Towards continuous domain models in Spatial Epidemiology

Garyfallos Konstantinoudis

Institute of Social and Preventive Medicine

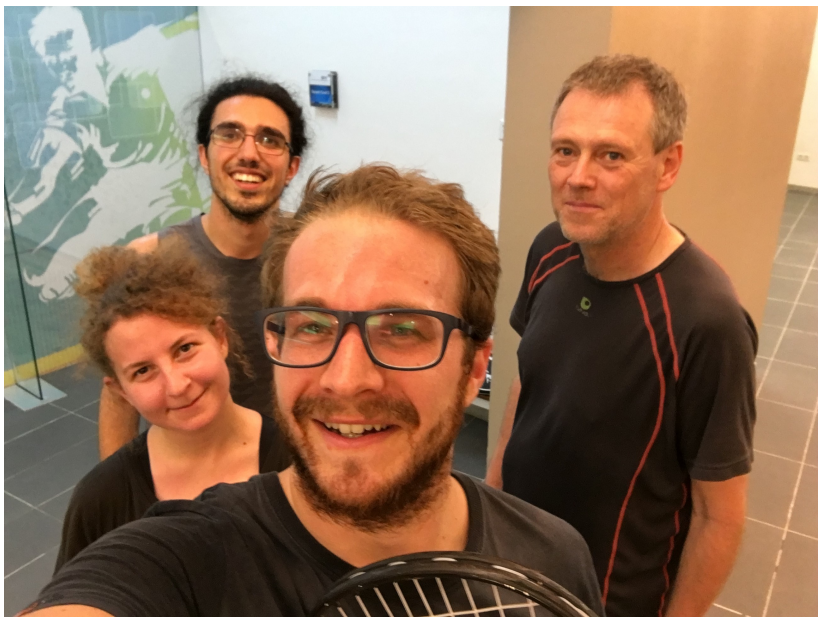
June 17, 2018

Outline

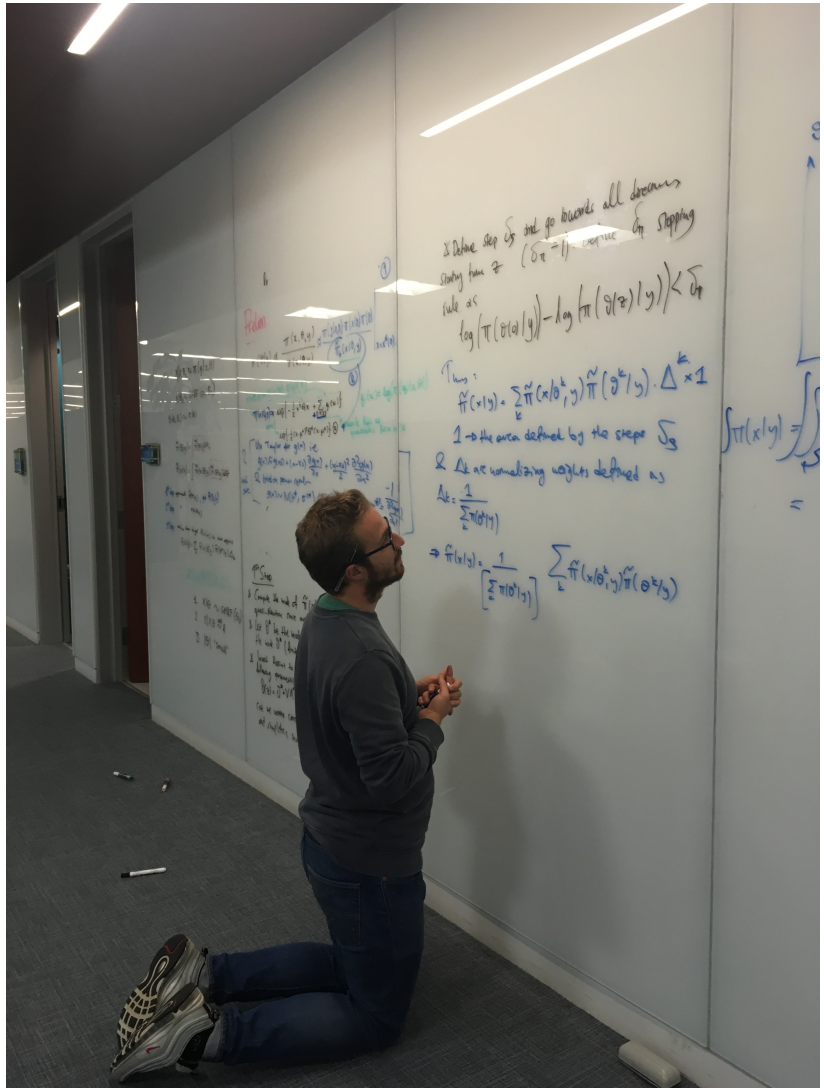
The campus

The food

The People



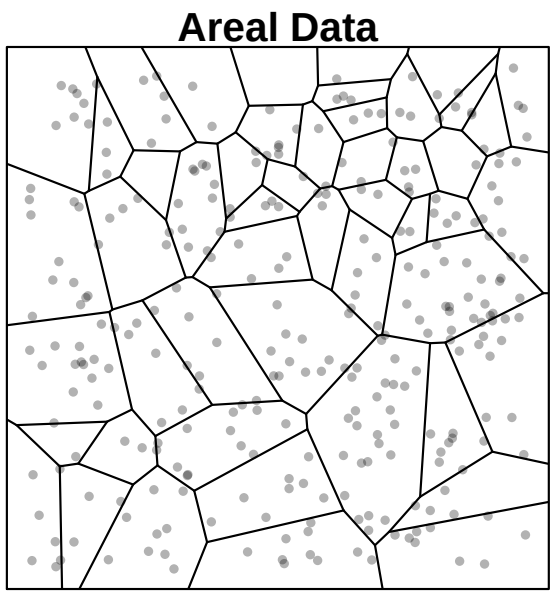
The Frustration

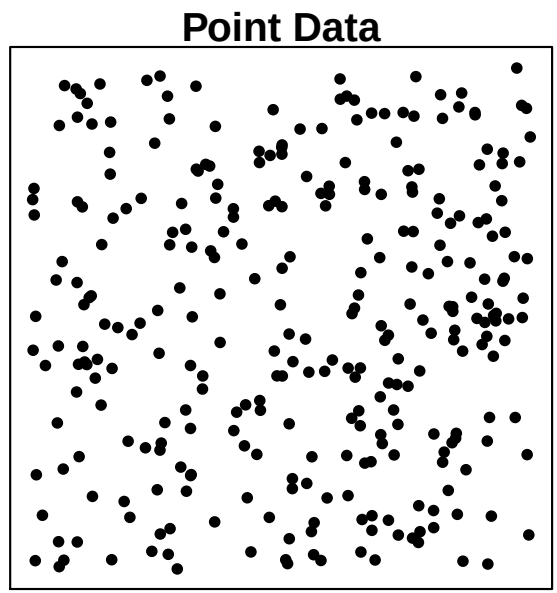


Introduction

Background

- Disease mapping of cancers
  - generating hypotheses
  - hotspots of environmental pollution
  - target areas for health intervention
- Areal data and BYM models
- Point data and LGCPs





Motivation

- Childhood leukaemia: 5.4 per $100,000$ person years
- Leukaemia clusters: Sellafield, Woburn, Fallon
- Possible risk factors: air-pollution, pesticides, ionising radiation etc.
- High resolution geographical data in Switzerland


EXCLUSIVE

Nineteen children living within 20 miles of the Sellafield nuclear plant have developed leukaemia, a cancer affecting their blood and bone marrow. In a few weeks' time their parents will begin proceedings to sue British Nuclear Fuels for compensation. If they win, it will make legal history. Four of these families have told their story exclusively to the Observer Magazine. Report by Alison Whyte and Annabel Ferriman. Photographs by Christopher Pillitz

