisson distribution.
  - Dispersion: On neighbouring grid points proportional to population density.
  - Duration: Over 16 weeks (approx. duration of the second wave in Berlin)
- Different levels of aggregation for map construction: 12 Districts (BZK), 97 so-called Ortsteile (ORT), 447 Neighbourhoods (LOR).

Comparison of simulated Corona-infections in Week 1

(a) True density

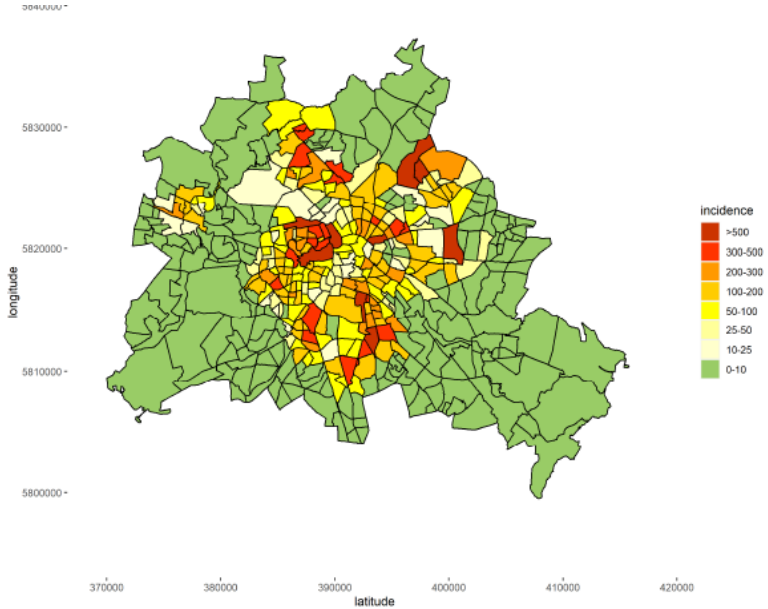


(b) True incidence

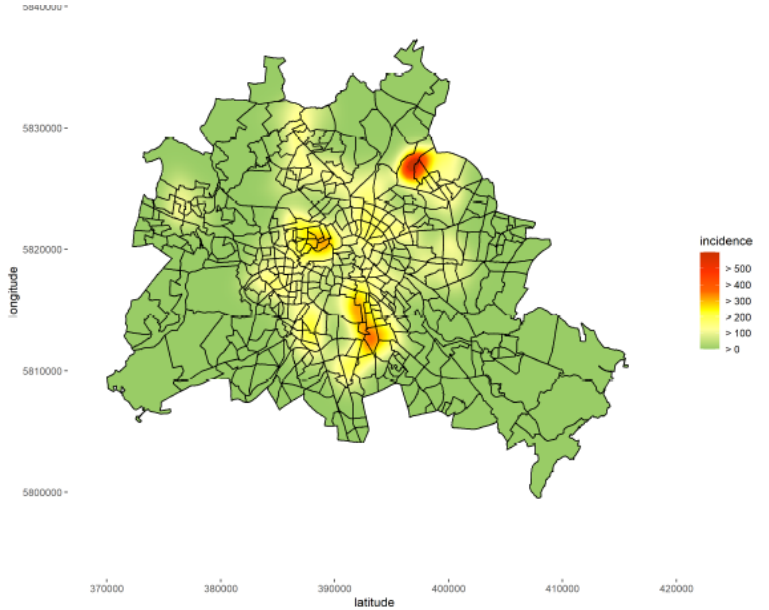
The Kernel density of simulated cases (true density) and the resulting local incidence (true incidence).

Freie Universität Berlin

Comparison of simulated Corona-infections in Week 1



(c) Choropleth incidence(LOR)

