School of Mathematics
FACULTY OF MATHEMATICS AND PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Modelling discontinuities in simulator output using Voronoi tessellations

John Paul Gosling (University of Leeds)

and

Chris Pope, Jill Johnson and Stuart Barber (University of Leeds)

Paul Blackwell (University of Sheffield)

Overview

UNIVERSITY OF LEEDS

1. Why bother with discontinuities?
2. Attempts to split the space of interest.
3. Sampling to find discontinuities.
4. Application in climate science.

DISCLAIMER: This is (still) work-in-progress: there are many aspects that we need to sort out and improve.

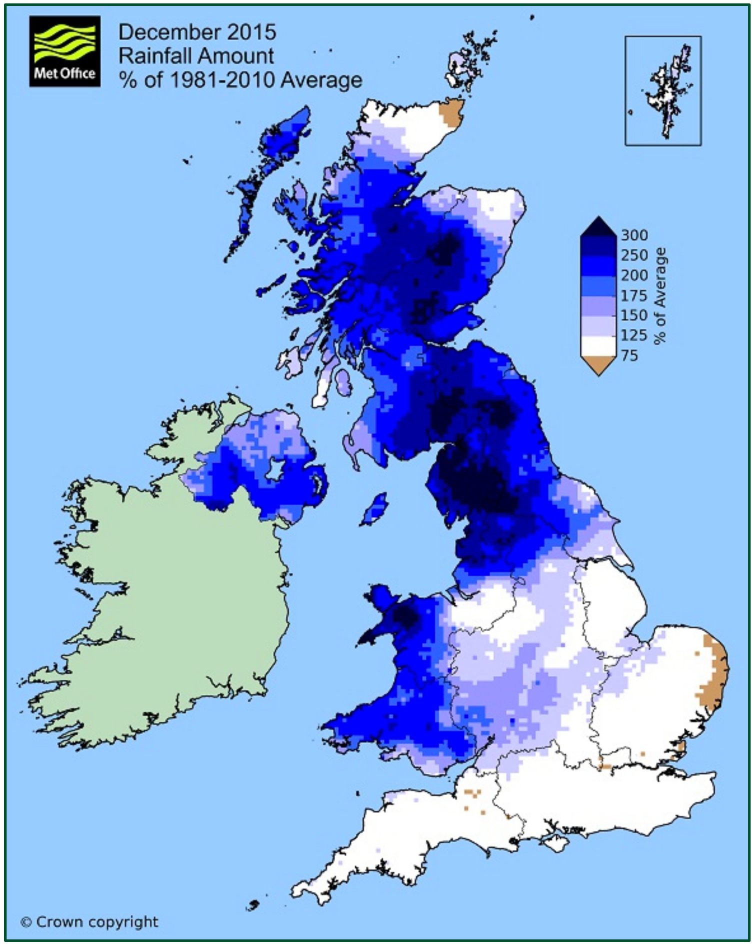
Motivation

UNIVERSITY OF LEEDS

Heterogeneity can occur in spatial processes.

Discontinuities can create challenges for modelling.

Transformations can only get us so far.



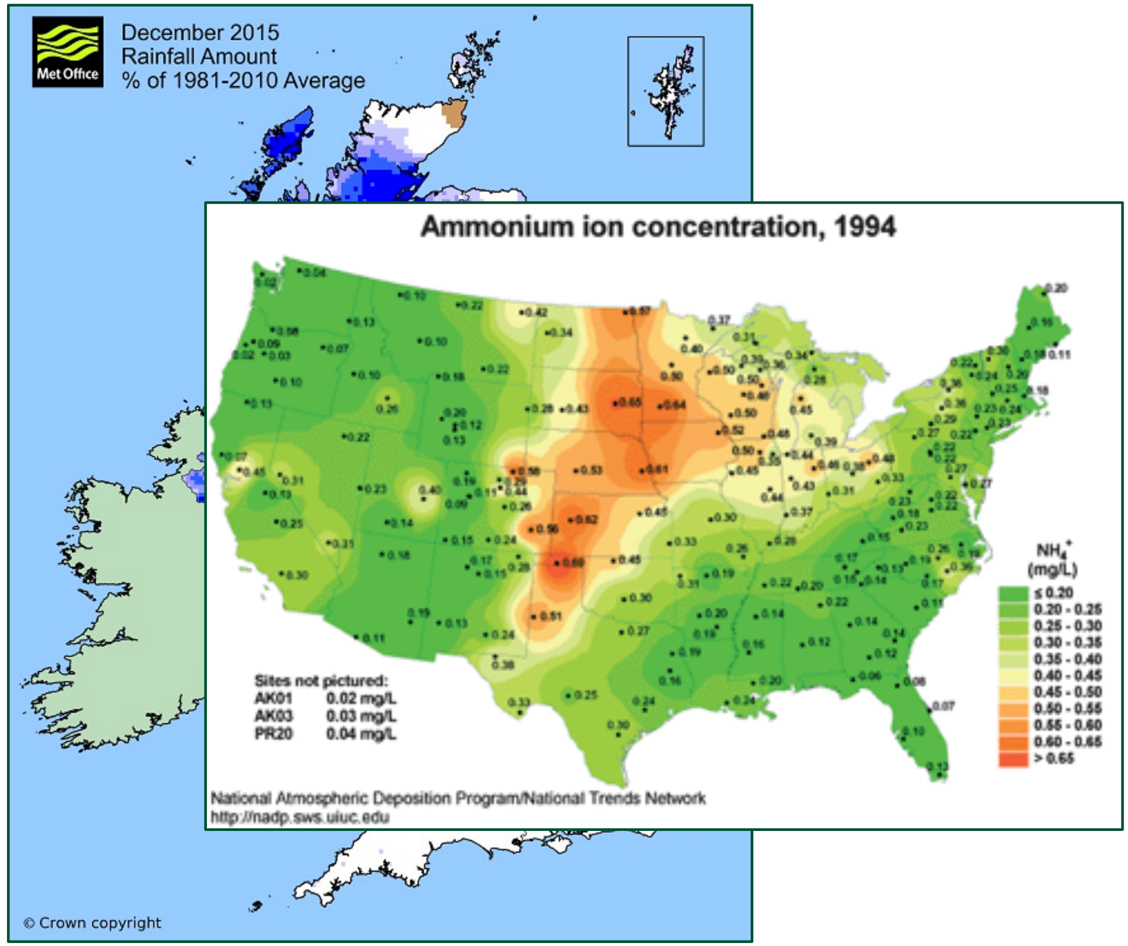
Motivation

UNIVERSITY OF LEEDS

Heterogeneity can occur in spatial processes.

Discontinuities can create challenges for modelling.

Transformations can only get us so far.



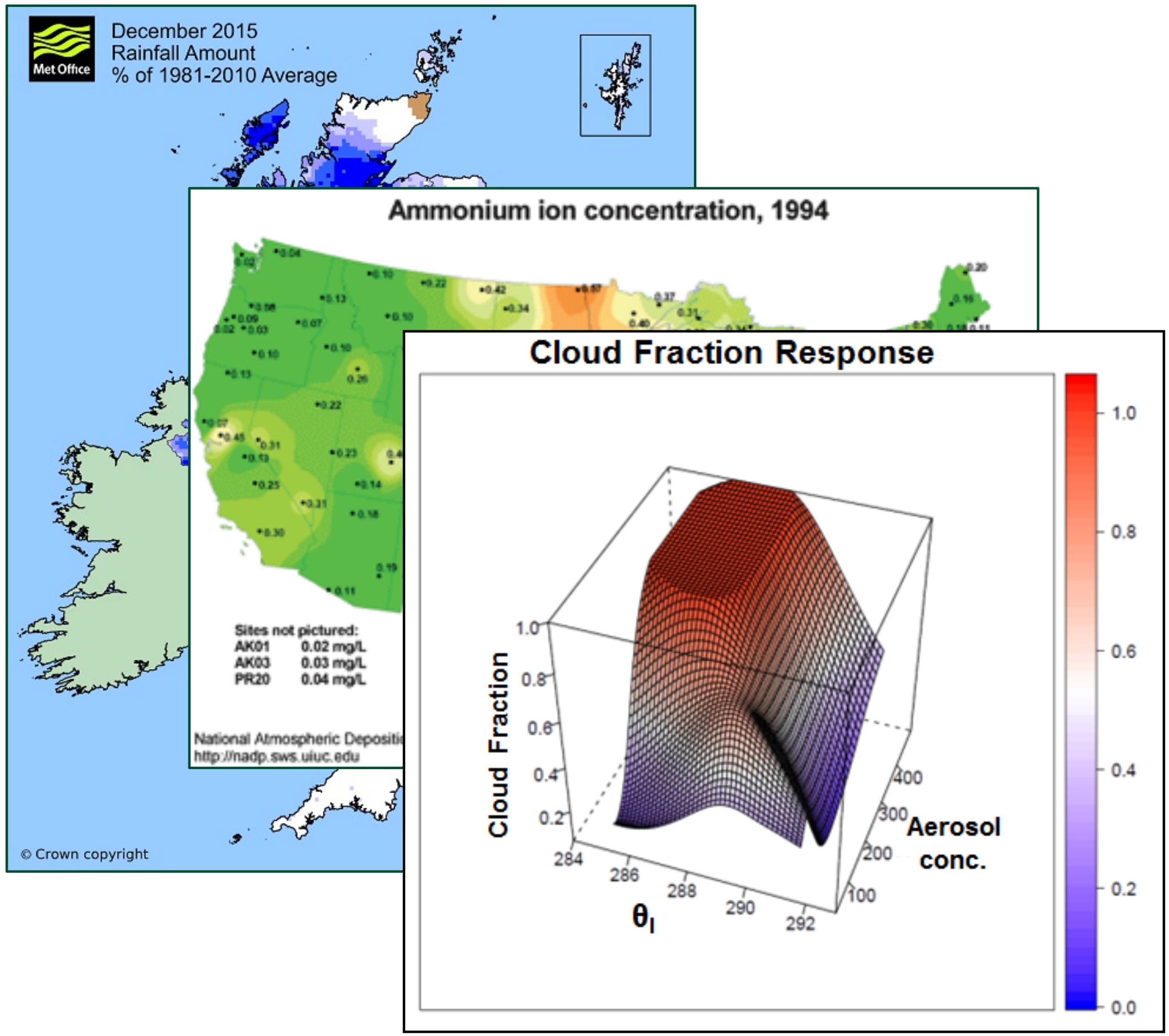
Motivation

UNIVERSITY OF LEEDS

Heterogeneity can occur in spatial processes.