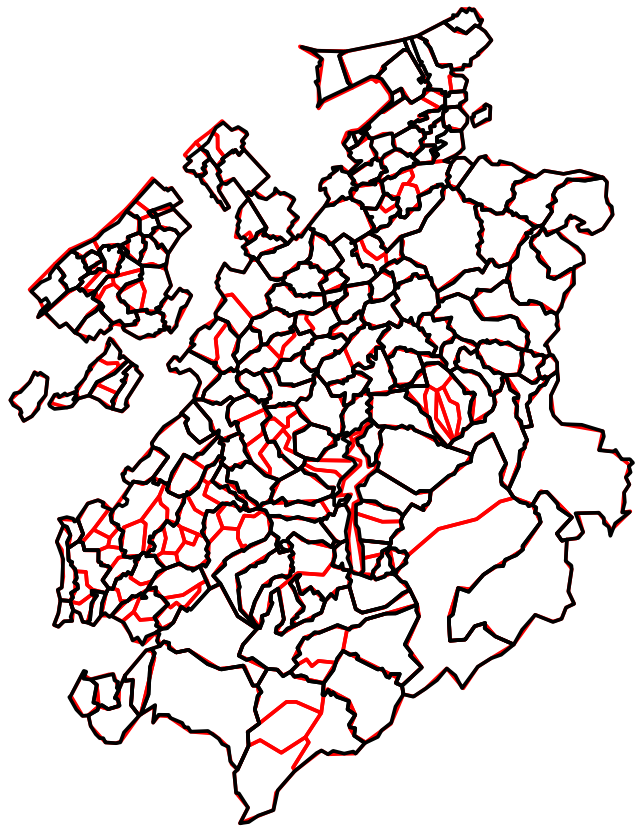
SELLAFIELD AND THE SUFFERING CHILDREN

Previous studies

- Areal data: Conditional Autoregressive (CAR) models
  - Besag Ann Inst Statist Math 1991
  - Faure et al. European Journal of Cancer Prevention 2009, Thompson et al. Cancer Causes & Control 2007, Manda et al. Eur J Epidemiol 2009.
- Precise data: Log Gaussian Cox process (LGCP)
  - Møller et al. Scand J Stat 1998
  - Cancer mapping: Lung cancer in Spain (Diggle et al. Stat Sci 2013), Colon and rectum in Minesota (Liang et al. Ann Appl Stat 2008)
  - none for childhood cancers.
- Simulation studies
  - Lung and stomach cancer (Li et al. J R Stat Soc C-Appl 2012)
  - Syphilis (Li et al. Methods in Medical Research 2012)
  - Cancer mortality (Kang et al. PLOS one 2013)

LGCPs have some apparent advantages

- Bypass all the problems arisen when selecting arbitrary boundaries
- Step-wise risk function might be a strong assumption
- Use all data sources available (spatial misalignment)
- Ecological bias



Aim

Does LGCP povide additional benefits over the BYM model on aggregated data when:

- Quantifying the risk in space
- Identify high-risk areas

Methods

Data Availability

Cases

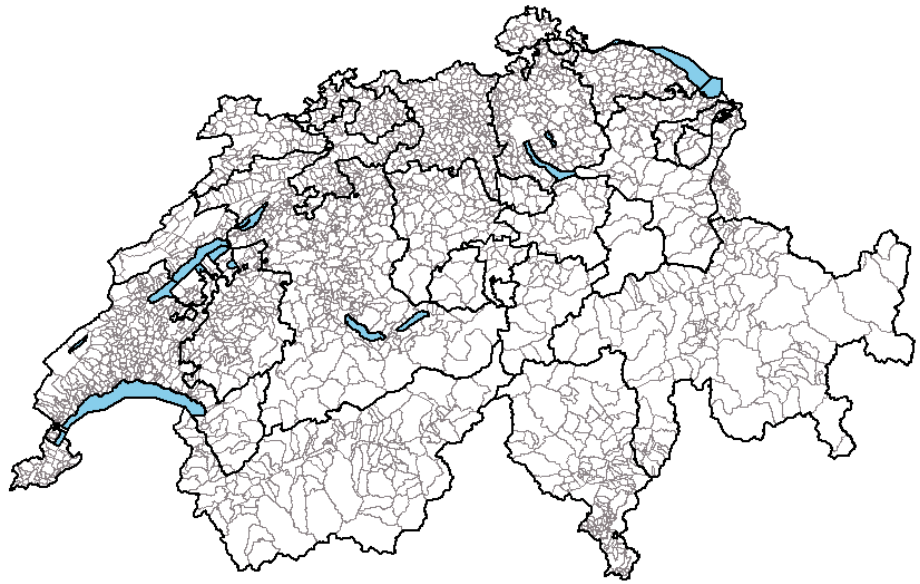
- Swiss Childhood Cancer Registry (SCCR)
- $>$ 90% completeness since 1985
- Precise location

