Spatial data analysis

Data Science Africa 2016
Ricardo Andrade

UCSF Global Health Sciences
Global Health Group

Outline

- The geographic context
- Geostatistics
- Non-linear models
- Discrete processes
- Time interactions
- Showcase: catchment area models in Zambia

So far…

- Different models:
  - Regression
  - Classification
  - Clustering
- How they work:
  - Relation between $\boldsymbol{X}$ {inputs/features} and $\boldsymbol{y}$ {output}
  - Distance/similarity between objects inputs
  - Distance/similarity between output instances

Geographic context

- Some problems have a spatial context that should not be dismissed
  - Road accidents
  - Pollution studies
  - Household income
  - Health studies
- Any difference if $\boldsymbol{X}$ are coordinates?

Geostatistics

- Geostatistics is a field concerned with continuous spatial variation.
- We have random realizations of a process z at any location $\boldsymbol{x} = (x_1, x_2)$,

$$z(\boldsymbol{x}) = \mu + \varepsilon(\boldsymbol{x})$$

- There is a spatial correlation at some scale given by

$$C(\boldsymbol{h}) = \langle \varepsilon(\boldsymbol{x})\varepsilon(\boldsymbol{x}+\boldsymbol{h}) \rangle$$

Geostatistics

- Geostatistics is a field concerned with continuous spatial variation.
- We have random realizations of a process z at any location $\boldsymbol{x} = (x_1, x_2)$,

$$y = \mu + f_{\boldsymbol{x}}$$

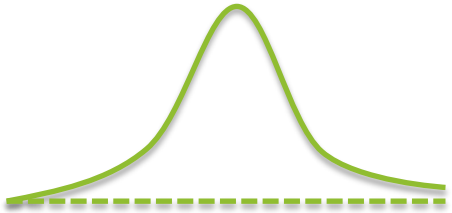
$$z(\boldsymbol{x}) = \mu + \varepsilon(\boldsymbol{x})$$

- There is a spatial correlation at some scale given by

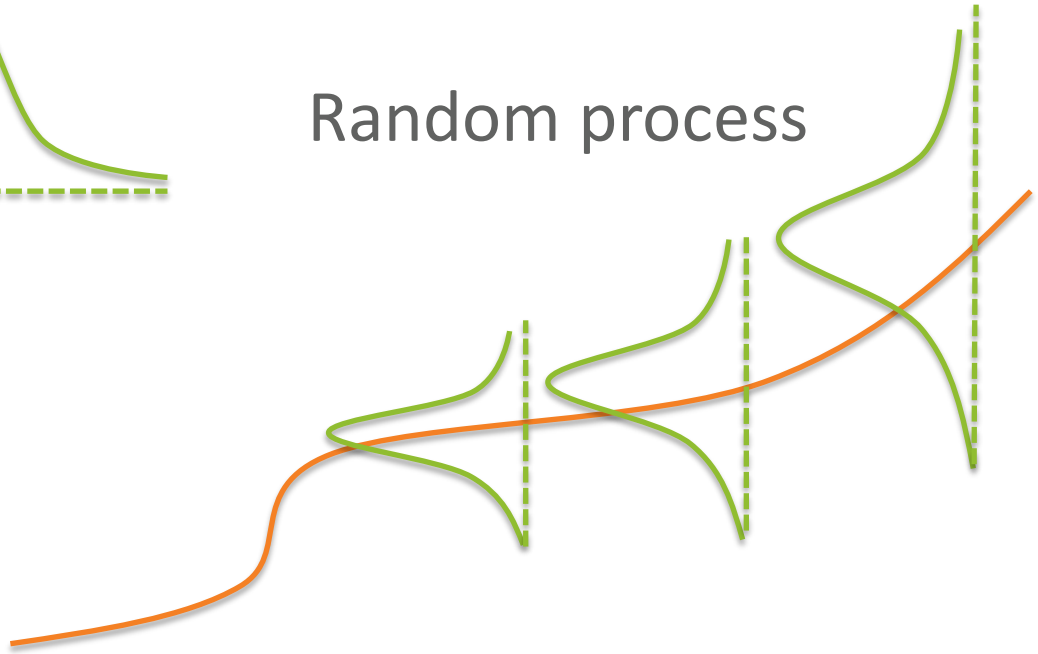
$$\boldsymbol{K}_{ff} = K(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

$$C(\boldsymbol{h}) = \langle \varepsilon(\boldsymbol{x})\varepsilon(\boldsymbol{x}+\boldsymbol{h}) \rangle$$

Random variable



Random process



Stationarity

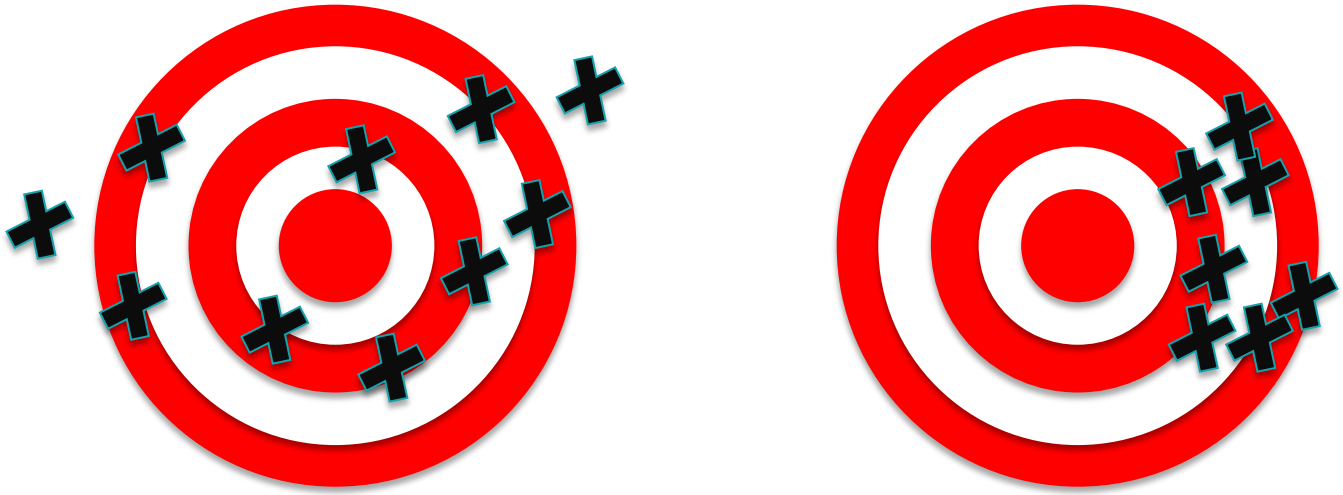
- Stationarity allows to assume the same degree of variation from place to place.
- Weaker assumption:
  - Constant mean
  - Covariance depends on the separation of points, but no their location

All that matters are the orientated distances between the points

- Gaussian distributions are defined by their first two moments.
- Centered Gaussian distributions are uniquely defined by their covariance.
- The study of Gaussian processes is in many ways the study of covariance functions.

Kriging

- A generic term for a range of least square methods to compute *best linear unbiased predictors (BLUP)* for spatial modelling.



Only linear?