Discontinuities can create challenges for modelling.

Transformations can only get us so far.



GP emulation (or kriging)

UNIVERSITY OF LEEDS

Our regression building block for this talk is a Gaussian process regression model:

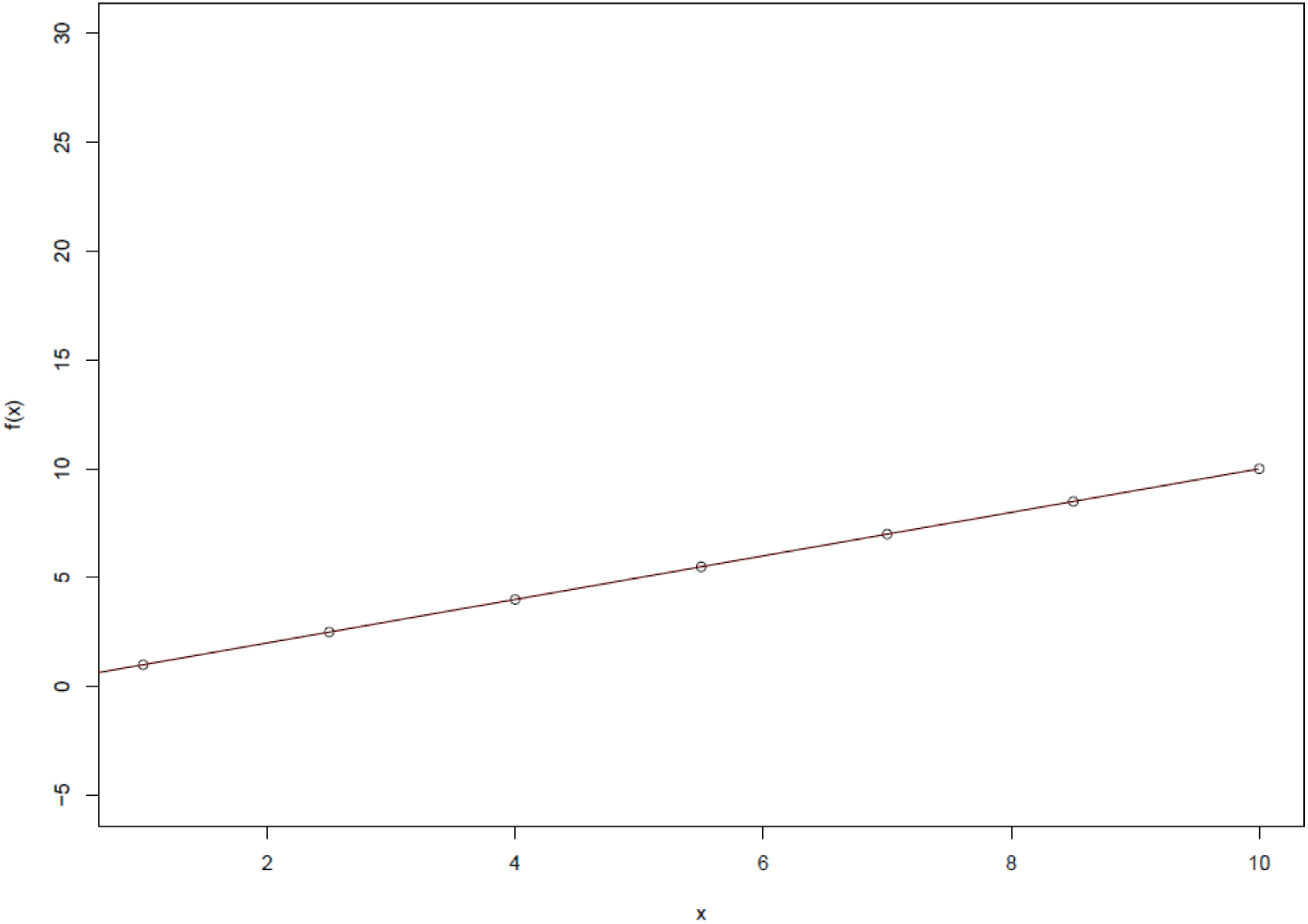
$$f(.) \sim GP\left(m(.), \sigma^2 c(.,.)\right).$$

We observe $f(.)$ at a limited number of points, and we can update this prior.

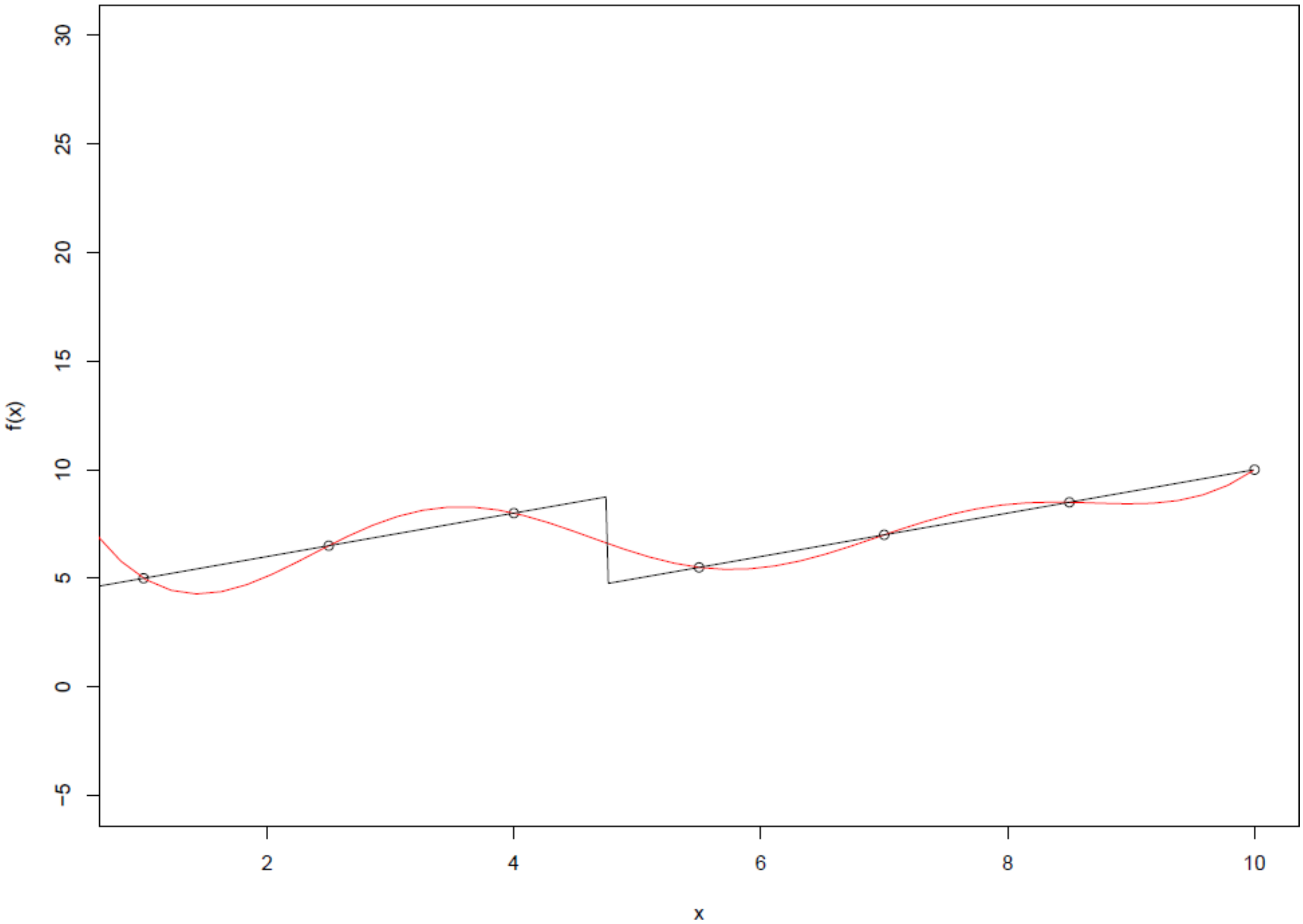
We have used both Gaussian and Matérn correlation functions.