Population at risk

- Census (1990, 2000, 2010 onwards)
- Precise location

Geographical units in Switzerland

- 26 Cantons
- 2352 municipalities



BYM model

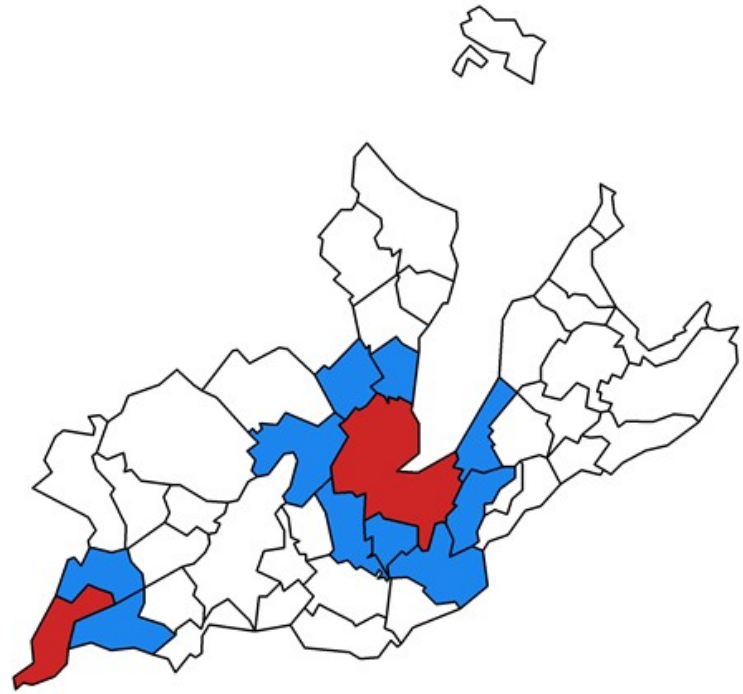
Let $\mathcal{W}$ an observation window, $A_1, ..., A_N$ a partition of $\mathcal{W}$, Y_i be the disease counts P_i the population and λ_i the risk in A_i:

$$Y_i|\lambda_i, P_i \sim \text{Poisson}(\lambda_i P_i)$$

$$\log(Y_i) = log(P_i) + \beta_0 + u_i + v_i$$

$$u_i|\mathbf{u}_{-i} \sim \mathcal{N}\Big(\frac{\sum_{j=1}^{N} w_{ij} u_j}{\sum_{j=1}^{N} w_{ij}}, \frac{1}{\tau_1 \sum_{j=1}^{N} w_{ij}}\Big)$$

$$v_i \sim \mathcal{N}(0, \tau_2^{-1})$$



LGCP model

Let $\mathcal{W}$ an observation window and Ξ a point process with intensity $\lambda(s)$ on the s location:

$$\Xi|\lambda(s) \sim \text{Poisson}(\int_{\mathcal{W}} \lambda(s)ds)$$

$$\log(\lambda(s)) = log(\lambda_0(s)) + \beta_0 + u(s)$$

$$u(s) \sim \text{GF}(0, \mathbf{\Sigma}(h, \tau, \phi))$$

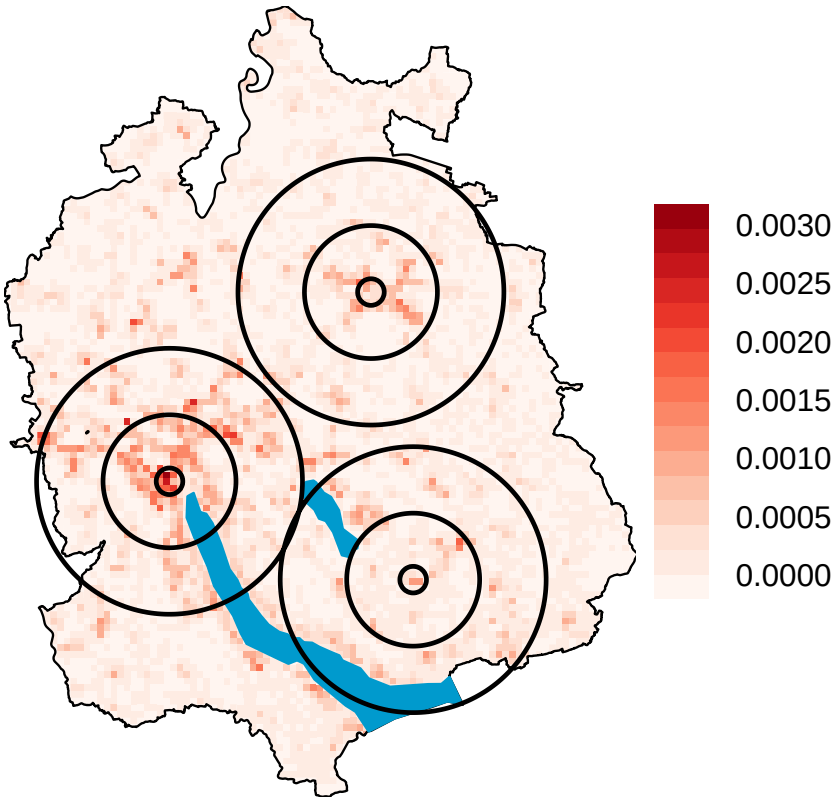
$$\kappa(h) = \tau^2 \rho_\nu(h/\phi), \rho_\nu(\cdot) \text{ Matern}$$

- Inference for both models was conducted with Integrated Nested Laplace Approximation (INLA)