- Linear regression model

$$y = \mu + f_{\boldsymbol{x}}$$

- Generalized linear model

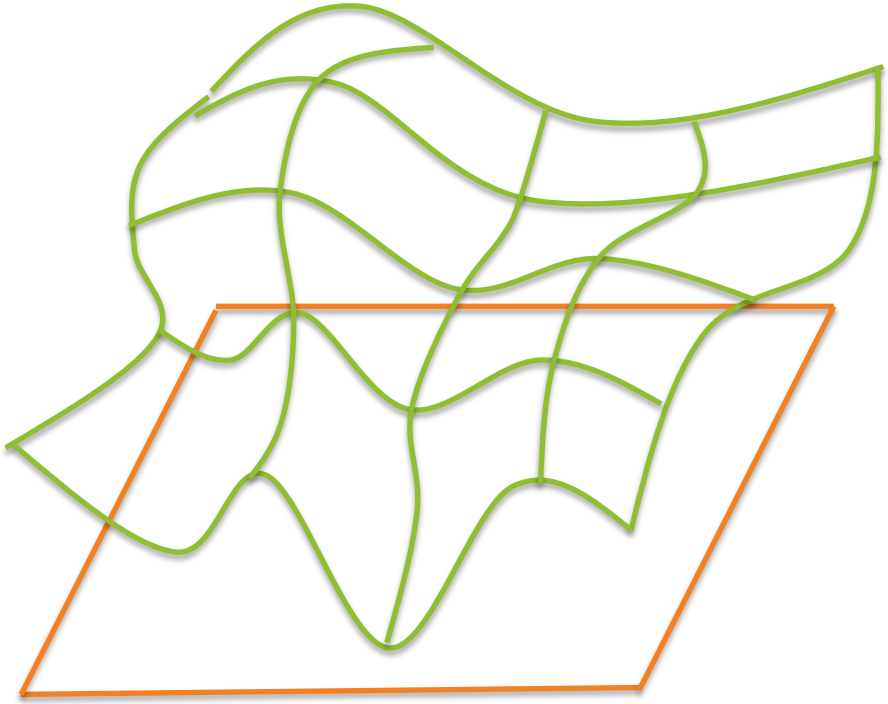
$$\eta = \mu + f_{\boldsymbol{x}}$$

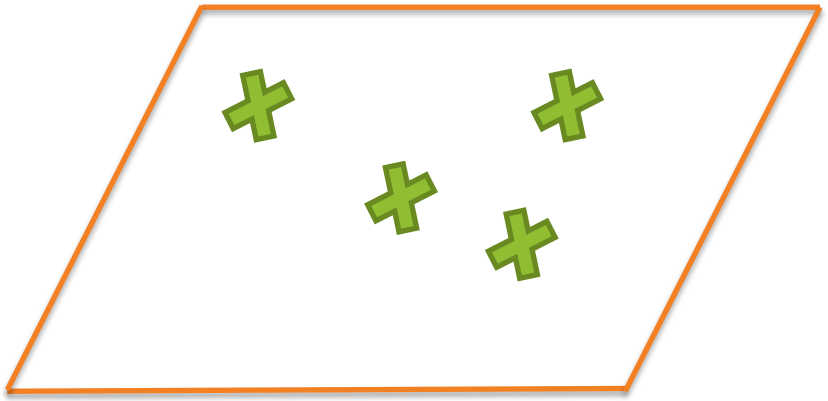
$$y = g(\eta)$$

- With a non-linear transformation exact inference becomes analytically intractable.
- Solutions: MCMC, Laplace approximation, expectation propagation, etc.

Only continuous processes?

- A point process is a stochastic process characterized by generating a countable set of events $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$ across a region.





Only continuous processes?

- A point process is a stochastic process characterized by generating a countable set of events $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$ across a region.
- Poisson process

$$y \sim \text{Poisson}(\lambda)$$

- Log-Gaussian Cox process

$$\lambda = \exp(\mu + f_{\boldsymbol{x}})$$

Time context

- Temporal variation can be equally important.

$$\eta = \mu + f_{\boldsymbol{x}} + [t]$$

  - Parametric

$$\eta = \mu + f_{\boldsymbol{x}} + \beta t$$

  - Non-parametric

$$\eta = \mu + f_{\boldsymbol{x}} + f_t$$

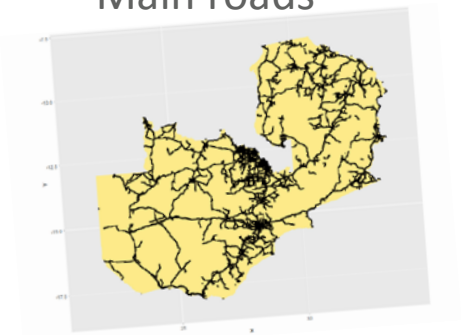
$$\eta = \mu + f_{(\boldsymbol{x},t)}$$

Catchment area model in Zambia

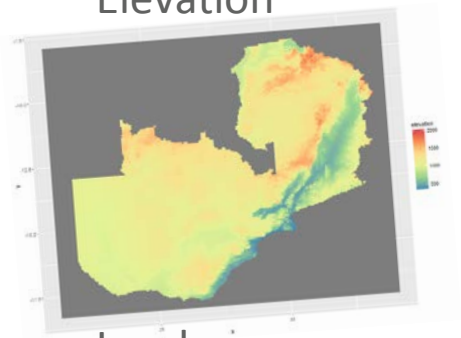
- Hospitals records are an imperfect measure of disease incidence:
  - Only individuals who sought treatment
  - Patients home address is often unknown
- Objectives:
  - Understand the **drivers of treatment seeking**
  - Estimate the **spatial distribution of the patients** attending a health facility

Information used

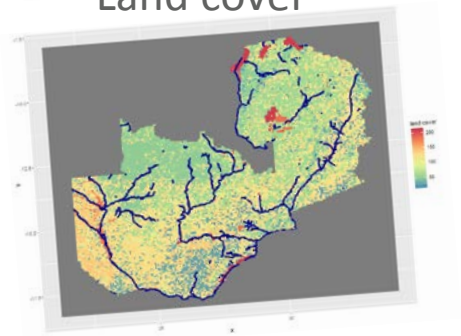
Main roads



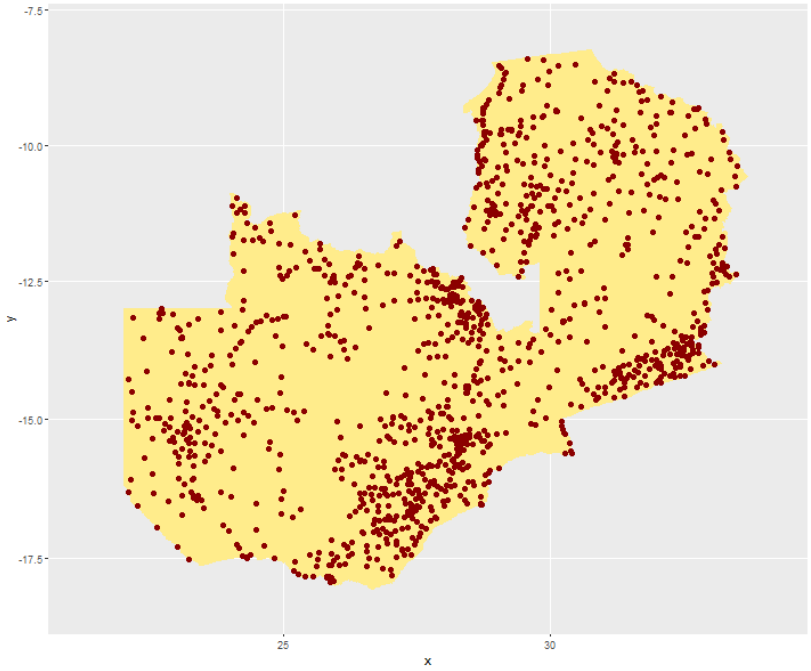
Elevation



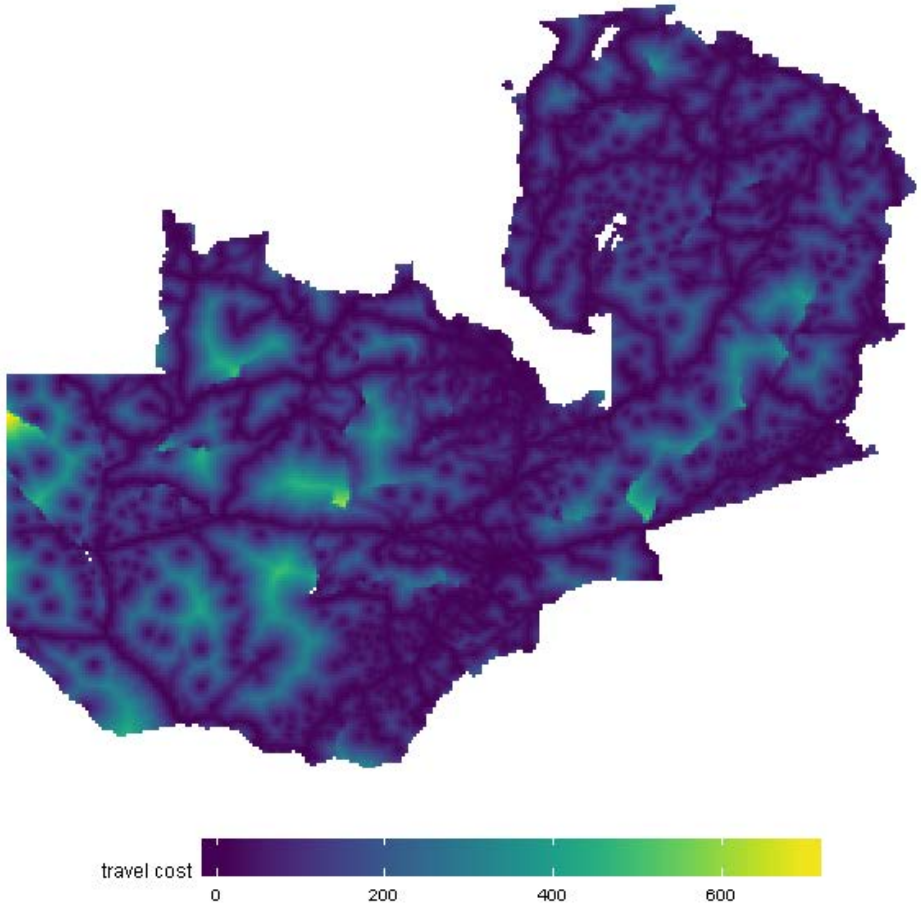
Land cover



Health facilities location Sample of household location and facilities attended



Travel cost to reach the closest facility




Seek treatment?