1d example

UNIVERSITY OF LEEDS

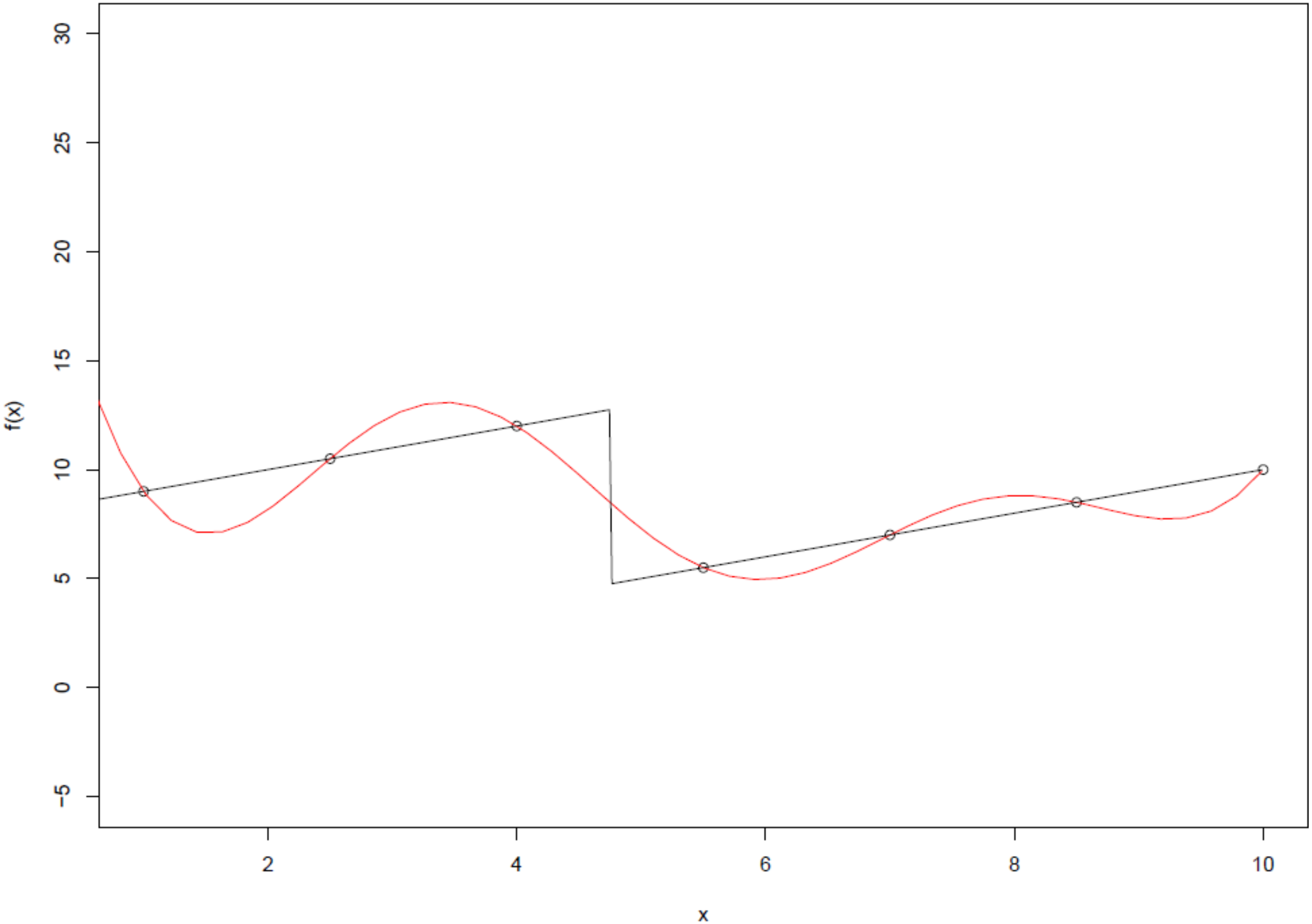


1d example



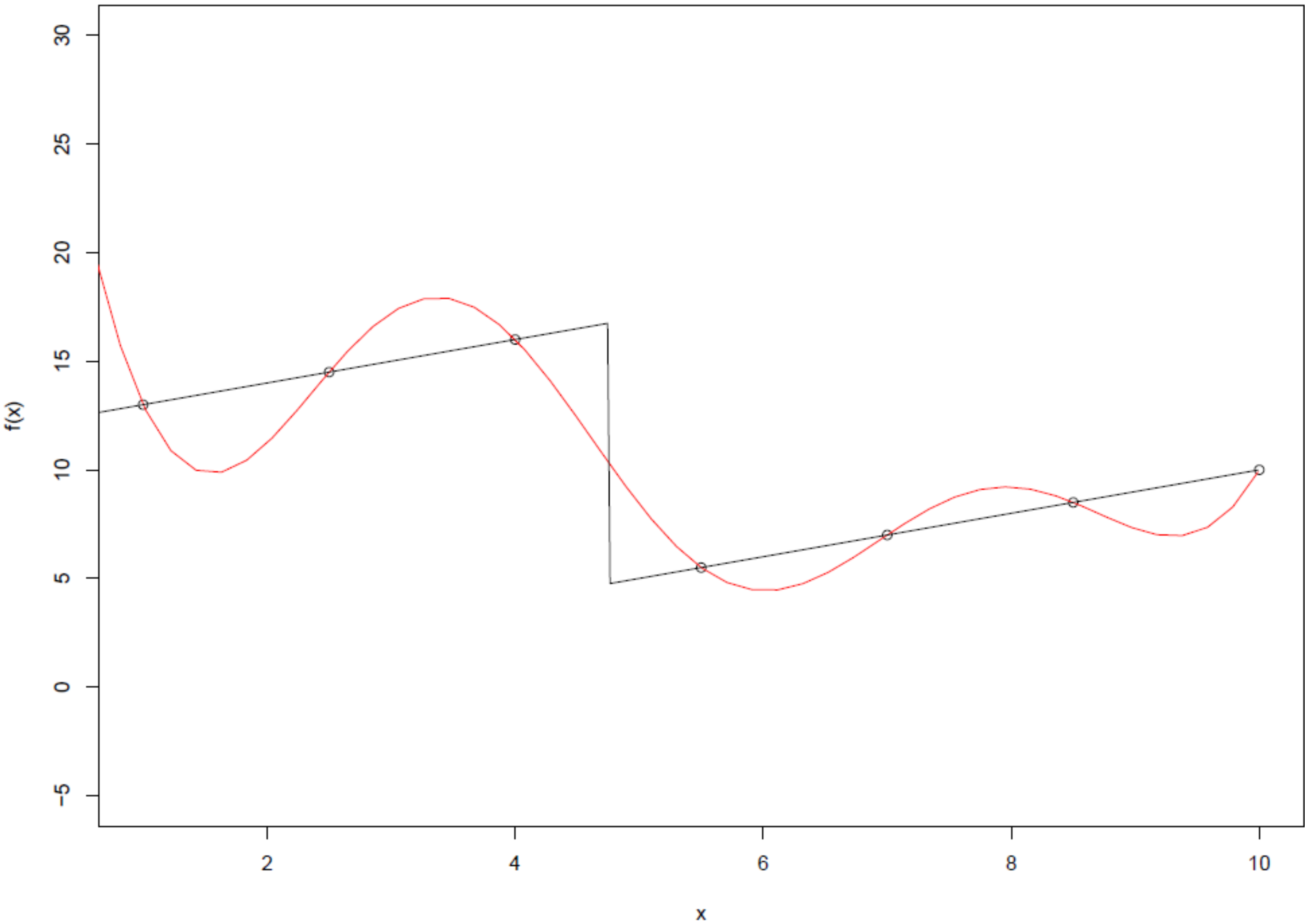
1d example

UNIVERSITY OF LEEDS



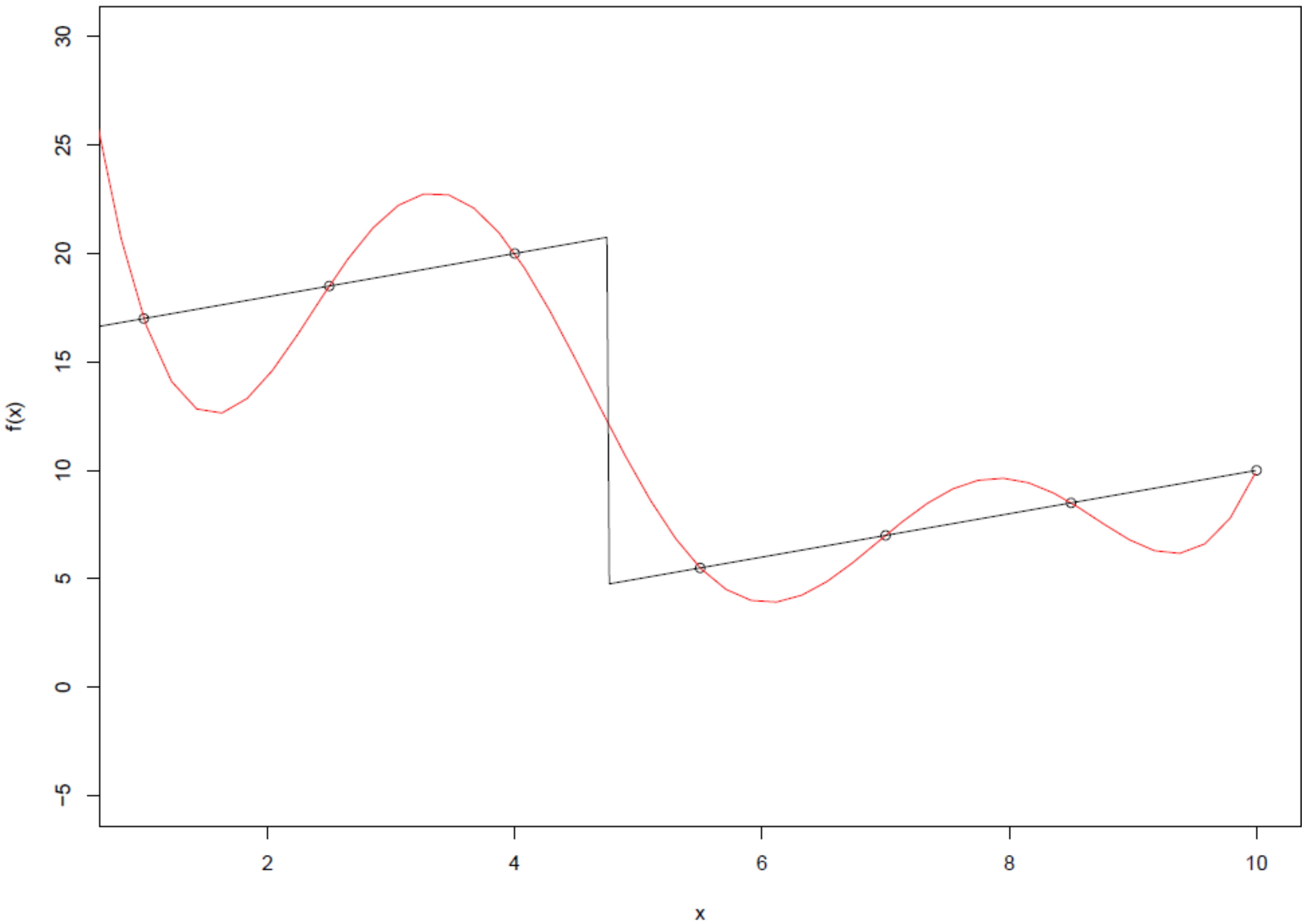
1d example

UNIVERSITY OF LEEDS



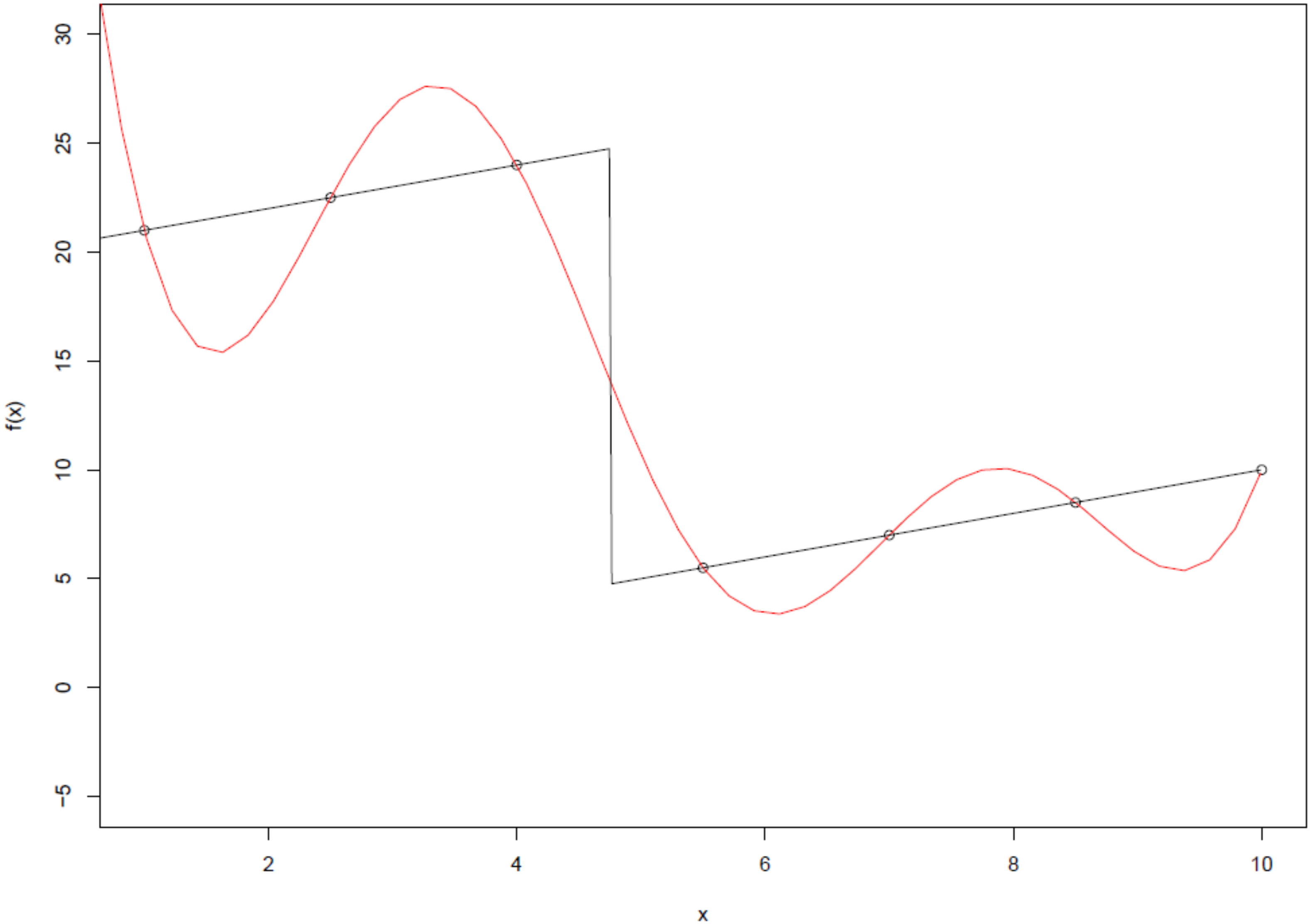
1d example