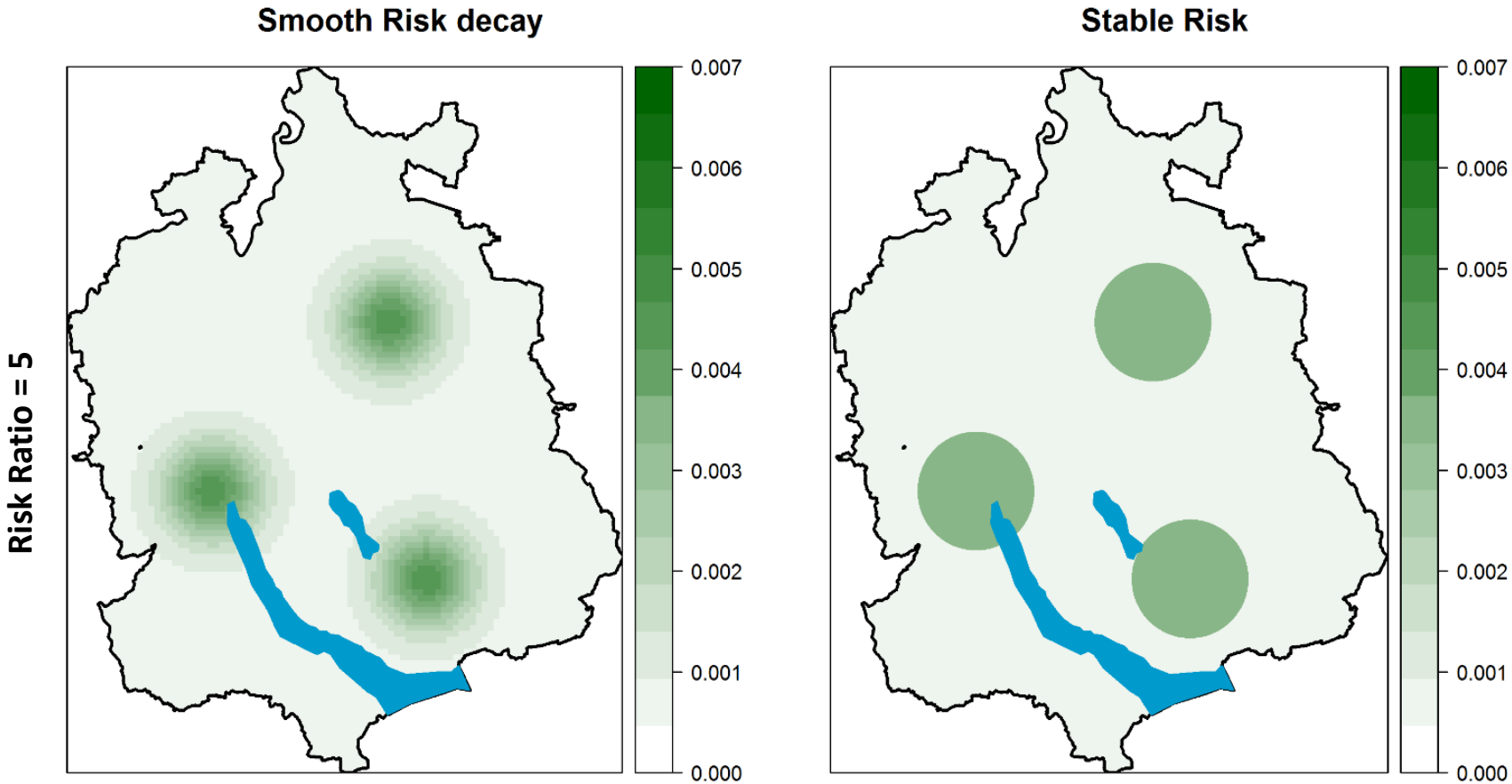
Data simulation: Different scenarios

- Canton of Zurich
- $N = 205,242$ (15%) children
- Leukaemia incidence 1985-2015 ($n = 334$)

Radius	RR	times n	decay
1km	2	1	step function
5km	5	5	smooth function
10km	-	10	-



Data Simulation: True risk for RR = 5 & radii = 5km



Data simulation: Simulation metrics

- Root mean integrated square error (RMISE)

$$\text{RMISE} = \left(\mathbb{E} \int_{\mathcal{W}} b(s)(log(\hat{\lambda}(s)) - log(\lambda(s)))^2 ds \right)^{1/2} =$$

$$\left(\mathbb{E} \sum_{g=1}^{G} b_g |D_g| (log(\hat{\lambda}_g) - log(\lambda_g))^2 \right)^{1/2}$$

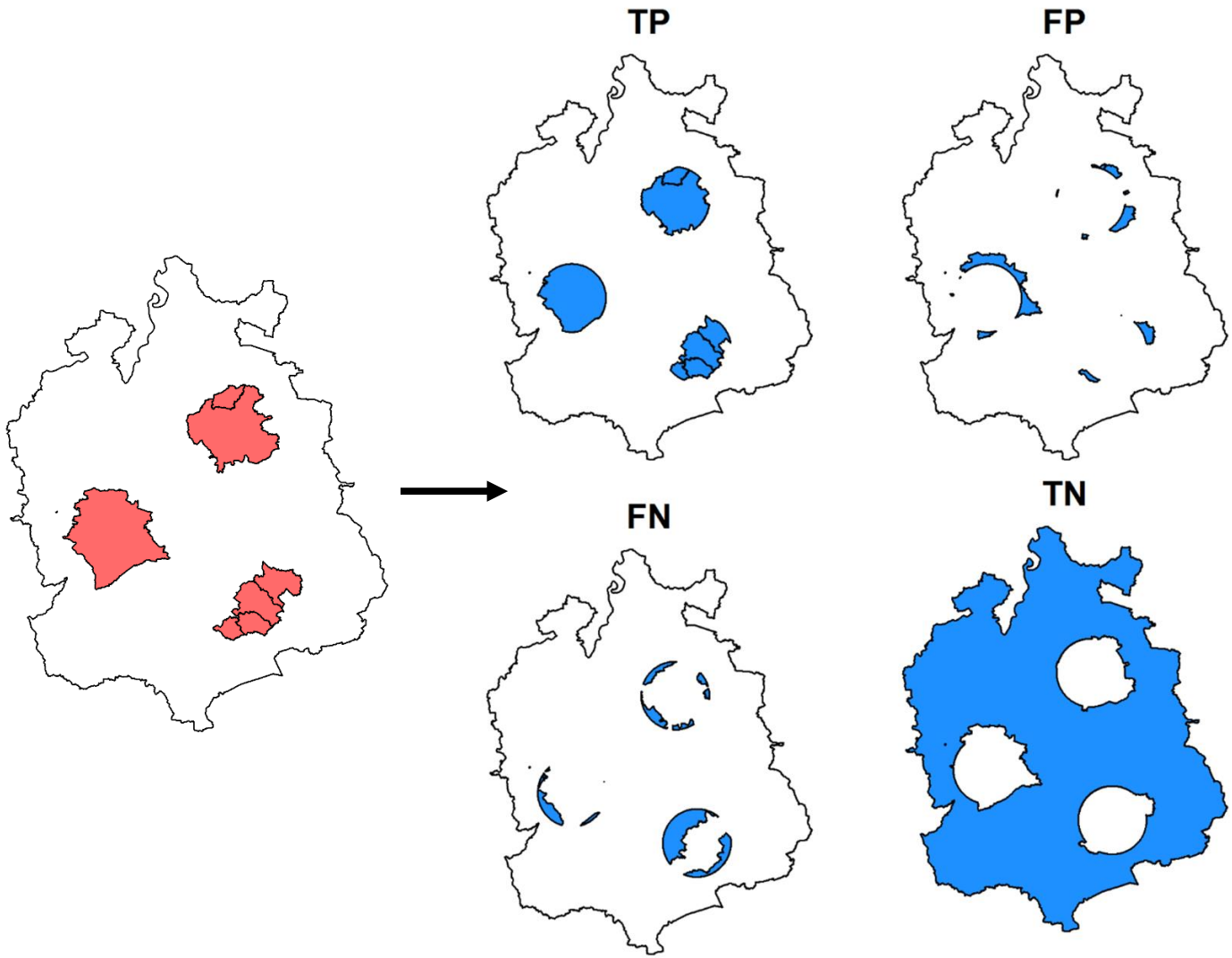
- Coverage probability: $p = \frac{\sum_{g=1}^{G} \kappa_g}{G}$, where