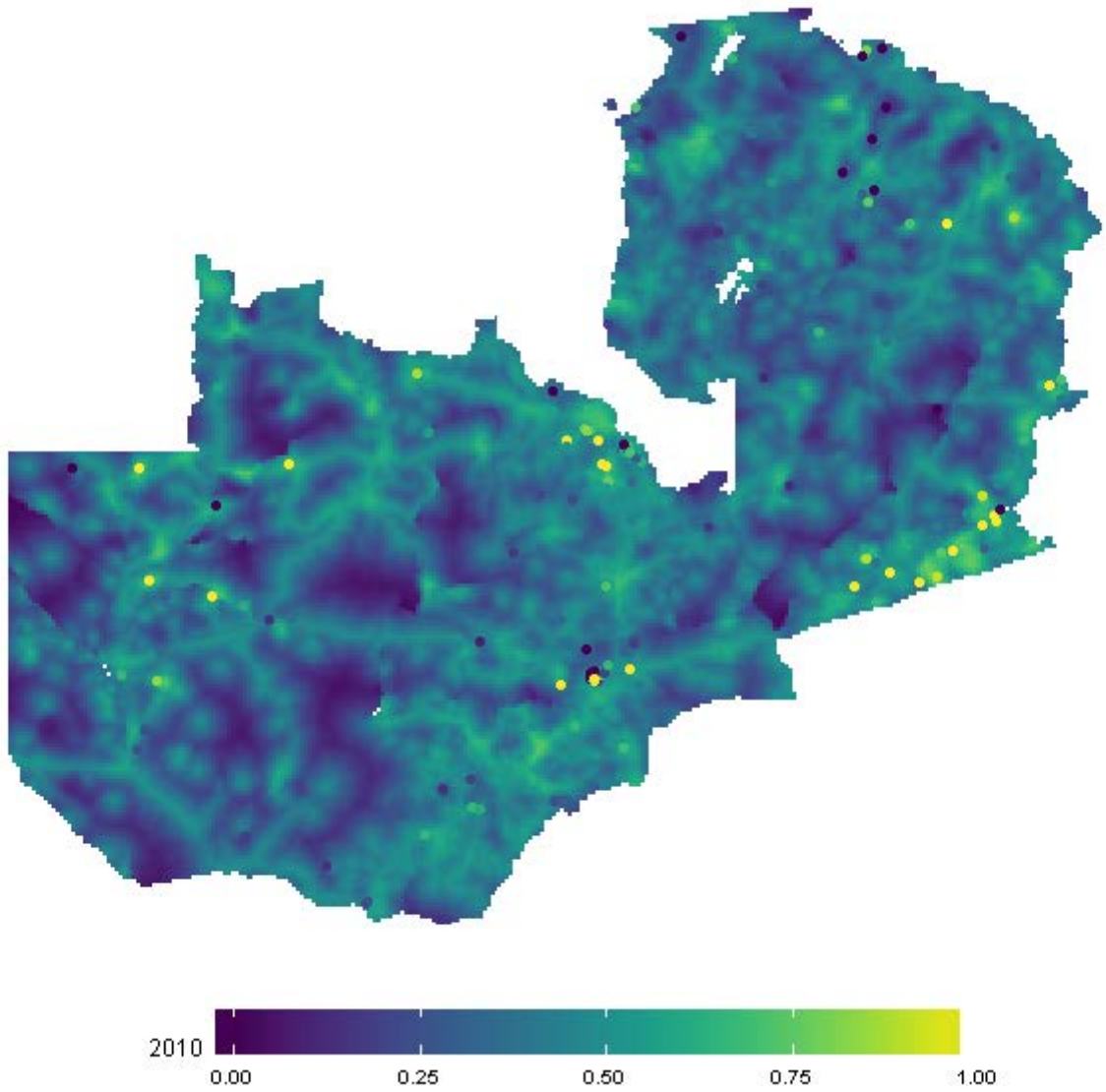
Spatiotemporal model to define the likelihood of seeking treatment

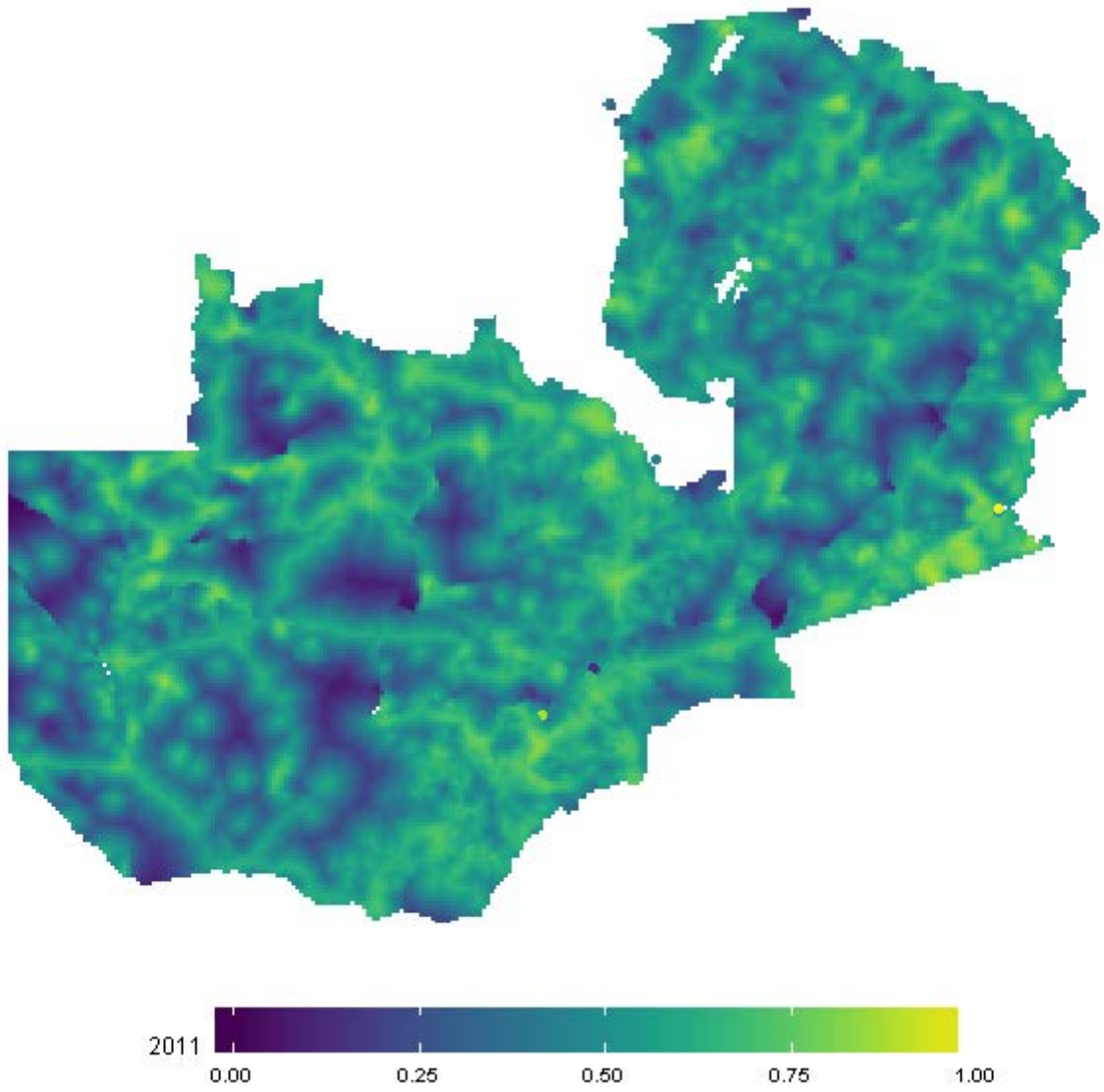
- Let the number of people seeking for treatment be modeled as