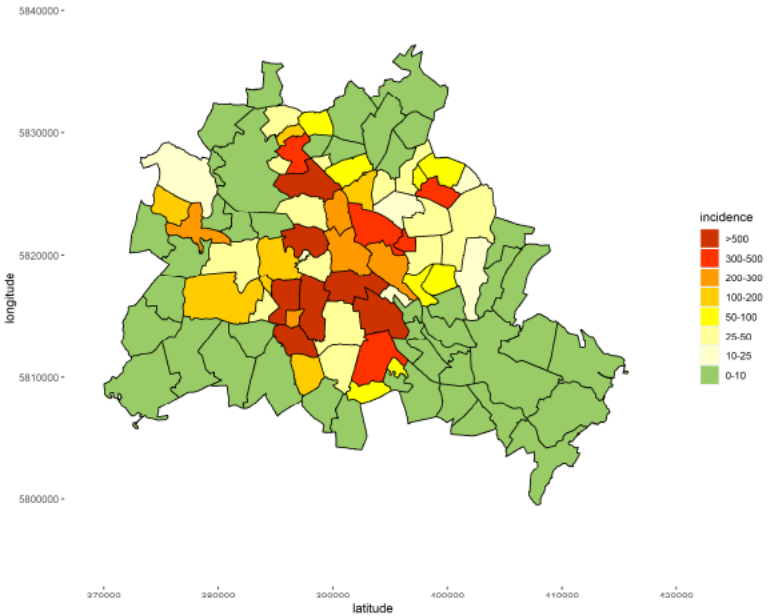


(d) Kernel heaping incidence(LOR)

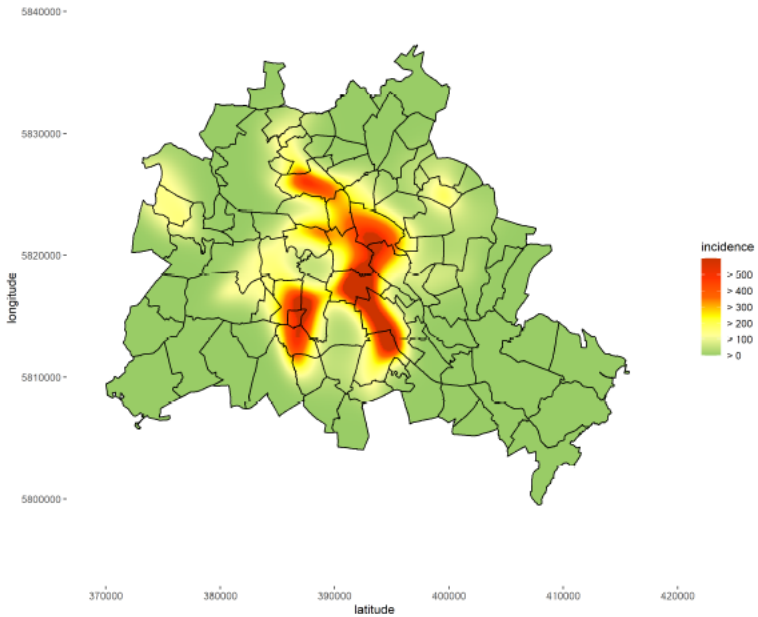
Aggregation level LOR: Choropleth map (left) and Kernel heaping map (right).

Freie Universität Berlin

Comparison of simulated Corona-infections in Week 1



(e) Choropleth incidence(ORT)

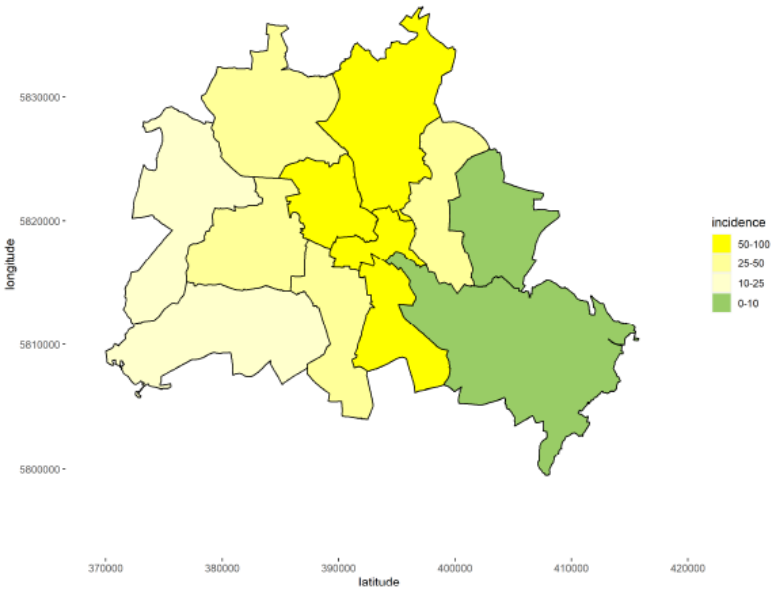


(f) Kernel heaping incidence(ORT)