$$\kappa_g = \begin{cases} 1, & \lambda_g \text{ lies inside the credibility region of } \hat{\lambda_g} \\ 0, & \text{otherwise} \end{cases}$$

- Sensitivity, Specificity and area under the curve (AUC)

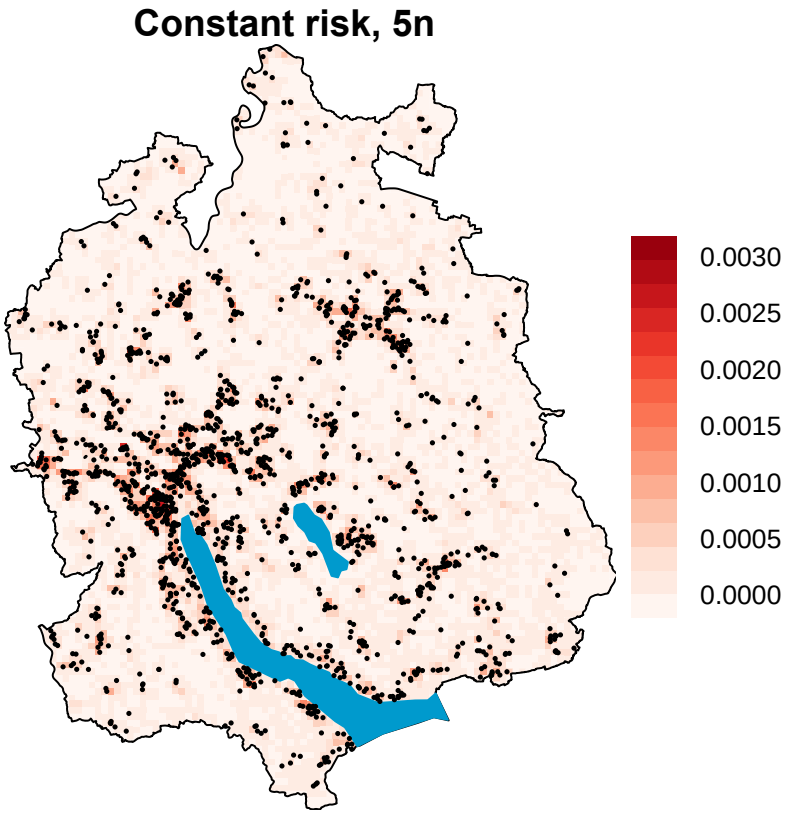
Data simulation: Sensitivity Specificity

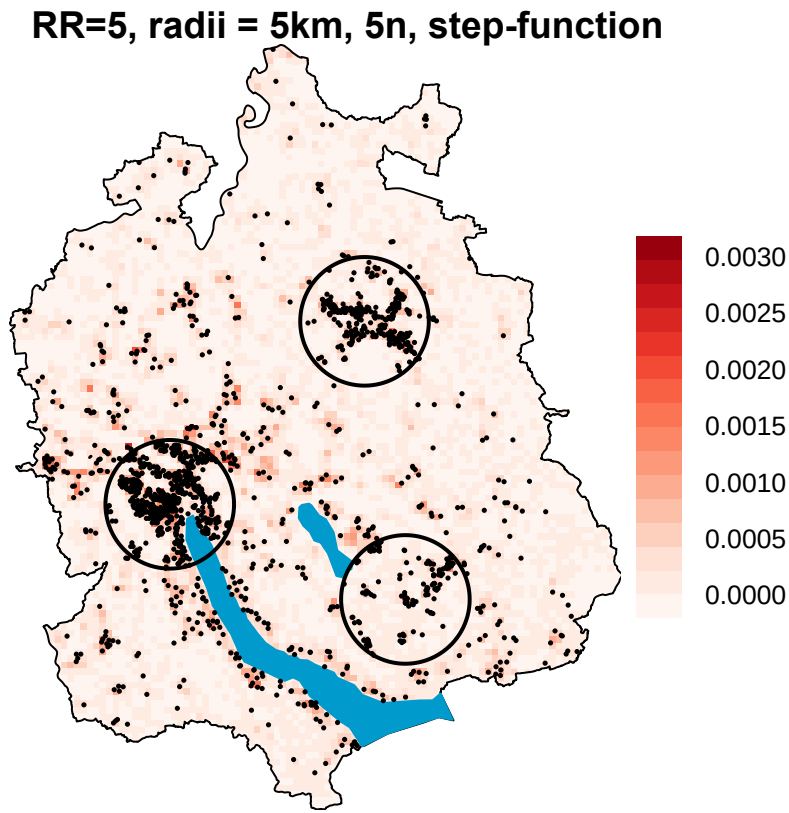
- Exceedance probability, $\Pr[\lambda_g > \frac{n}{N}] > \alpha$, $\alpha = 0, 0.05, 0.10, ...$