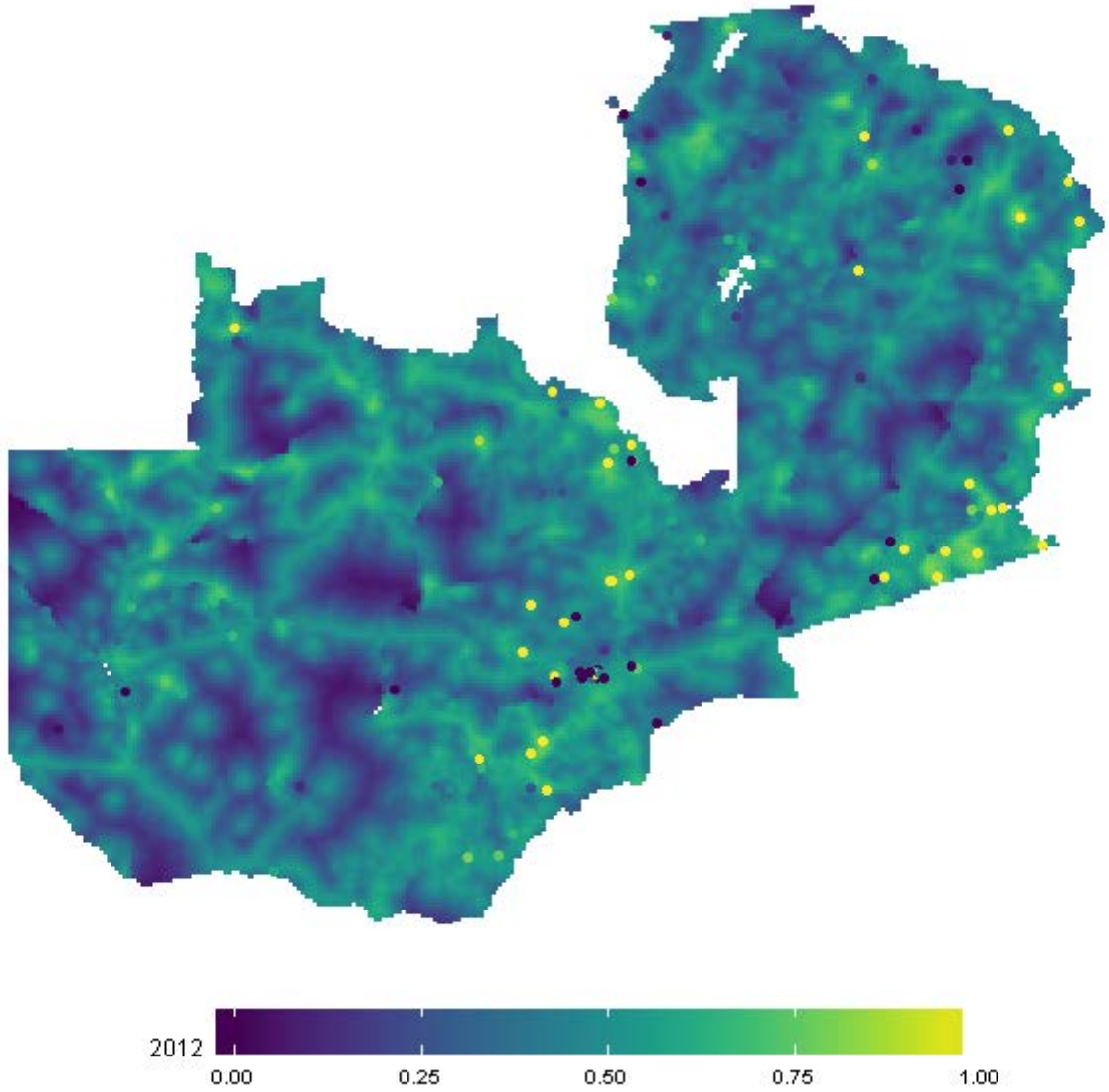
$$y \sim \text{Binomial}(p, n)$$

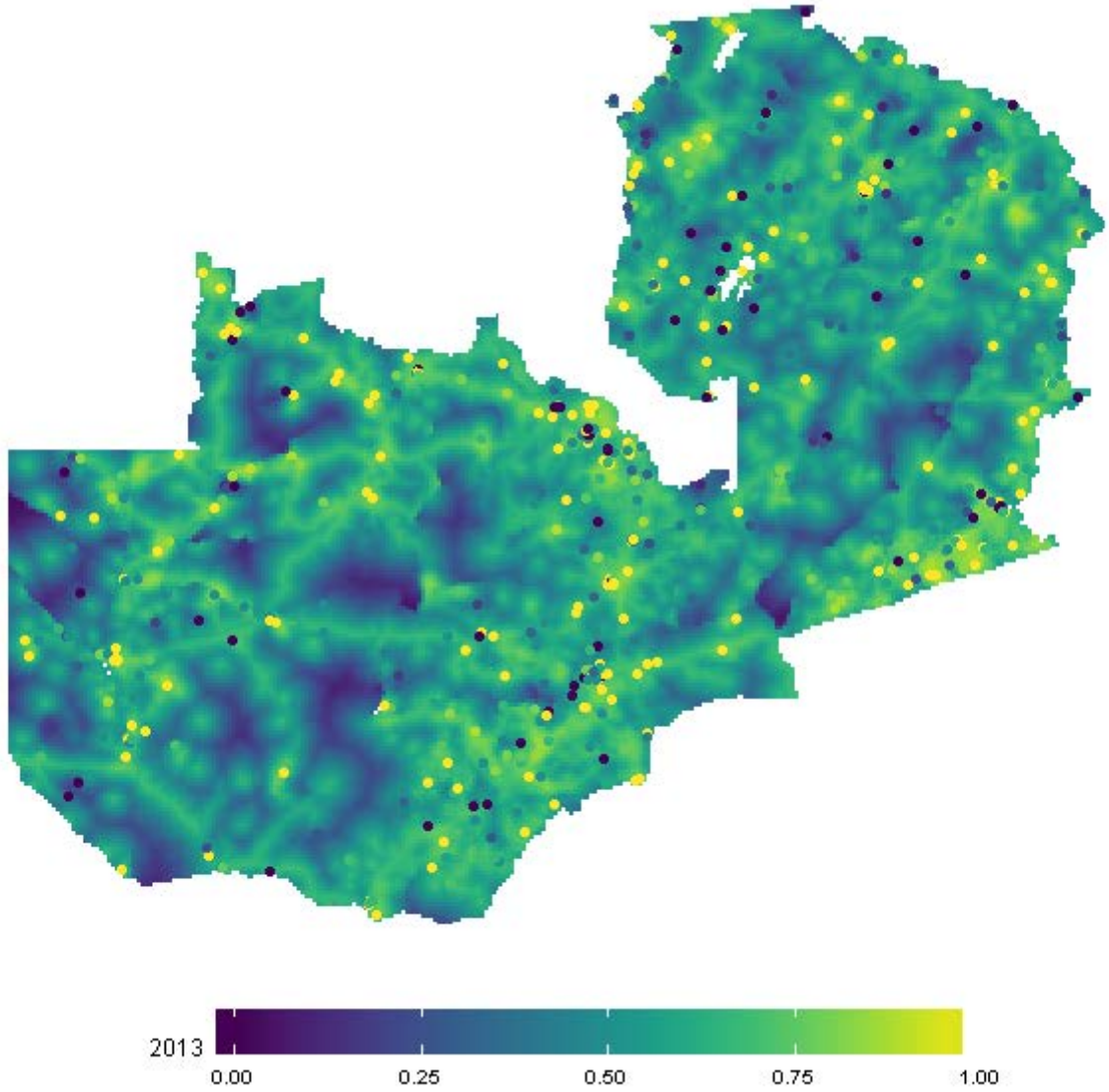
$$\text{logit}(p) = \mu + \beta d_{\boldsymbol{x}} + f_{\boldsymbol{x}} + f_t$$

- Where μ and β are constants
- $d_{\boldsymbol{x}}$ is the travel cost of a patient located in $\boldsymbol{x}$ to get to a health facility
- $f_{\boldsymbol{x}}$ and f_t are spatial and temporal random processes