UNIVERSITY OF LEEDS



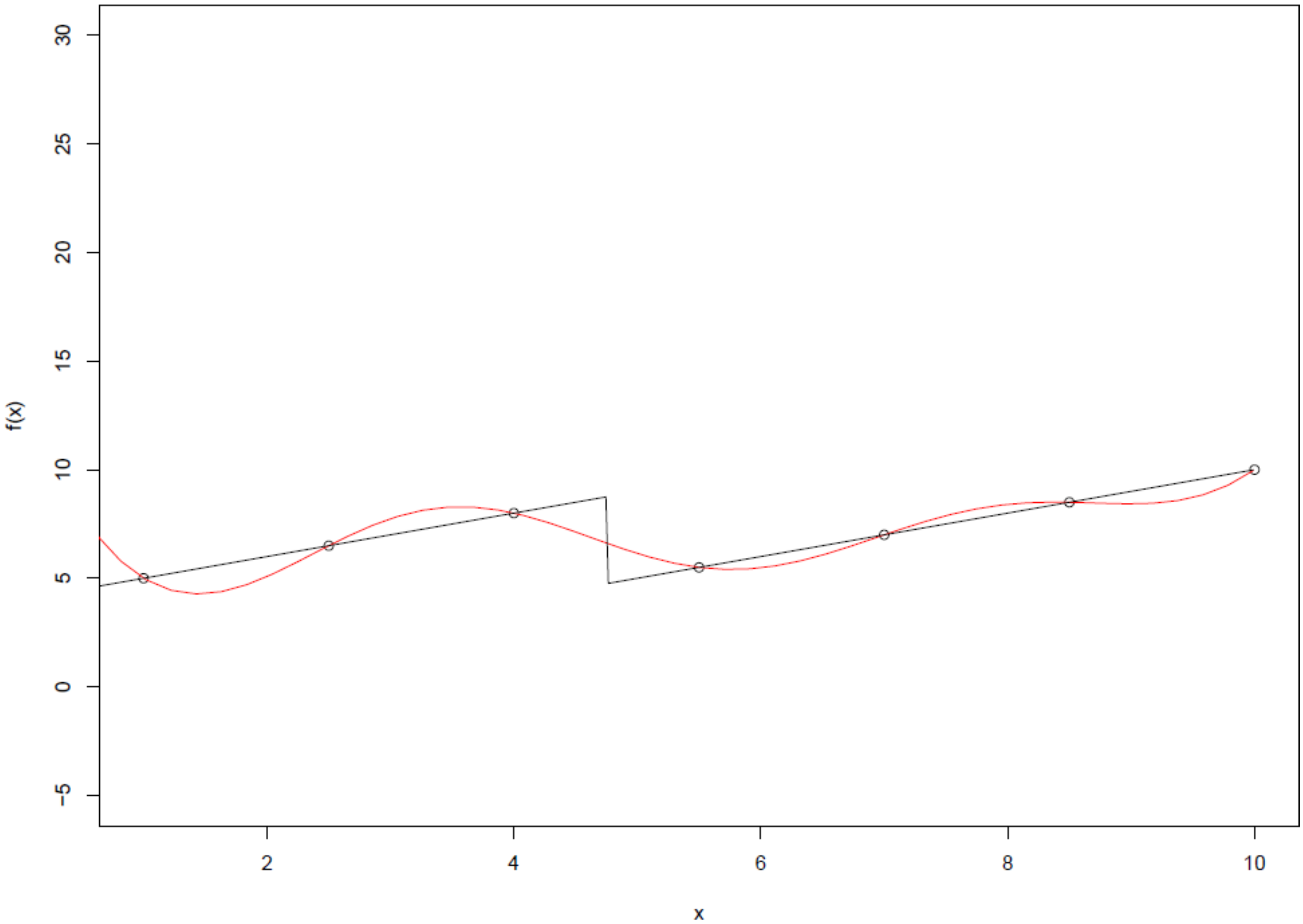
1d example

UNIVERSITY OF LEEDS



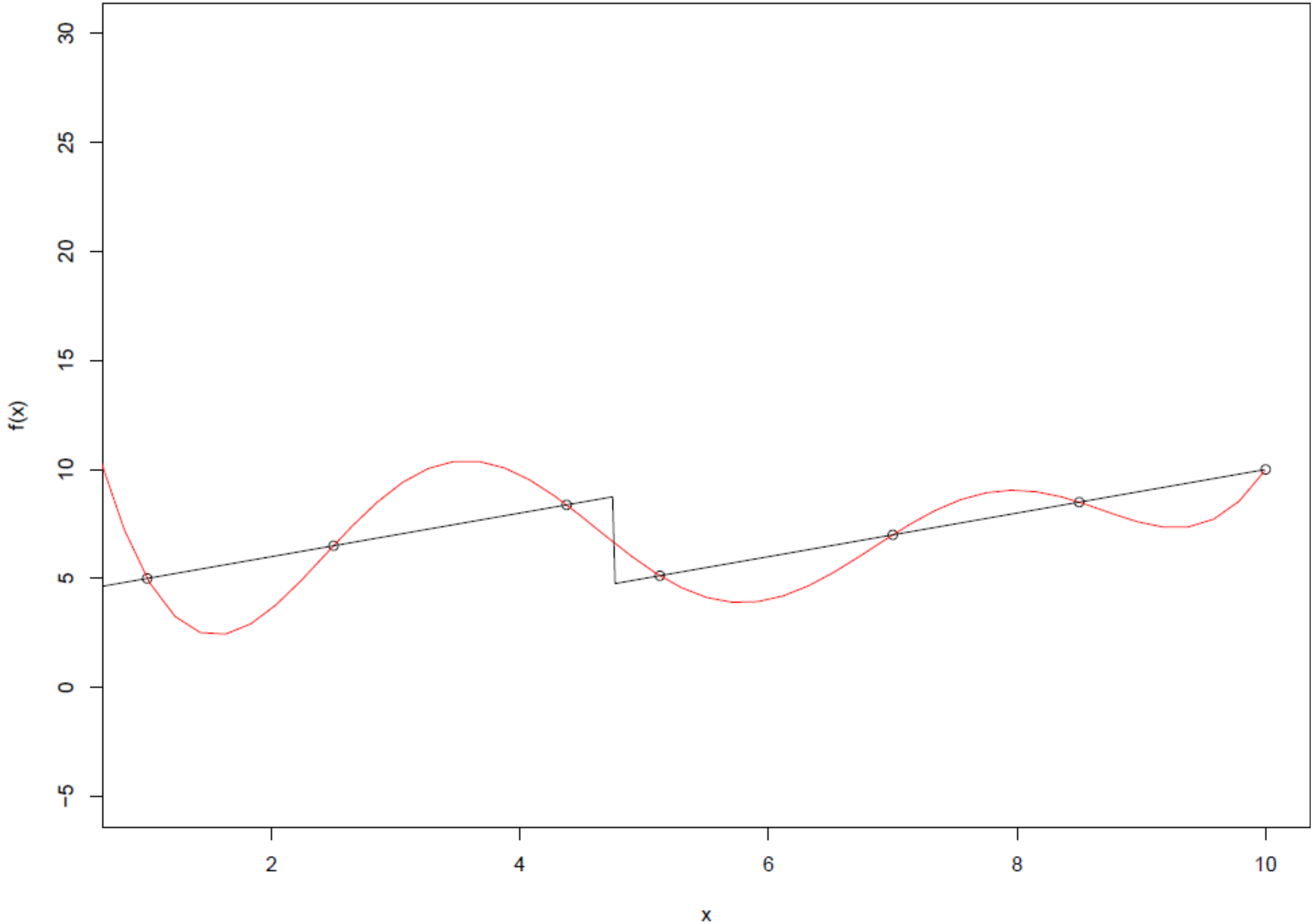
1d example

UNIVERSITY OF LEEDS



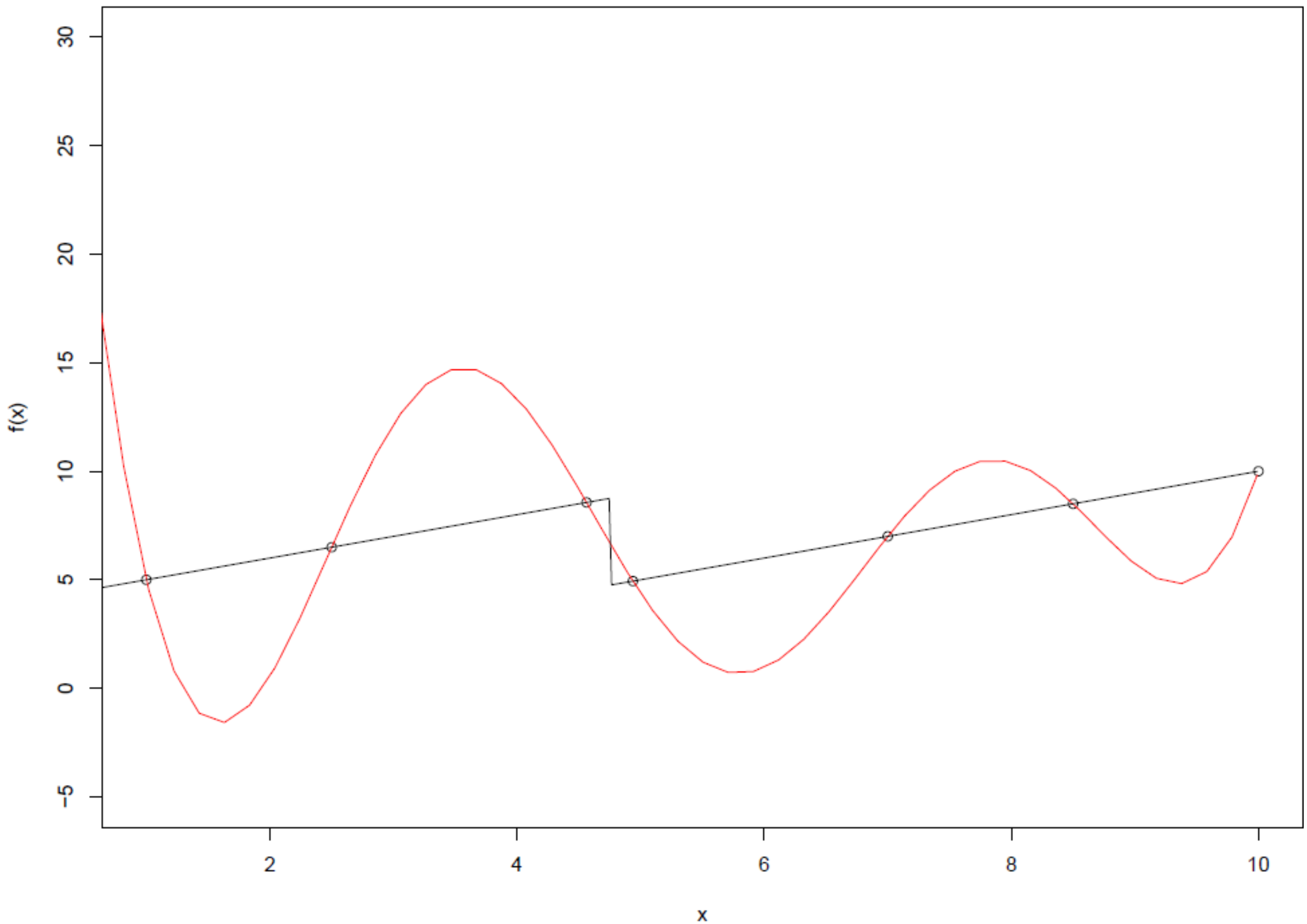
1d example

UNIVERSITY OF LEEDS



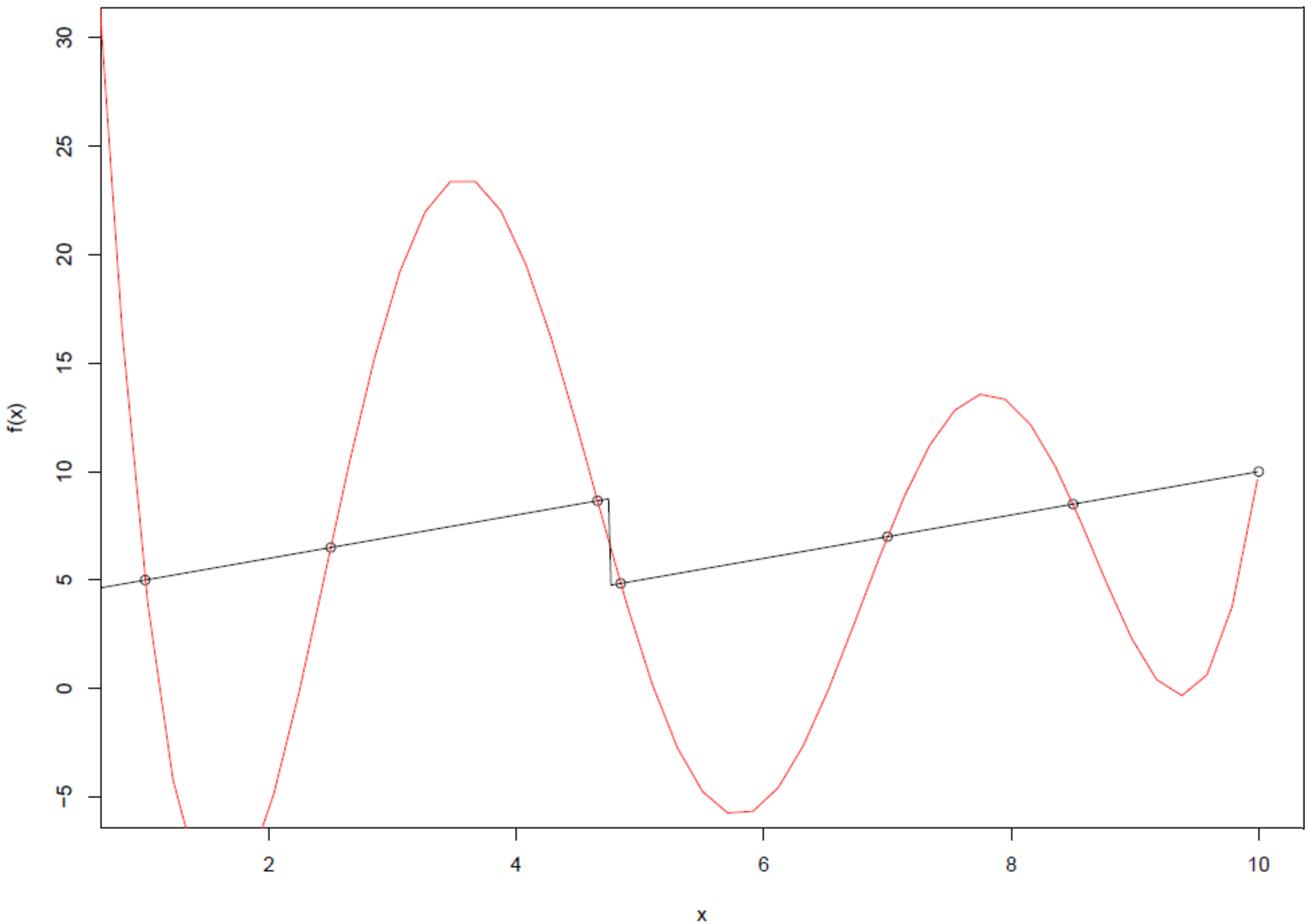
1d example