Results

Example of simulated datasets





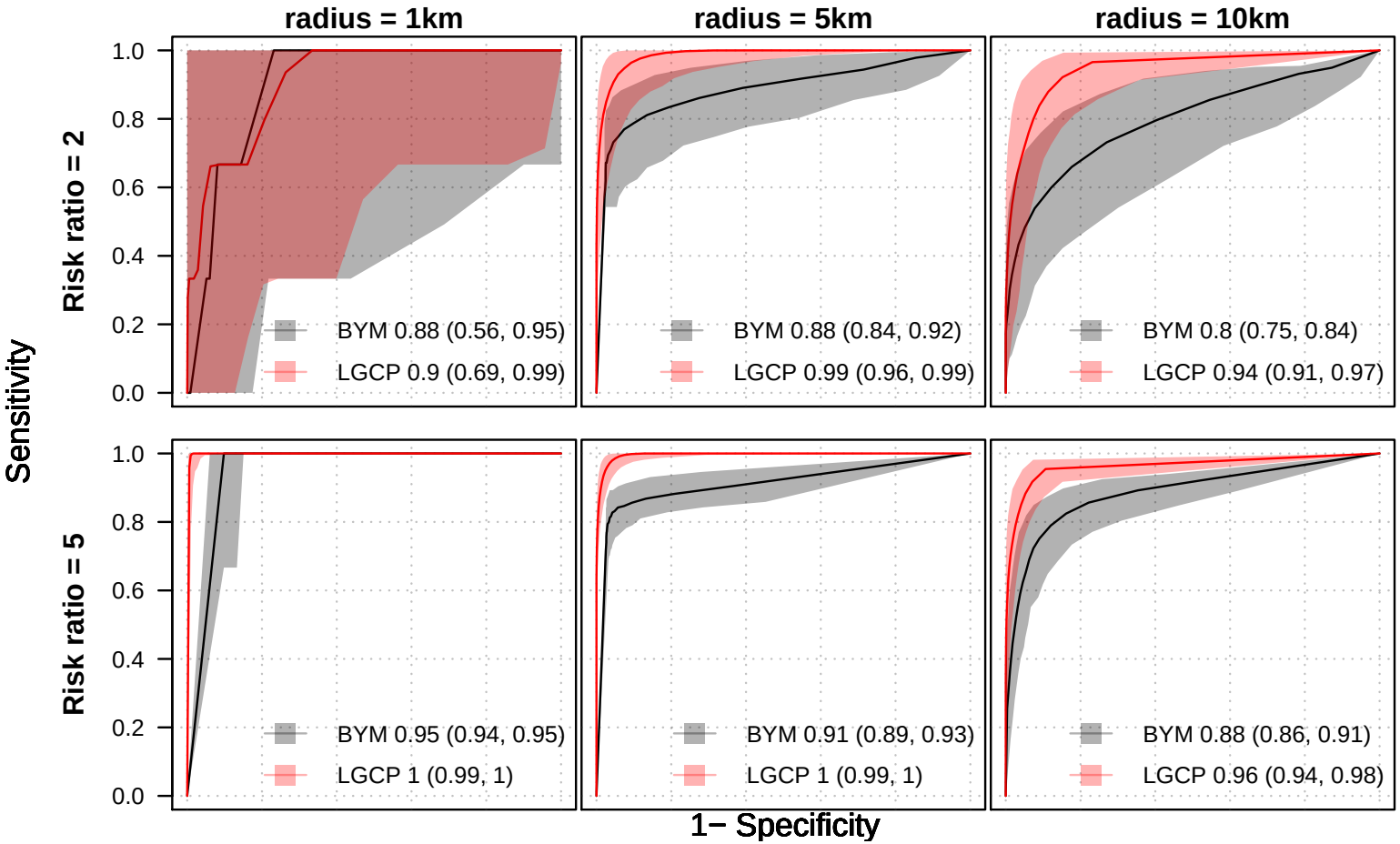
RMISE, $5n$

	BYM	LGCP
Step function		
Radius = 1km		
RR = 2	4.47 (3.17, 6.81)	6.62 (4.24, 9.88)
RR = 5	10.4 (8.77, 12.5)	14.8 (13.1, 17.1)
Radius = 5km		
RR = 2	11.6 (10.6, 13.1)	12.2 (10.8, 14.7)
RR = 5	22.8 (21.4, 24.5)	21.5 (19.6, 24.6)
Radius = 10km		
RR = 2	14.9 (14.3, 15.8)	12.1 (11, 14.4)
RR = 5	28.4 (27.3, 29.8)	22.3 (20.8, 24.6)
Smooth function		
Radius = 1km		
RR = 2	4.48 (3.1, 6.88)	6.51 (4.27, 9.9)
RR = 5	10.8 (8.82, 12.5)	14.8 (13, 16.8)
Radius = 5km		
RR = 2	10.4 (9.32, 12)	11 (9.33, 14.3)
RR = 5	19.2 (18, 20.6)	16.8 (14.8, 19.9)
Radius = 10km		
RR = 2	12.3 (11.5, 13.4)	10.1 (8.57, 12.7)
RR = 5	21.8 (21, 22.8)	13.9 (12.1, 17)