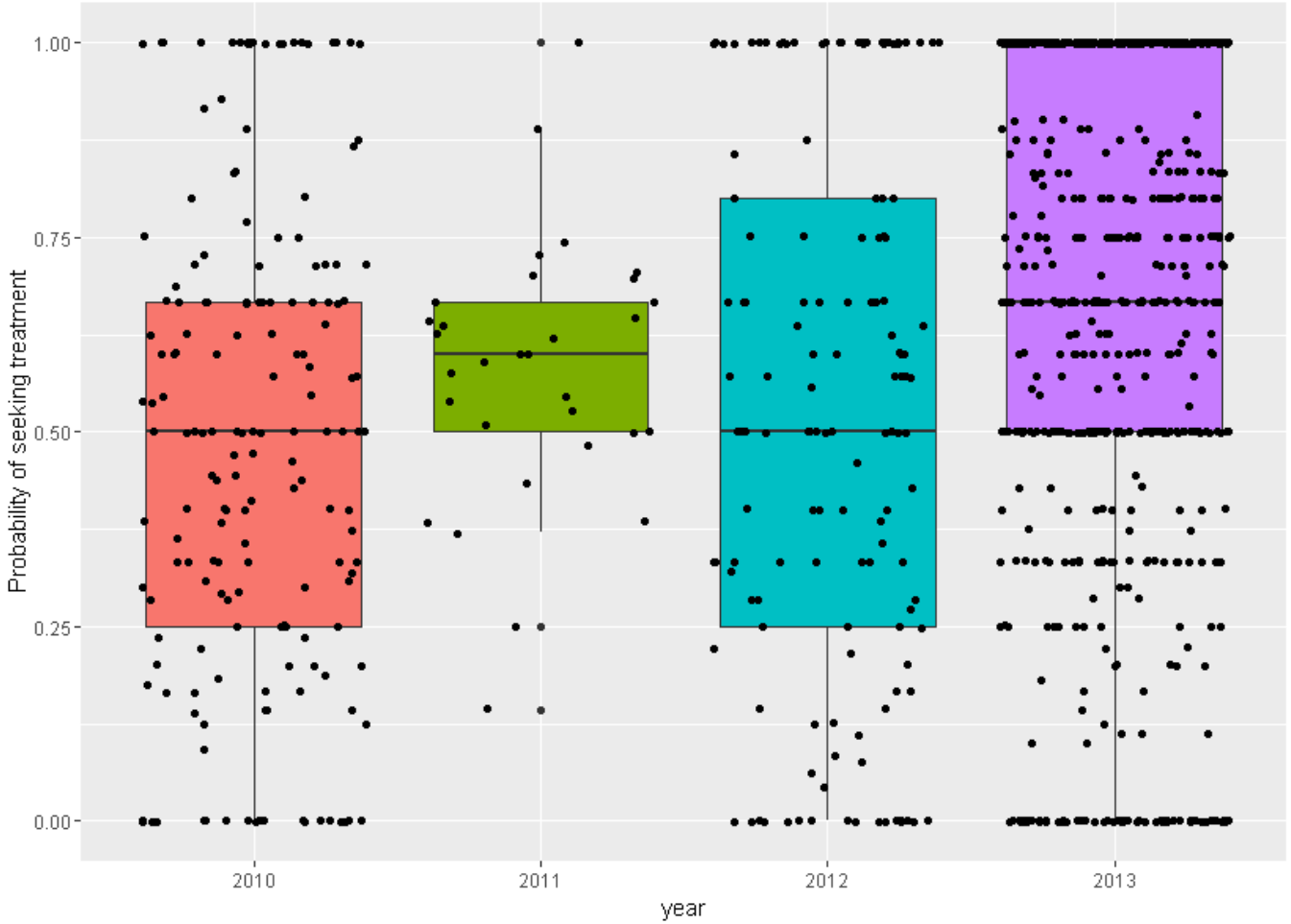

2010
0.00
0.25
0.50
0.75
1.00


2011
0.00
0.25
0.50
0.75
1.00

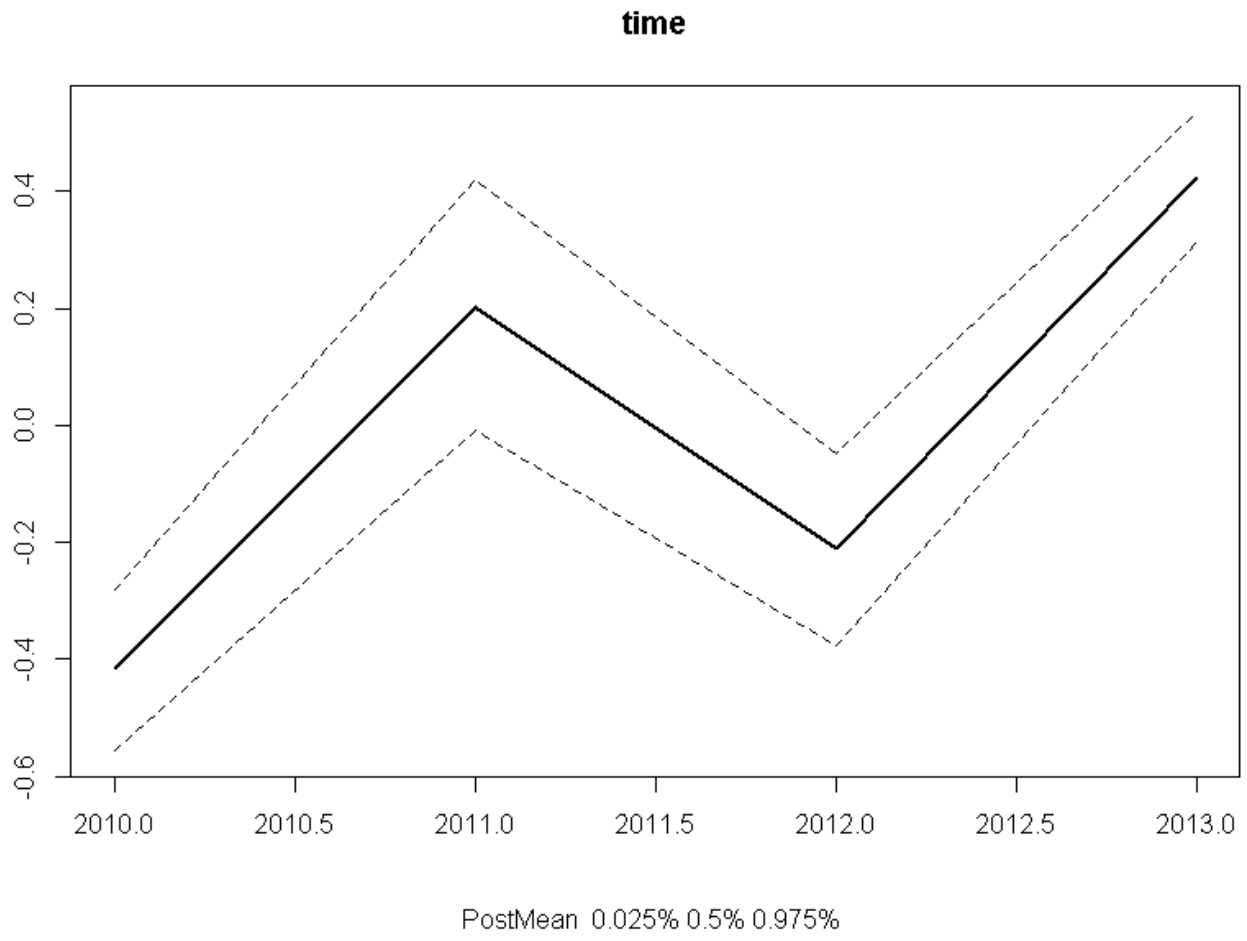

2012
0.00
0.25
0.50
0.75
1.00


2013
0.00
0.25
0.50
0.75
1.00

Observed data per year

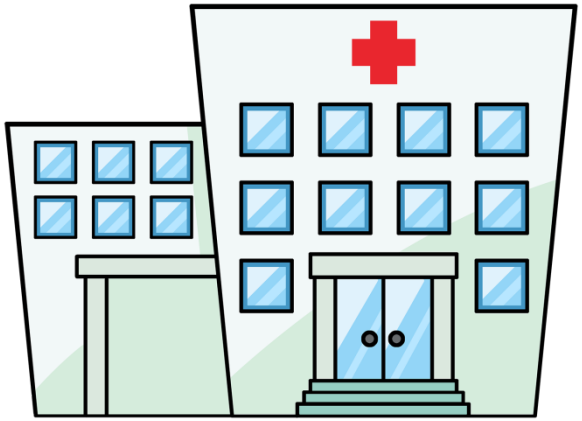


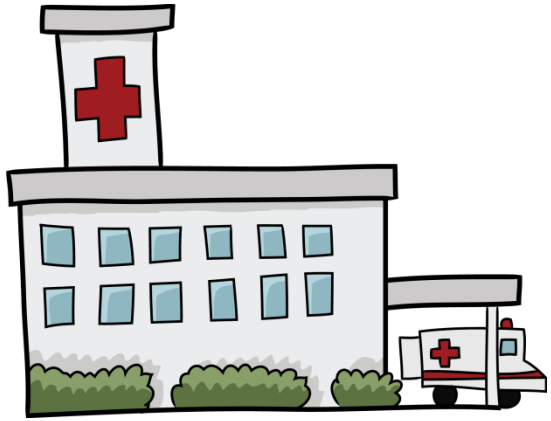
Time process



Information across years was captured by different types of surveys.
This introduces a variation that is independent from the spatial process.

Where to go?





Huff model

- Let the probability of a patient attending a health facility be given by

$$p_{ij} = \frac{u_{ij}}{\sum_j u_{ij}}$$

- Where u_{ij} is the utility level of patient i when attending facility j.
- We assume the utility is a function of the health facility attributes.

Our implementation

$$u_{ij} = e^{\tau_j} b_{ij}^{\beta} d_{ij}^{\delta}$$

- Where
  - τ is a parameter that depends on the type of health facility,
  - b_{ij} is the number of health facilities with a lower travel cost, wrt. patient i, than facility j.
  - d_{ij} is the travel cost for patient i of visiting facility j.
  - β and δ are parameters of the model.

Linear model

- Then we have that

$$p_{ij} = \frac{e^{\tau_j}\, b_{ij}^{\beta}\, d_{ij}^{\delta}}{\sum_j e^{\tau_j}\, b_{ij}^{\beta}\, d_{ij}^{\delta}}$$