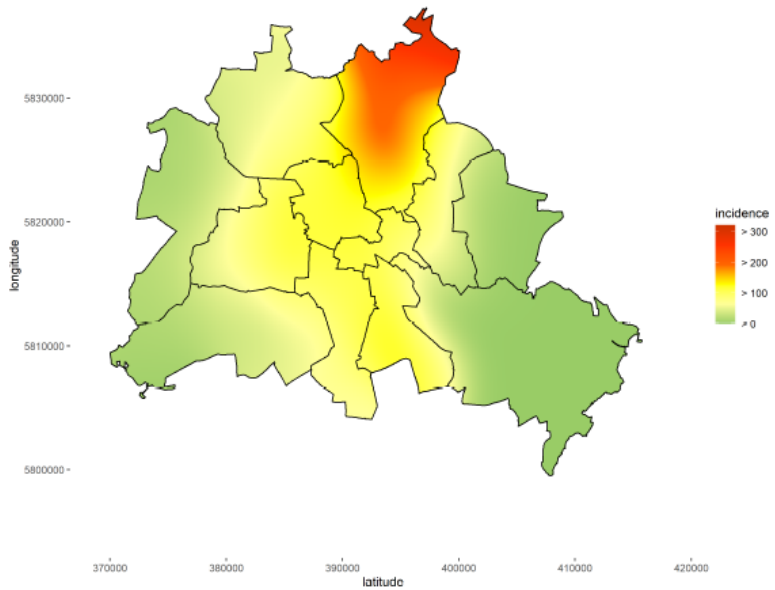
Aggregation level ORT: Choropleth map (left) and Kernel heaping map (right).

Freie Universität Berlin

Comparison of simulated Corona-infections in Week 1



(g) Choropleth incidence(BZK)