UNIVERSITY OF LEEDS



1d example

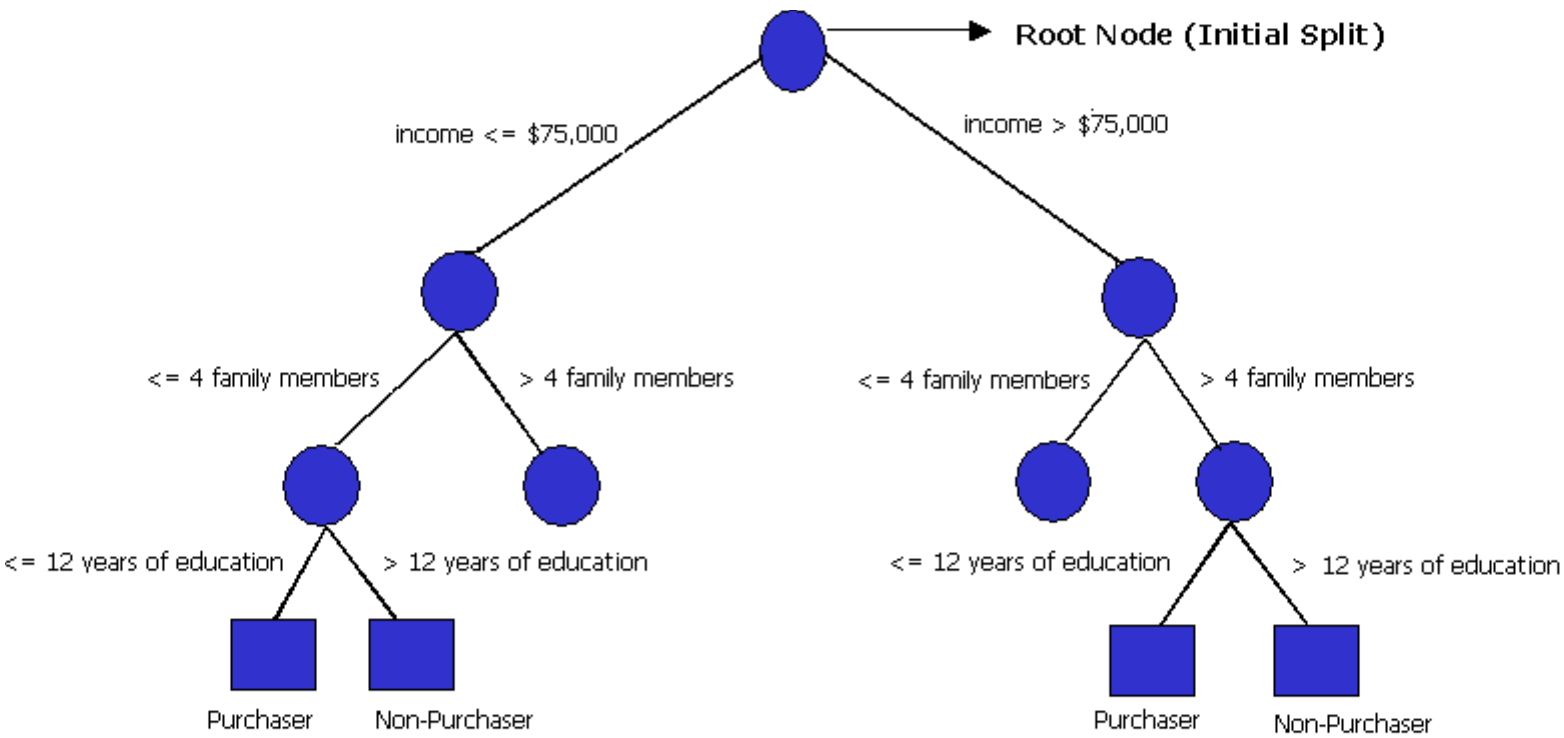
UNIVERSITY OF LEEDS



Classification trees

UNIVERSITY OF LEEDS

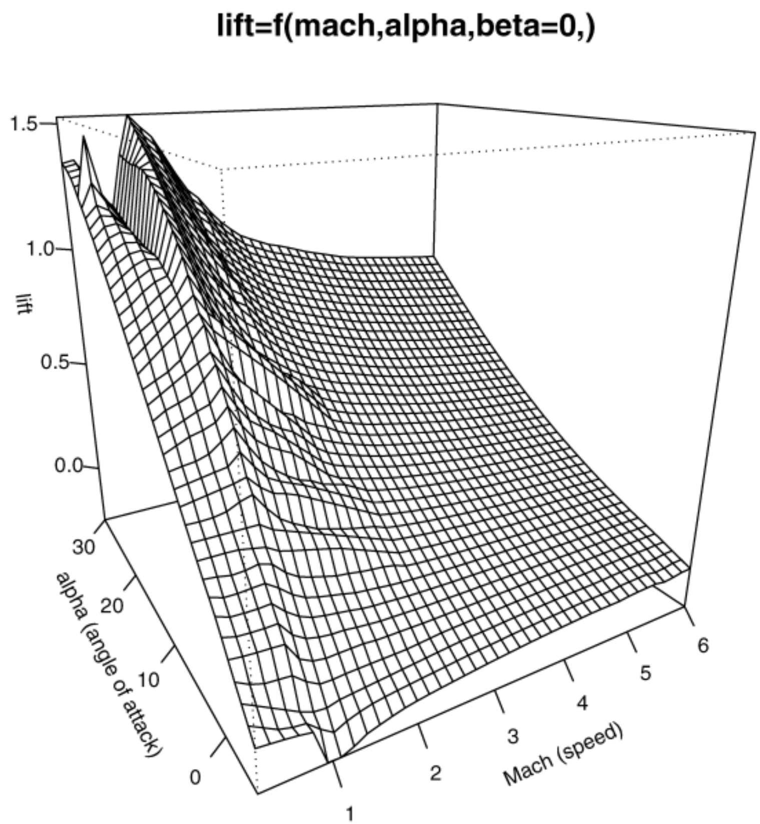
Classification trees are learning analogues of decision trees.



Classification trees

UNIVERSITY OF LEEDS