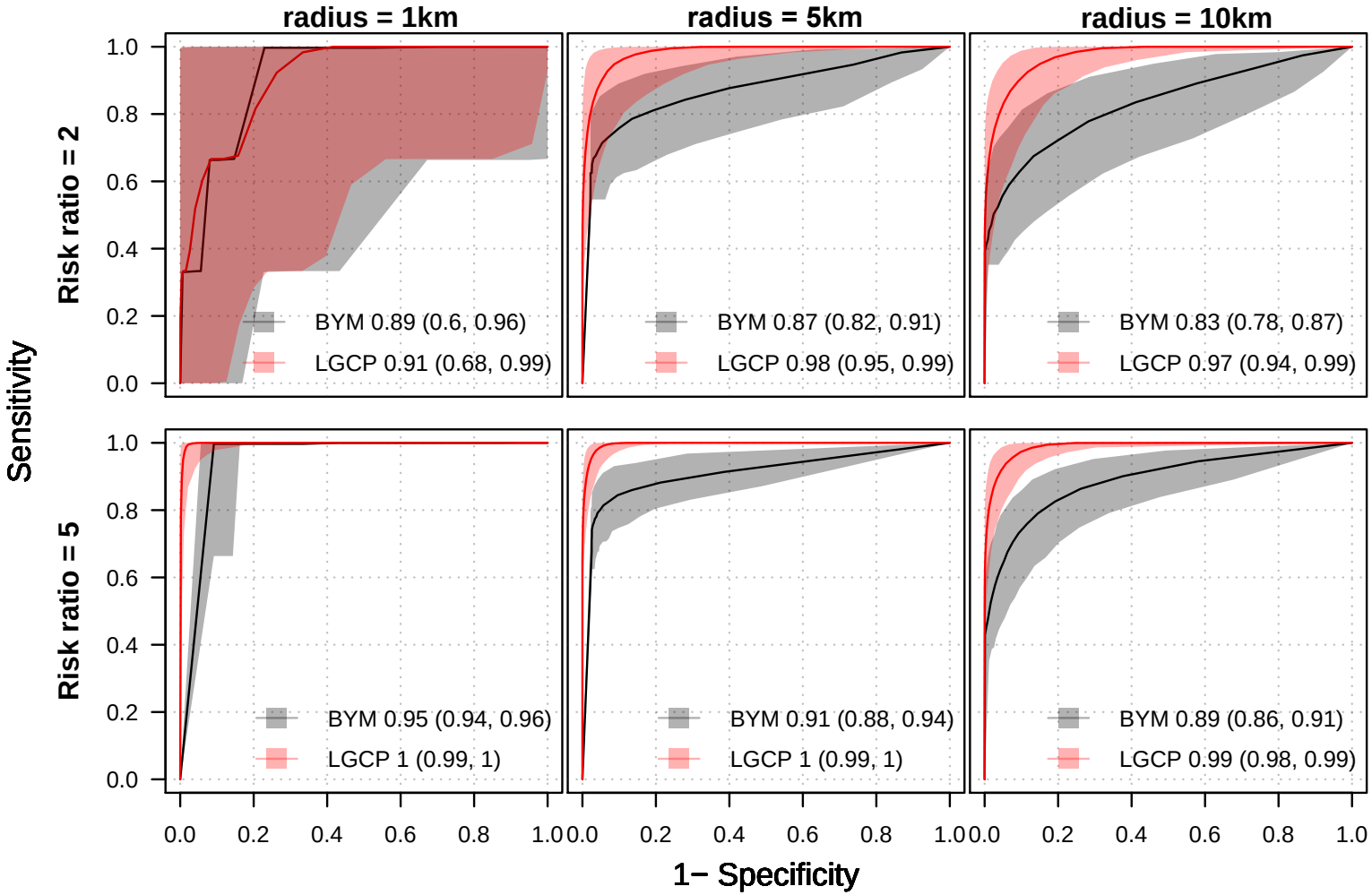
Coverage probability, $5n$

	BYM	LGCP
Step function		
Radius = 1km		
RR=2	0.94(0.91,0.99)	0.99(0.89,1)
RR=5	0.9(0.88,0.9)	0.99(0.98,0.99)
Radius = 5km		
RR=2	0.9(0.85,0.94)	0.95(0.89,0.97)
RR=5	0.88(0.84,0.9)	0.92(0.89,0.94)
Radius = 10km		
RR=2	0.88(0.52,0.95)	0.88(0.8,0.94)
RR=5	0.85(0.76,0.9)	0.84(0.78,0.88)
Smooth function		
Radius = 1km		
RR = 2	0.95(0.91,0.99)	0.99(0.88,1)
RR = 5	0.9(0.89,0.93)	0.99(0.98,1)
Radius = 5km		
RR = 2	0.92(0.9,0.93)	0.99(0.93,1)
RR = 5	0.87(0.85,0.89)	1(0.96,1)
Radius = 10km		
RR = 2	0.93(0.82,0.95)	1(0.94,1)
RR = 5	0.85(0.71,0.9)	0.99(0.96,1)