- Dividing by the geometric mean $\tilde{\mu}(p_{i:})$, we have

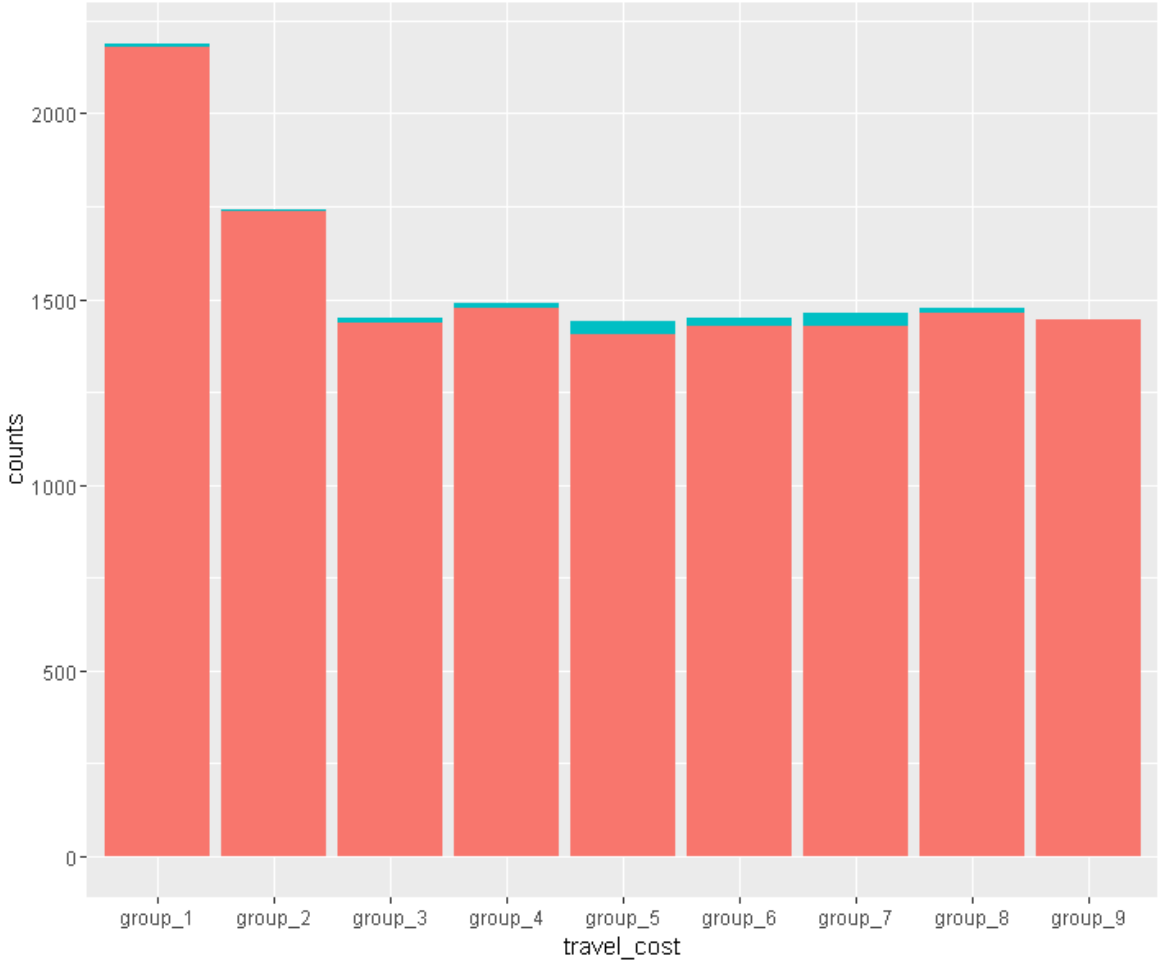
$$y_{ij} = \tau_j + \delta \log \frac{d_{ij}}{\tilde{\mu}(d_{i:})} + \beta \log \frac{b_{ij}}{\tilde{\mu}(b_{i:})}$$

- Where $y_{ij} = \log \frac{p_{ij}}{\tilde{\mu}(p_{i:})}$

How to sample?

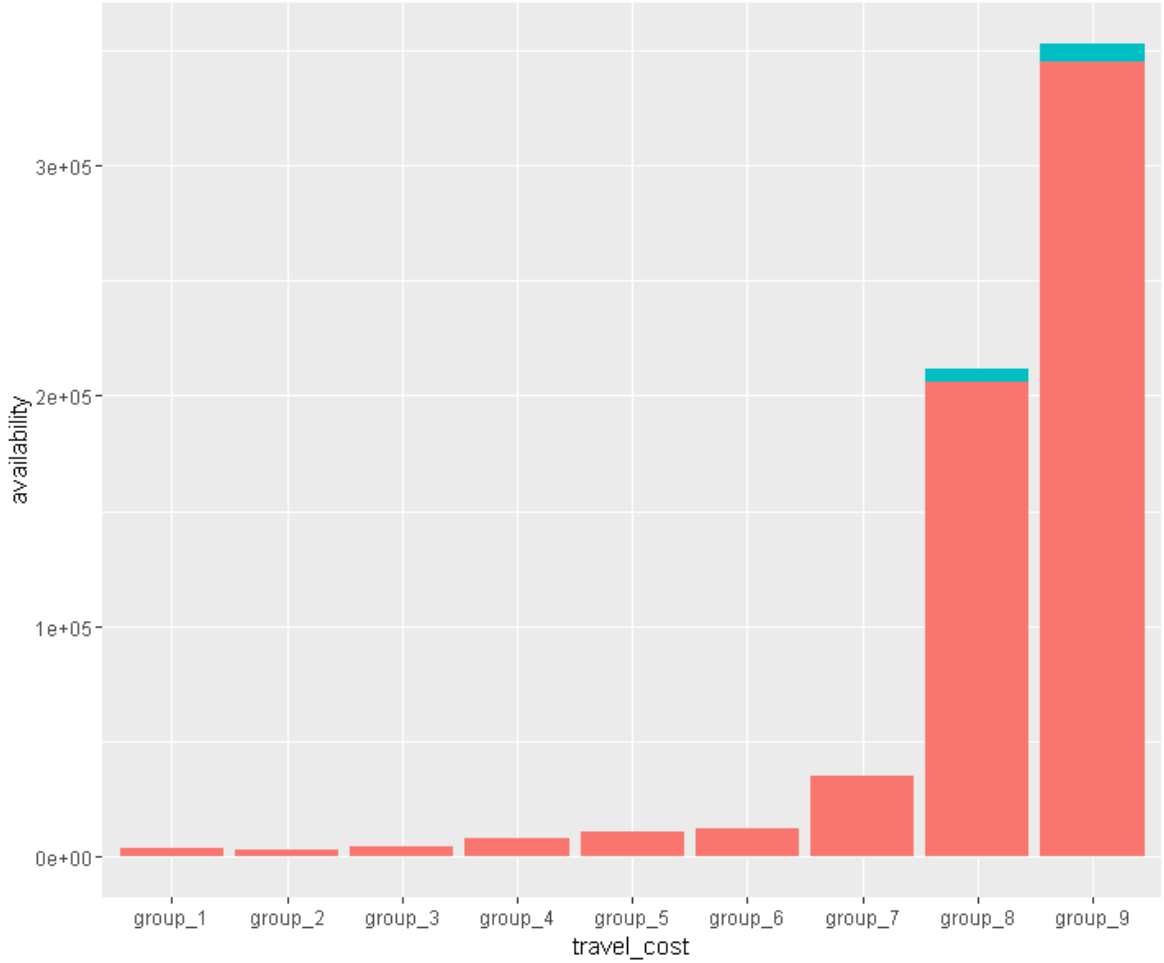
- We do not have a set of preferences p_{ij} for each patient.
- Just the one facility they visited.
- Assume every patient chooses a health facility in a similar way.
- Then the set of facilities attended are a realization of the same decision process.

Number of facilities attended by travel cost and type



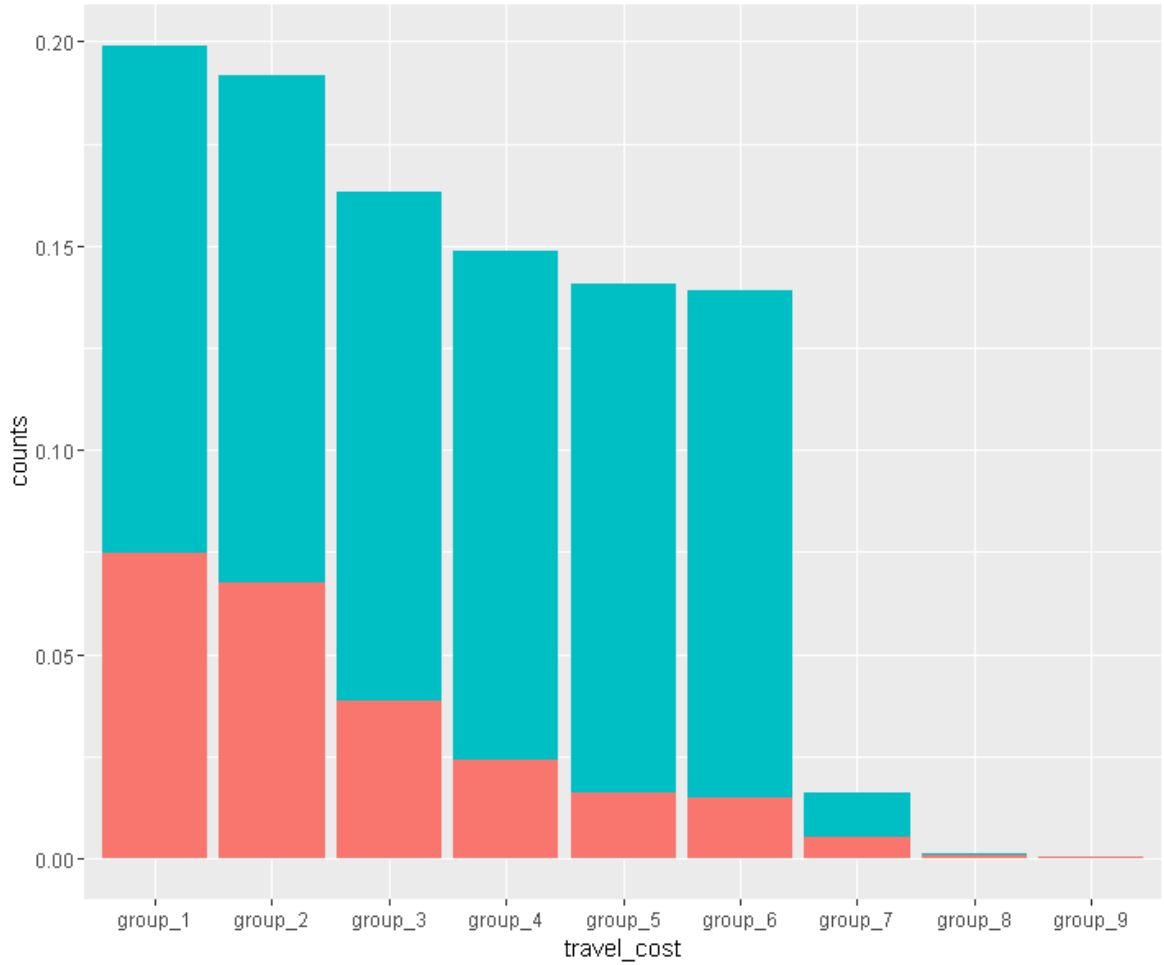
The travel costs between each patient's household and the facility they attended