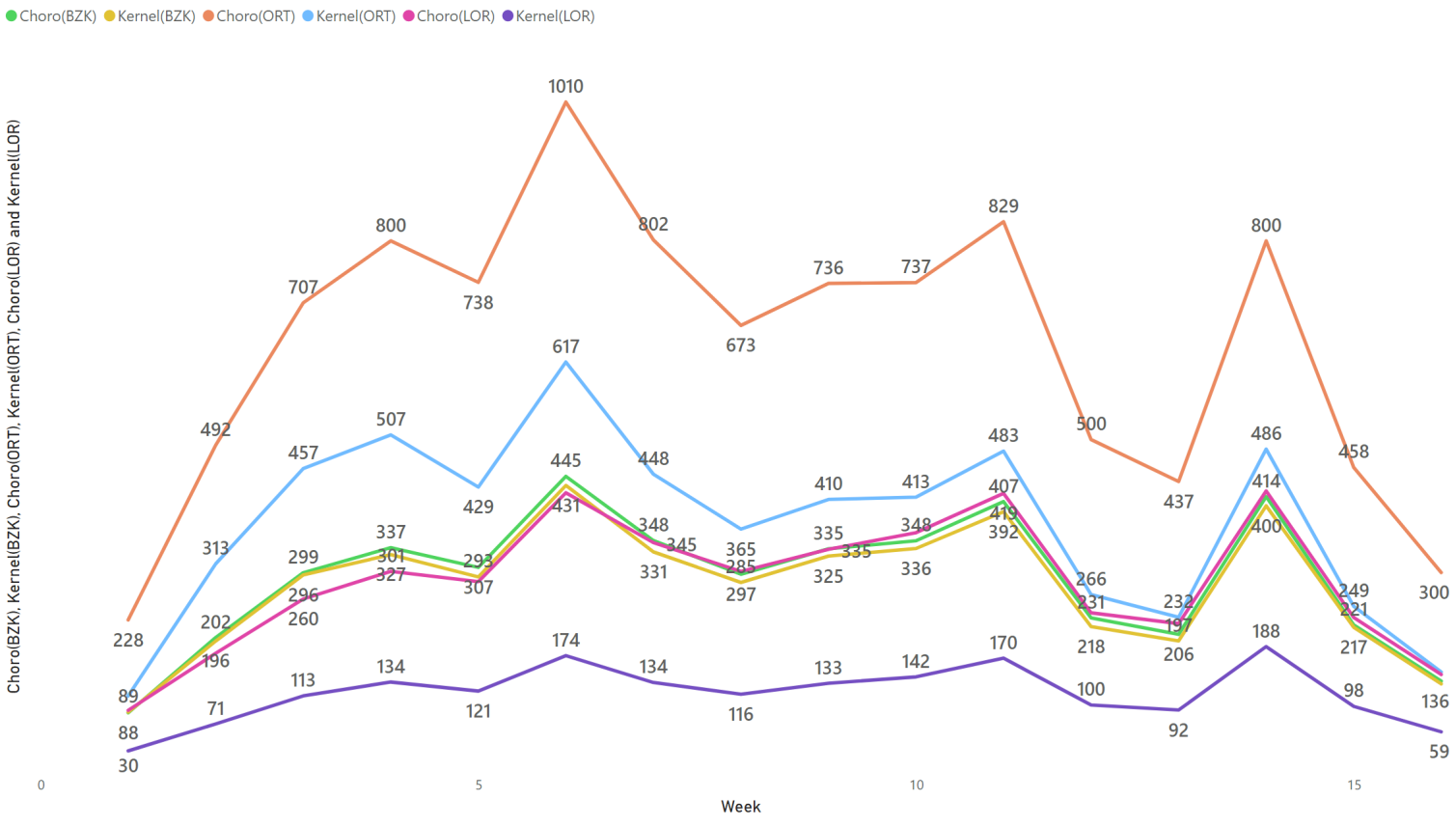


(h) Kernel heaping incidence(BZK)

Aggregation level BZK: Choropleth map (left) and Kernel heaping map (right).

Freie Universität Berlin

Comparison of RMSE over 16 simulation weeks



Upmost lines: ORT level (Red=Choropleth, Lightblue=Kernel), **middle lines:** BZK level (Green=Choropleth, Orange=Kernel) LOR level for Choropleths (Violet), **bottom line:** LOR level of the Kernel heaping map (Darkblue)

Lessons learnt from simulations

- Kernel heaping map has at all levels and at all times smaller RMSE values than the Choropleth map.
- It does **not** hold: the smaller the area level the better the map.
- RMSE of Choropleth maps: There is almost no difference between the LOR level and the BZK level!

Outlook

- Extension of approach to **3-dim Kernel densities**:
  - Longitude, latitude and time
  - Longitude, latitude and age
- **Geostatistics**: Smoothing by Kriging of area centroid values. Ignores the shape of the areas!
- **A complain**: Public media in Germany + RKI: 100 percent use of Choropleth maps!
  I tried to contact newspapers (Spiegel, Tagesspiegel) and the RKI. No reaction!

References incidence maps

- App: https://www.inwt-statistics.com/read-blog/covid-19_heat-map_of-local_7-day_incidences_over_time.html
- Use of App: https://www.inwt-statistics.com/read-blog/the-representation-of-corona-incidence-figures-in-space-and.html
- Use of App in German: Rendtel, U.; Neudecker, A.; Fuchs, L.(2021): Die Darstellung von Inzidenzgebieten mit simulierten Geokoordinaten. AStA Wirtschafts- und Sozialstatistisches Archiv, 15, https://doi.org/10.1007/s11943-021-00288-x
- *Kernelheaping* Package: Version 2.2.8 (May 2021) https://cran.r-project.org/web/packages/Kernelheaping/Kernelheaping.pdf

References SEM algorithm

- **Grouped income data**: Walter,P.; Groß,M.; Schmid,T.; Tzavidis, N. (2021): Domain prediction with grouped income data. JRSS A https://doi.org/10.1111/rssa.12736
- **Location of special populations**: Groß,M.; Rendtel, U.; Schmid,T.; Schmon,S.; Tzavidis,N. (2017): Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error. JRSS A , 180, 161 – 183.
- **Change of support**: Groß, M.; Kreutzmann, A.-K.; Rendtel, U; Schmid, T.; Tzavidis, N. (2020): Switching between different area systems via simulated geo-coordinates: A case study for student residents in Berlin. J. Official Statistics, 36, 297 – 314, http://dx.doi.org/10.2478/JOS-2020-0016
- **Service maps + open data**: Rendtel, U.; Ruhanen, M. (2018): Die Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarfs an Kinderbetreuung in Berlin. AStA Wirtschafts- und Sozialstatistisches Archiv, 12, 271–284 https://doi.org/10.1007/s11943-018-0235-y
- **Voting analysis**: Erfurth, K; Groß, M; Rendtel, U; Schmid, T. (2021): Kernel density smoothing of composite spatial data on