ROC-curves, Step-function, $5n$

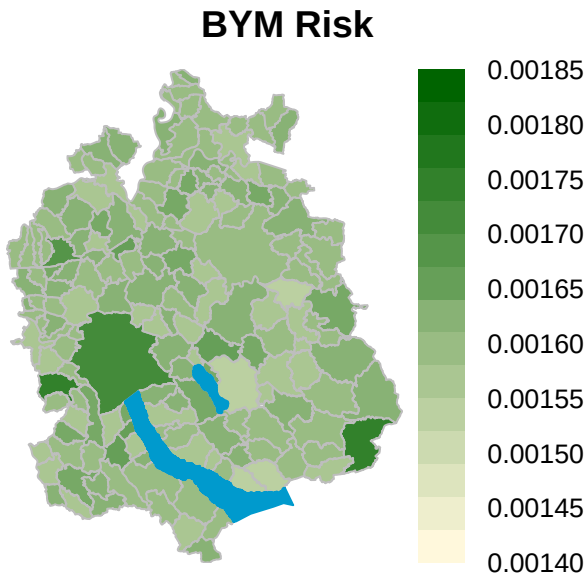


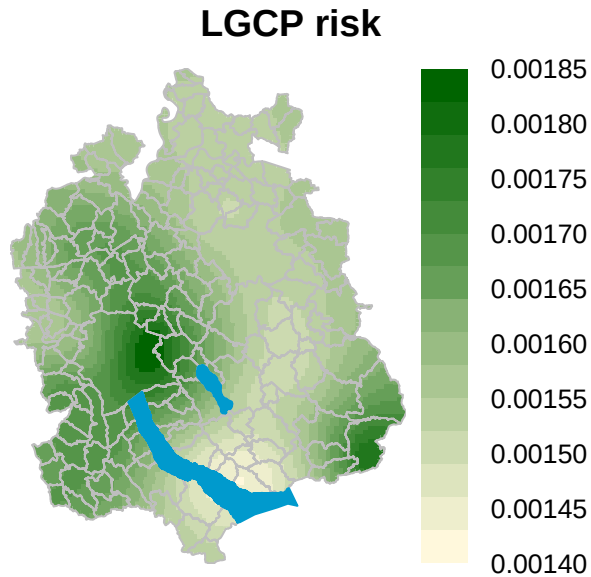
ROC-curves, Smooth-function, $5n$

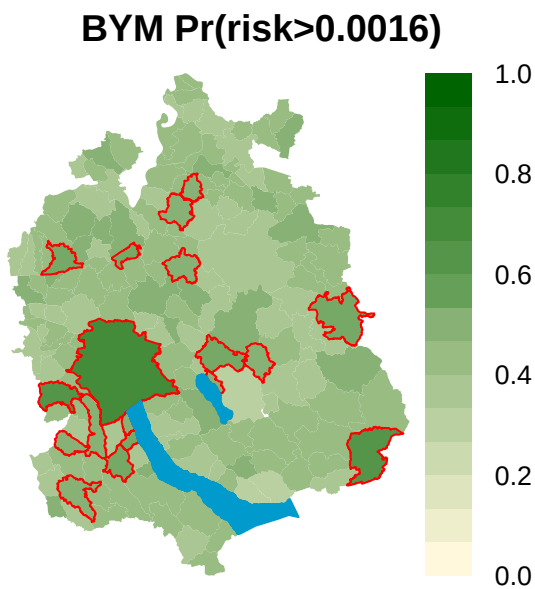


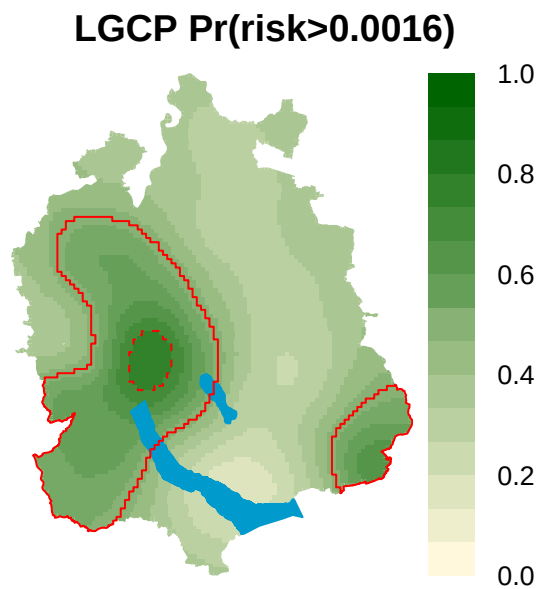
Example: Chilhdood leukaemia in the canton of Zurich

- $\Pr(\lambda(s) > \frac{n}{N}) > 0.50$ (red solid line)
- $\Pr(\lambda(s) > \frac{n}{N}) > 0.80$ (red dotted line)









Could this be chance?

Mean sensitivity and specificity for $\alpha = 0.50$ and $\alpha = 0.80$ based on the simulation keeping the number of cases n.

	$\alpha = 0.5$				$\alpha = 0.8$			
	BYM		LGCP		BYM		LGCP	
scenarios	sens	spec	sens	spec	sens	spec	sens	spec
Step;r = 1km;RR = 2	0.58	0.75	0.36	0.86	0.10	0.99	0.09	0.99
Smooth;r = 1km;RR = 2	0.59	0.75	0.36	0.86	0.10	0.99	0.08	0.99
Step;r = 5km;RR = 2	0.65	0.93	0.72	0.91	0.52	0.98	0.33	0.99
Smooth;r = 5km;RR = 2	0.64	0.93	0.69	0.92	0.50	0.98	0.31	1.00
Step;r = 10km;RR = 2	0.31	0.93	0.42	0.95	0.08	1.00	0.09	1.00
Smooth;r = 10km;RR = 2	0.46	0.95	0.59	0.92	0.30	1.00	0.21	1.00
Step;r = 1km;RR = 5	0.86	0.83	0.89	0.88	0.50	0.94	0.62	0.98
Smooth;r = 1km;RR = 5	0.90	0.83	0.92	0.89	0.59	0.94	0.63	0.99
Step;r = 5km;RR = 5	0.72	0.96	0.87	0.97	0.64	0.97	0.61	1.00
Smooth;r = 5km;RR = 5	0.70	0.96	0.83	0.97	0.62	0.98	0.55	1.00
Step;r = 10km;RR = 5	0.39	0.97	0.53	0.97	0.18	1.00	0.21	1.00
Smooth;r = 10km;RR = 5	0.50	0.96	0.71	0.96	0.38	1.00	0.38	1.00

Discussion

Main results

- Overall LGCPs perform better compared to BYM models in almost all scenarios considered
- Our results are consistent with the literature
- We identified an area of higher leukaemia risk in the canton of Zurich
- Possible explanations: Failure to correct for population density or environmental risk factors as air-pollution

Strengths and Limitations

Strengths

- Extensive simulation study assessing 39 scenarios
- Datasets with characteristics corresponding to a realistic population
- Fair comparison between the models

Limitations

- Did not incorporate the effect of spatially varying covariates
- Did not examine different inferential approaches (MCMC, MALA, HMC etc.)
- There are always more interesting scenarios to be considered

Discussion

- The study highlights the importance of availability of precise geocodes for disease mapping
- Data confidentiality considerations
- Present data on a resolution compatible with patient confidentiality requirements

Our study points towards continuous domain models in Spatial Epidemiology

Avenues for future research

- Extent the comparison by incorporating the effect of spatially varying covariates
- Examine different methods for identifying high-risk areas (excursion sets, quantile regression)
- Compare this model-based approach with the popular circular scan for cluster detection
- Examine methods for preserving the high resolution but also data confidentiality (integrate estimates in larger areal units, jittering, etc.)
- The study sets the scene for more research on the continuous domain models, with possible extension to continuous space and time models

Next steps

- Publish the paper
- Applying LGCPs in chilhdood cancer incidence in Switzerland
- Apply for a grant to address some of the previous slide's objectives
- Defend on the 7th of December

Thanks to..

Any questions?