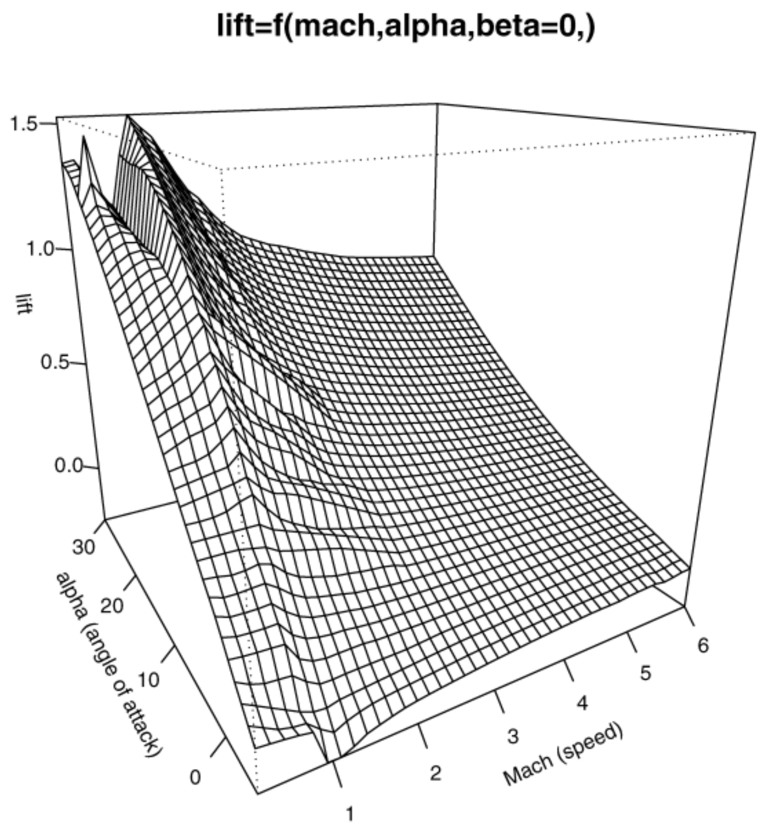
Treed Gaussian processes were designed to split space into heterogeneous areas.

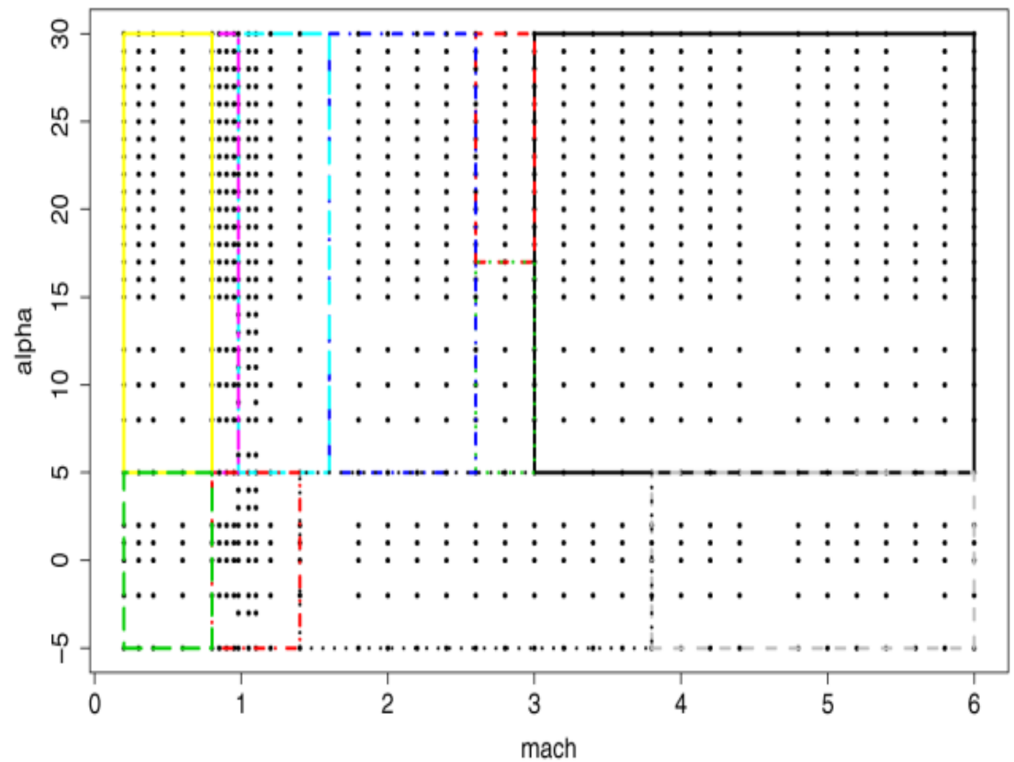


Nice R implementation: `tgp` package.

Classification trees

Treed Gaussian processes were designed to split the space into heterogeneous areas.