Availability of facilities by travel cost and type



The travel costs patient's household and all the health facilities in the survey

Weighted frequency by travel cost and type



The number of facilities attended by group/type was scaled by the number of facilities available by group/type and divided so that they add up to 1.

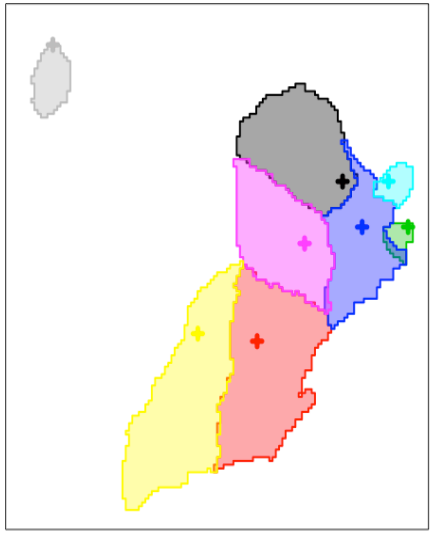
Now we can learn the model parameters

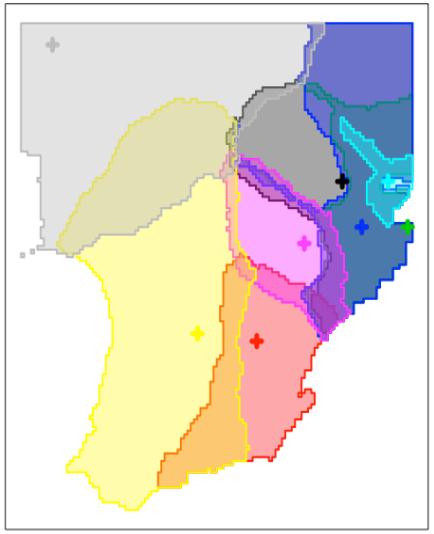
```
Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
factor(huff.data$X.type)C -0.009053   0.001028   -8.806   <2e-16 ***
factor(huff.data$X.type)P  0.464875   0.006741   68.960   <2e-16 ***
huff.data$X.cost          -1.597340   0.003730 -428.214   <2e-16 ***
huff.data$X.numb          -0.521516   0.002982 -174.862   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8142 on 641918 degrees of freedom
Multiple R-squared:  0.7791,    Adjusted R-squared:  0.7791
F-statistic: 5.659e+05 on 4 and 641918 DF,  p-value: < 2.2e-16
```

Catchment areas

- For every point in the country we can now estimate the preferences towards the surrounding health facilities.
- These preferences define catchment areas with soft boundaries.





Catchment areas estimated for 8 facilities in Zambia.
Left: catchment areas defined by points with a probability >. 1 of attending that facility.
Right: catchment areas defined by points with a probability of > .05 of attending the facility

How is this linked to the incidence?

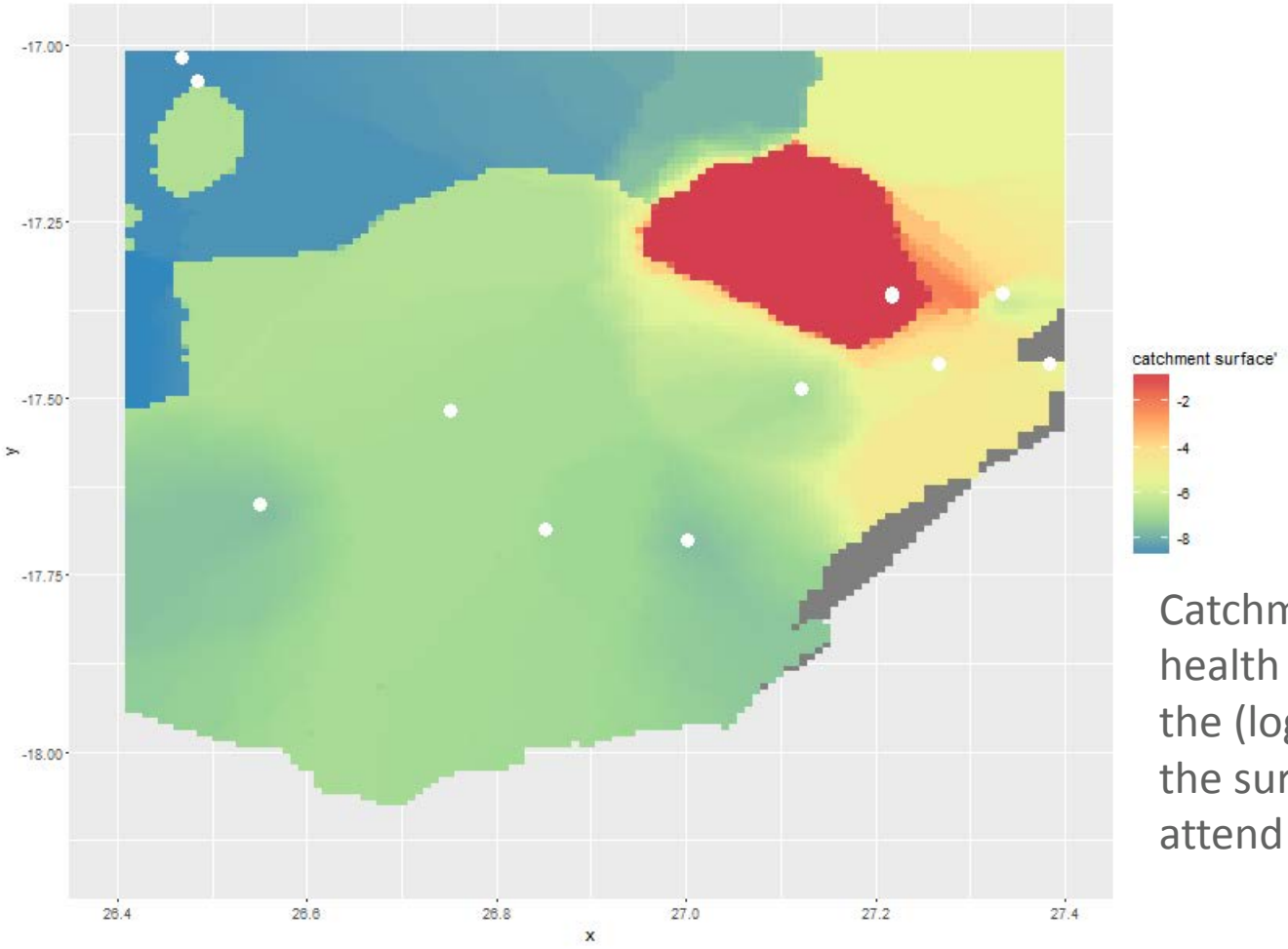
- Let the number of patients received by health facility j be

$$n_j = \text{Poisson}(\lambda_j(\boldsymbol{x}))$$

- Where $\lambda_j(\boldsymbol{x}) \propto f_{\text{Pop}}(\boldsymbol{x}) \times f_{\text{Seek}}(\boldsymbol{x}) \times f_{\text{Catch}}(\boldsymbol{x})$
- Since we know the number of patients $\dot{n}_j$ that visited each facility, we can set the constraint