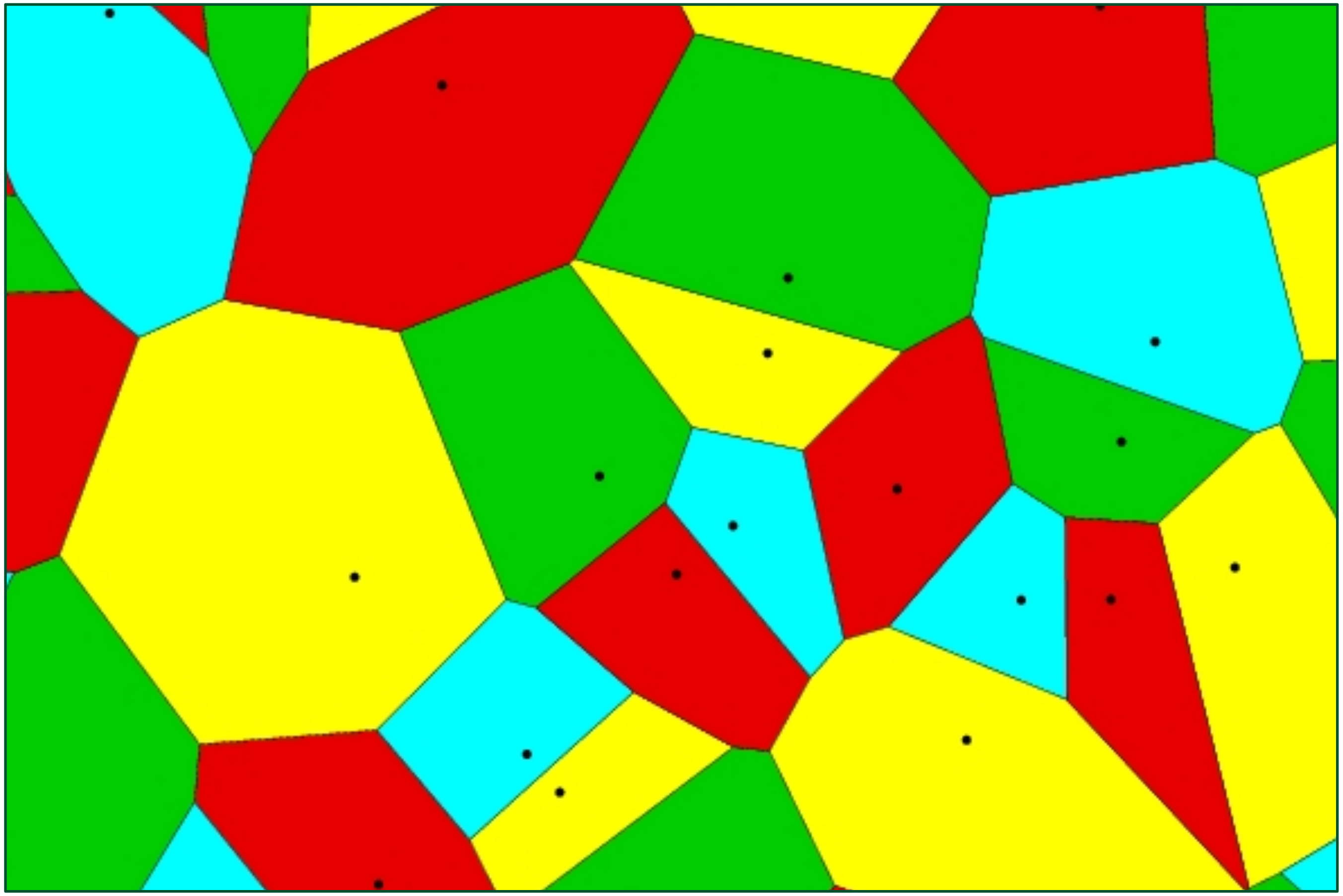


Nice R implementation: `tgp` package.

Voronoi tessellations

UNIVERSITY OF LEEDS

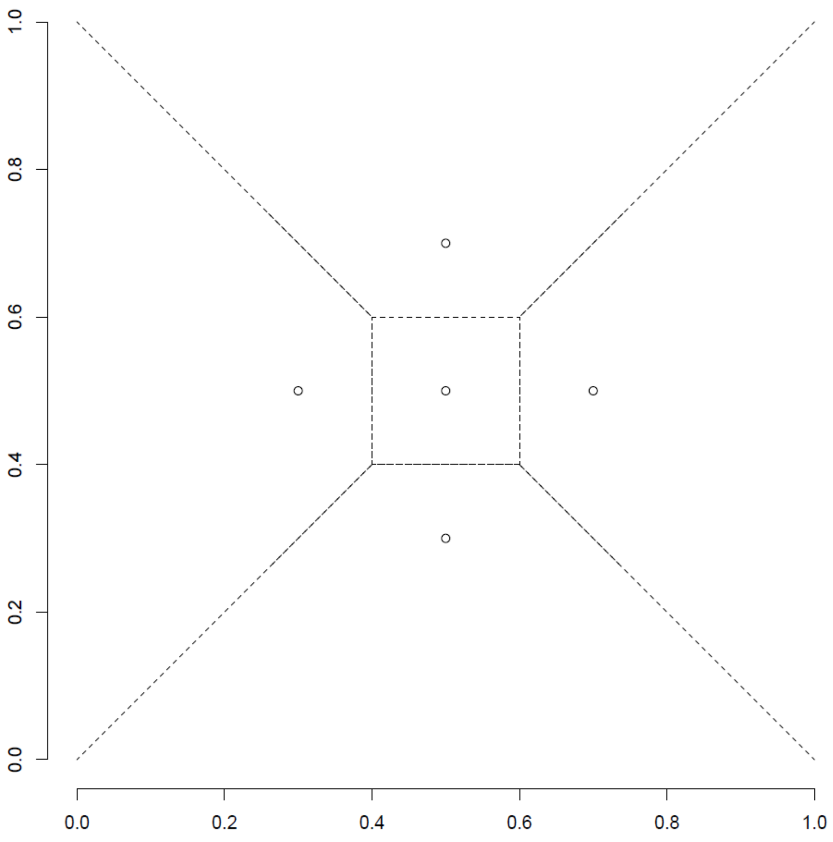
Tiles are defined completely by a set of centres.

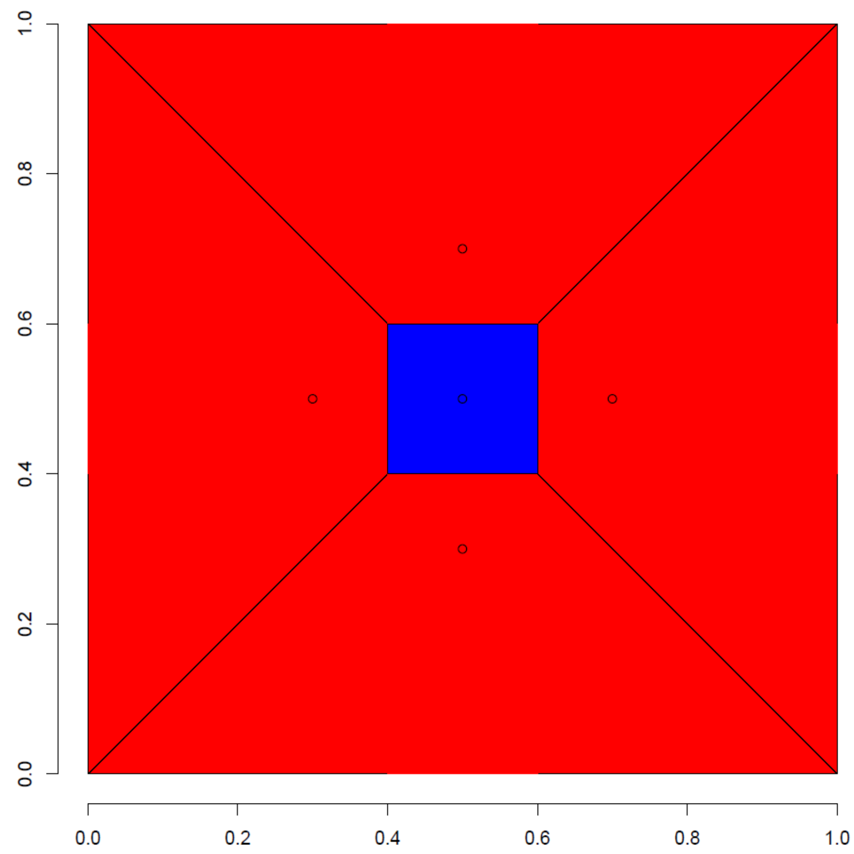


A point lies on a tile if it is closest to that tile's centre.

A point lies on a boundary if it is equally close to more than one centre.

Voronoi tessellations





Note that our "regions" do not need to be made up of neighbouring tiles.

Our model

UNIVERSITY OF LEEDS

Input space is divided into disjoint regions: each contain a number of Voronoi tiles.

Each region has an independent GP model:

$$l(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{t}) \propto \prod_{i=1}^{z} \pi\left(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i,\sigma_i^2,\boldsymbol{\beta}_i,\boldsymbol{t}\right),$$

where $\pi(.|.)$ denotes a multivariate normal pdf derived from the GP model.

We have extended the model of Kim *et al.* (2005) in several ways.

Priors

The prior for the region specification is

$$\pi(\boldsymbol{t}) = \pi(m, \boldsymbol{c})\pi(z|m)\pi(\boldsymbol{Q}|m, z).$$

Poisson point process with some sensible intensity

Discrete uniform over ordered partitions: $[1, m^{\text{th}} \text{ Bell number}]$

Discrete uniform over [1,m]

And an additional prior constraint that says we can only have a region if there are enough training points to fit a GP.