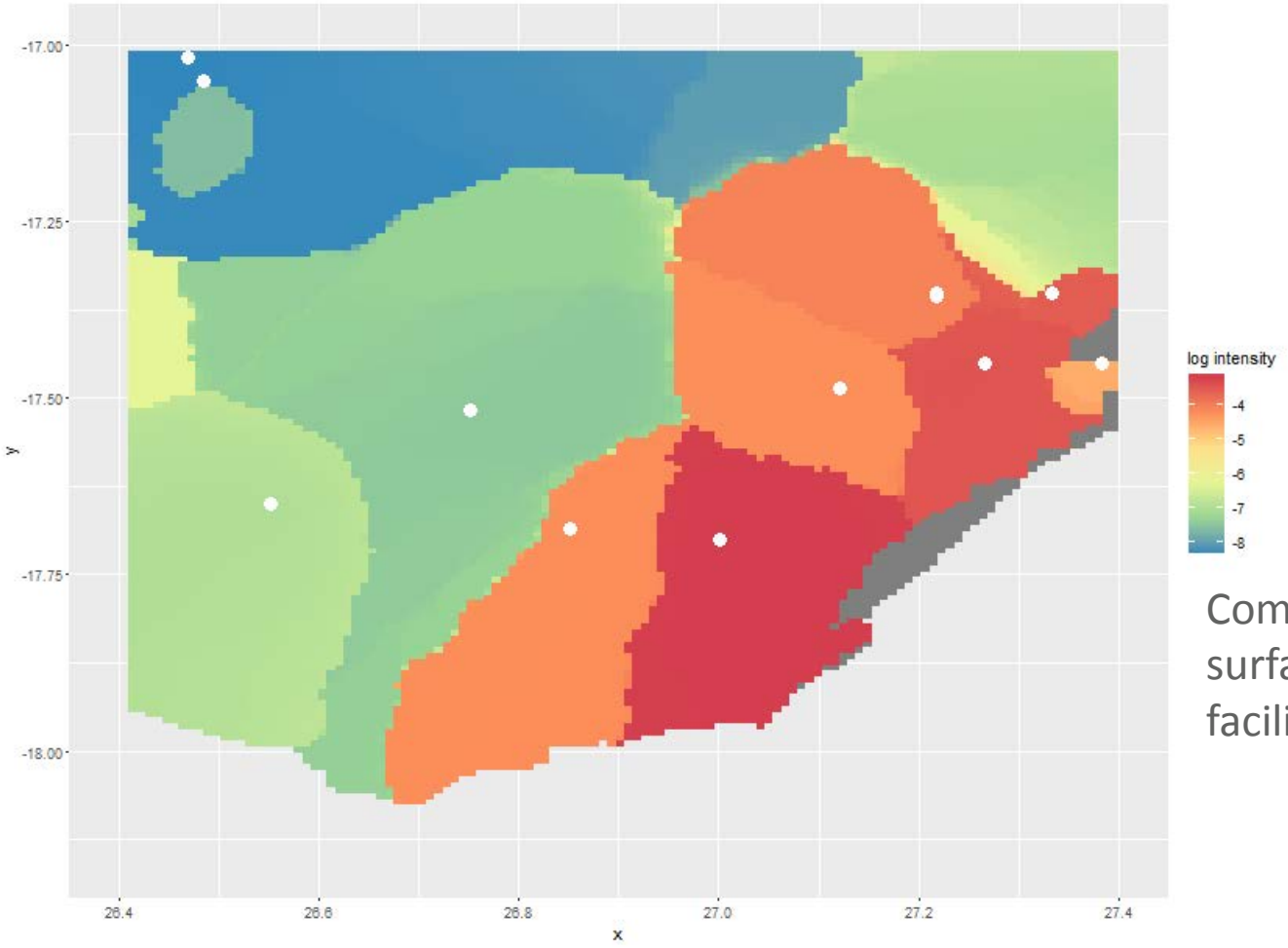
$$\dot{n}_j = \int \lambda_j(\boldsymbol{x}) d\boldsymbol{x}$$

Preliminary results



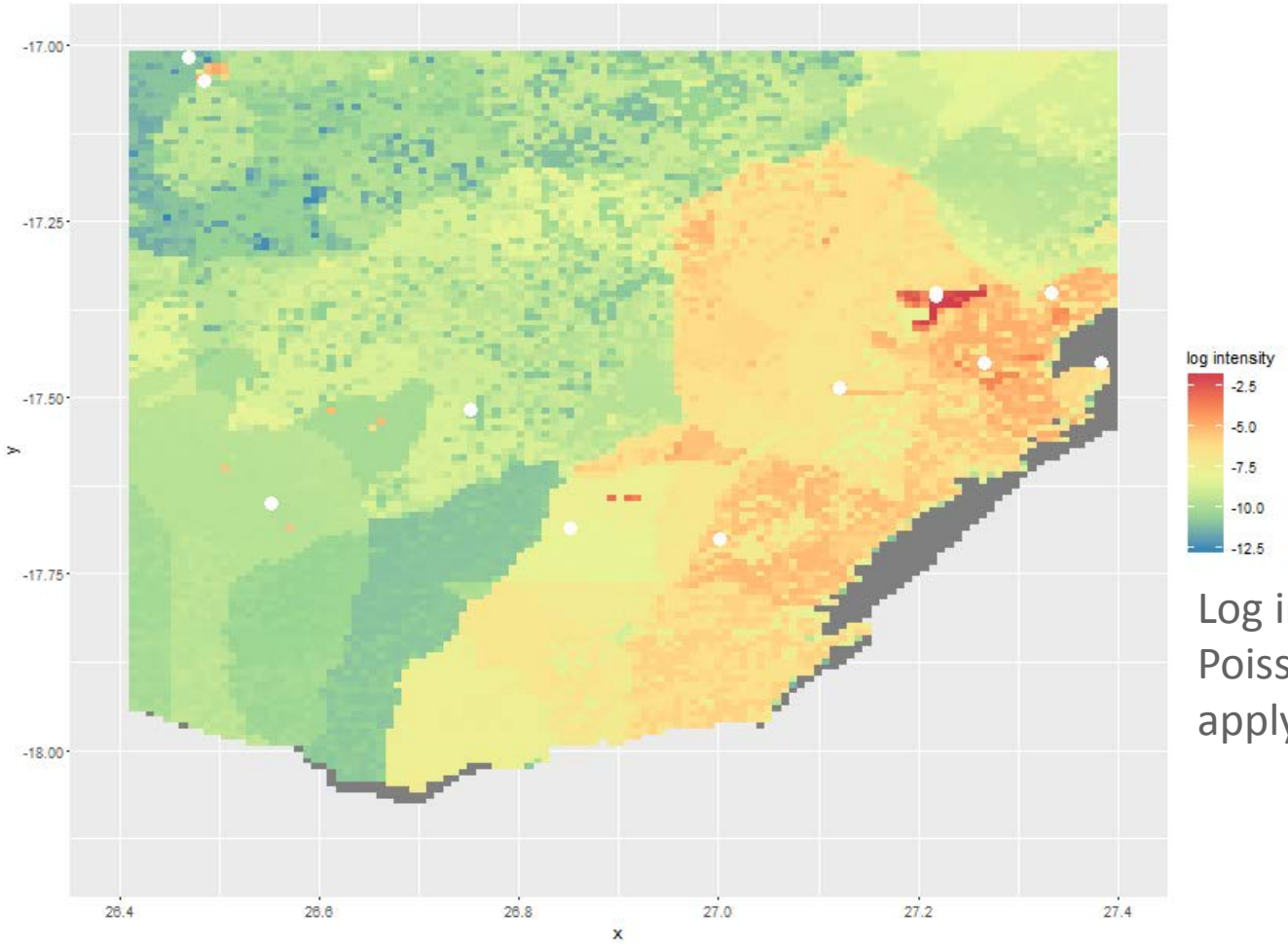
Catchment surface of a health facility defined by the (log) probability that the surrounding points attend it.

Preliminary results



Combined catchment surfaces of different facilities

Preliminary results



Log intensity of the Poisson process after applying the constraint

$$\dot{n}_j = \int \lambda_j(\boldsymbol{x}) d\boldsymbol{x}$$

Acknowledgments

- Adam Bennett, MEI, Global Health Group, UCSF , San Francisco, CA, United States
- Busiku Hamainza, Zambia National Malaria Control Centre, Lusaka, Zambia
- Daniel J. Weiss, University of Oxford, Oxford, United Kingdom
- Hugh Sturrock, MEI, Global Health Group, UCSF , San Francisco, CA, United States
- John Miller, Malaria Control and Elimination Partnership in Africa, Lusaka, Zambia
- Kafula Silumbe, Malaria Control and Elimination Partnership in Africa, Lusaka, Zambia
- Pete Gething, University of Oxford, Oxford, United Kingdom
- Samir Bhatt, University of Oxford, Oxford, United Kingdom
- Thomas P. Eisele, Tulane University, New Orleans, LA, United States

References

- Abrahamsen, P. (1997). *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Gandin, L. S. (1963). *Ob“ektivnyi analiz meteorologicheskikh polei*. Gidrometeologicheskoe Izdatel’stvo, Leningrad. Translation (1965): *Objective analysis of meteorological fields*. Israel Program for Scientific Translations, Jerusalem.
- Huff, D. L. (1964). Defining and estimating a trading area. *The Journal of Marketing*, 34-38.

References

- Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.
- Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.
- Matheron, G. (1962). Traité de géostatistique appliquée, tome I. *Mémoires du Bureau de Recherche Géologiques et Minières*, No. 14. Editions Technip, Paris.
- Matheron, G. (1963). Traité de géostatistique appliquée, tome II: le krigeage. *Mémoires du Bureau de Recherche Géologiques et Minières*, No. 24. Editions Bureau de Recherche Géologiques et Minières, Paris.
- Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and Kriging. *Catena*, *113*, 56-69.