Implementation

UNIVERSITY OF LEEDS

Reversible-jump-MCMC:

GP MAP estimates within birth/death/move and relationship-change MH;

Implementation

UNIVERSITY OF LEEDS

Reversible-jump-MCMC:

GP MAP estimates within birth/death/move and relationship-change MH;

Start off **100 MCMC chains** from random points in model space;

Run each chain for **10,000 iterations** and checking for autocorrelation and convergence;

Implementation

Reversible-jump-MCMC:

GP MAP estimates within birth/death/move and relationship-change MH;

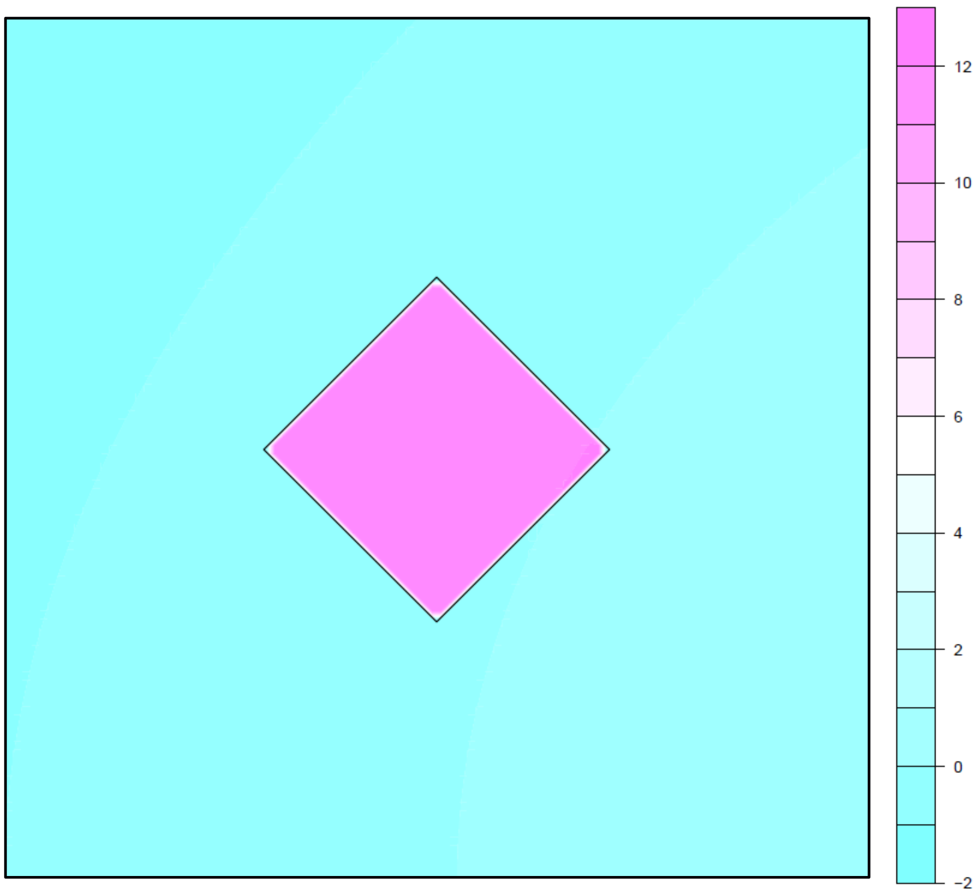
Start off **100 MCMC chains** from random points in model space;

Run each chain for **10,000 iterations** and checking for autocorrelation and convergence;

Hope that you have something that has converged…

We need to be wary of identifiability issues and local maxima.

Toy example

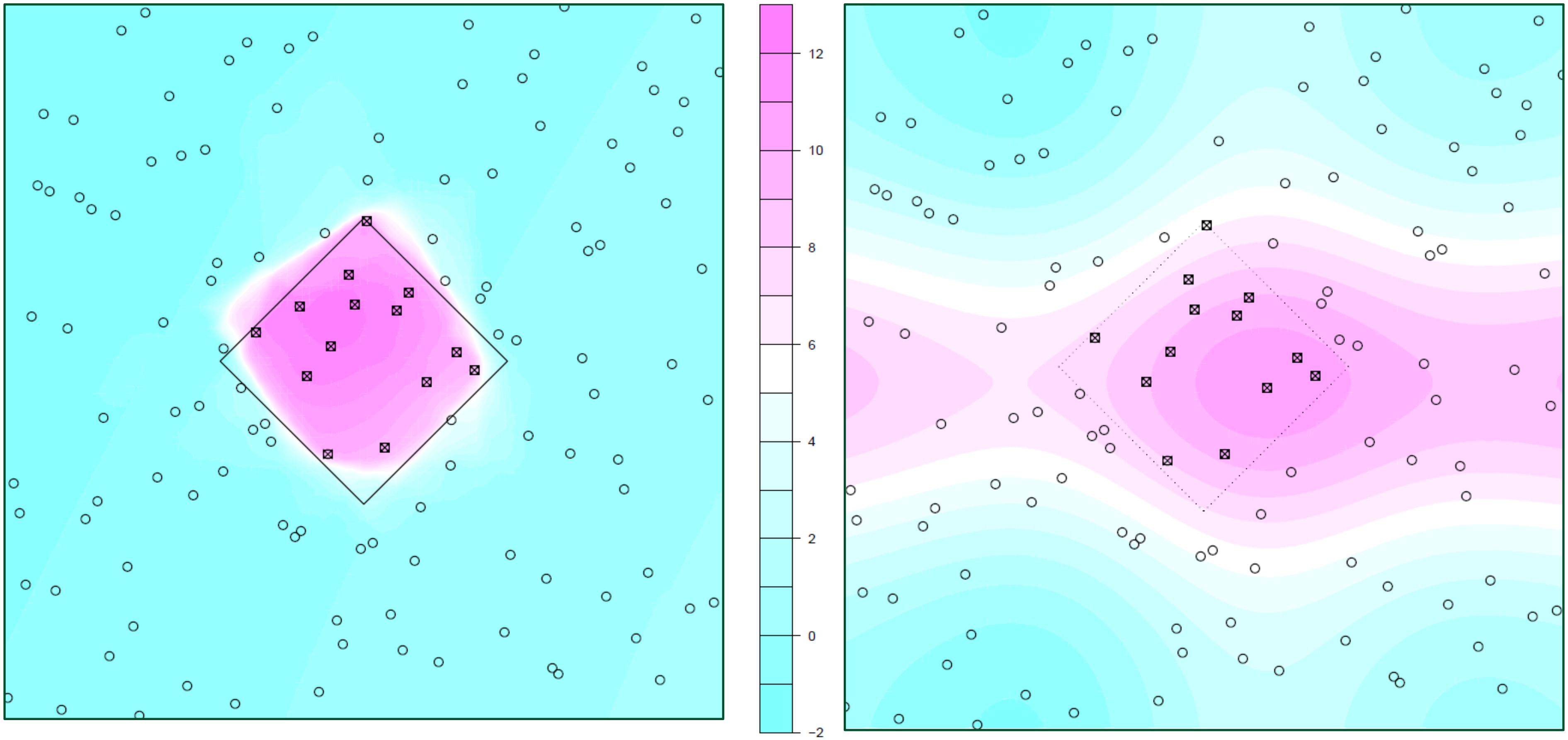


Another step function.

Toy example

UNIVERSITY OF LEEDS

Predictive mean

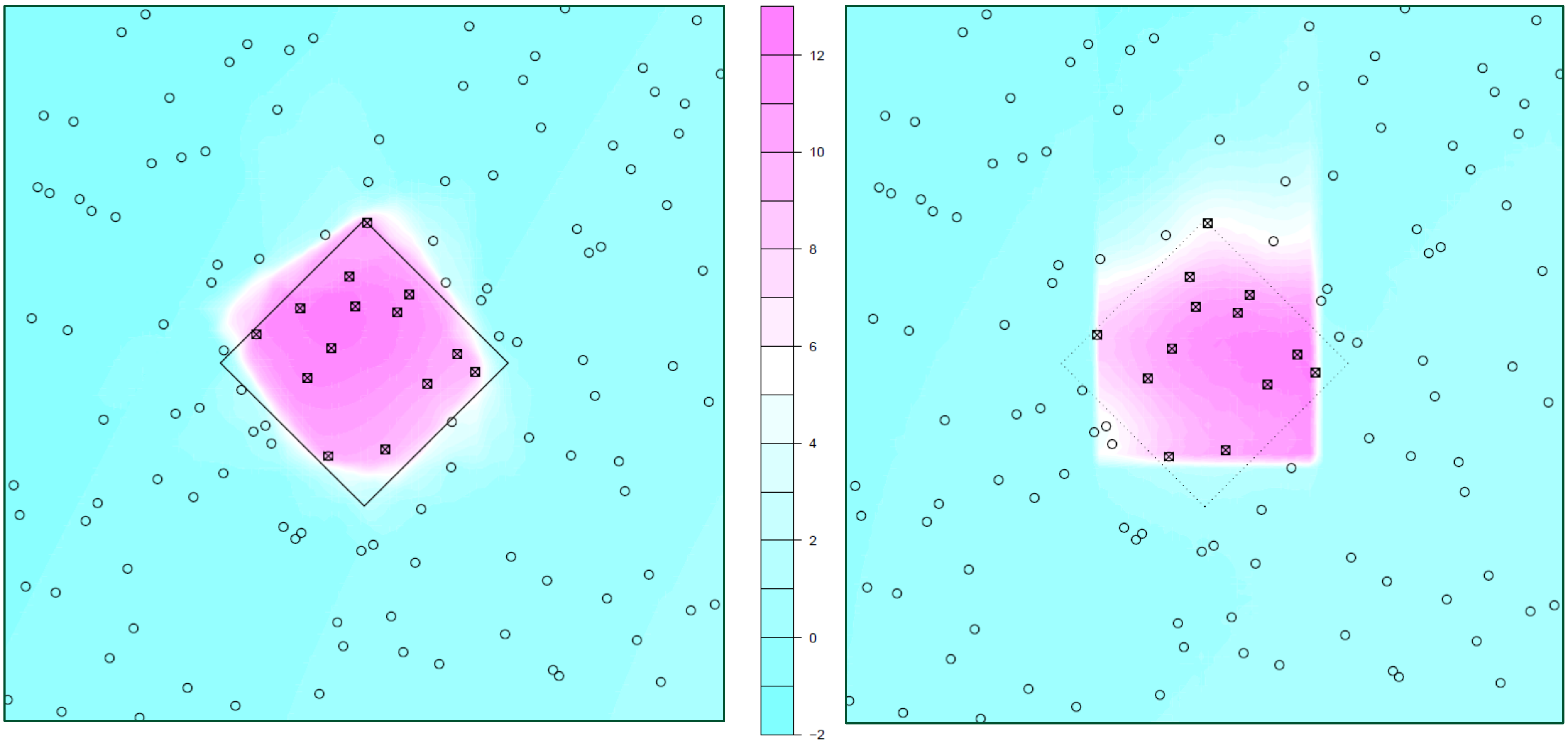


Voronoi GP model **Standard GP model**

Toy example

UNIVERSITY OF LEEDS

Predictive mean

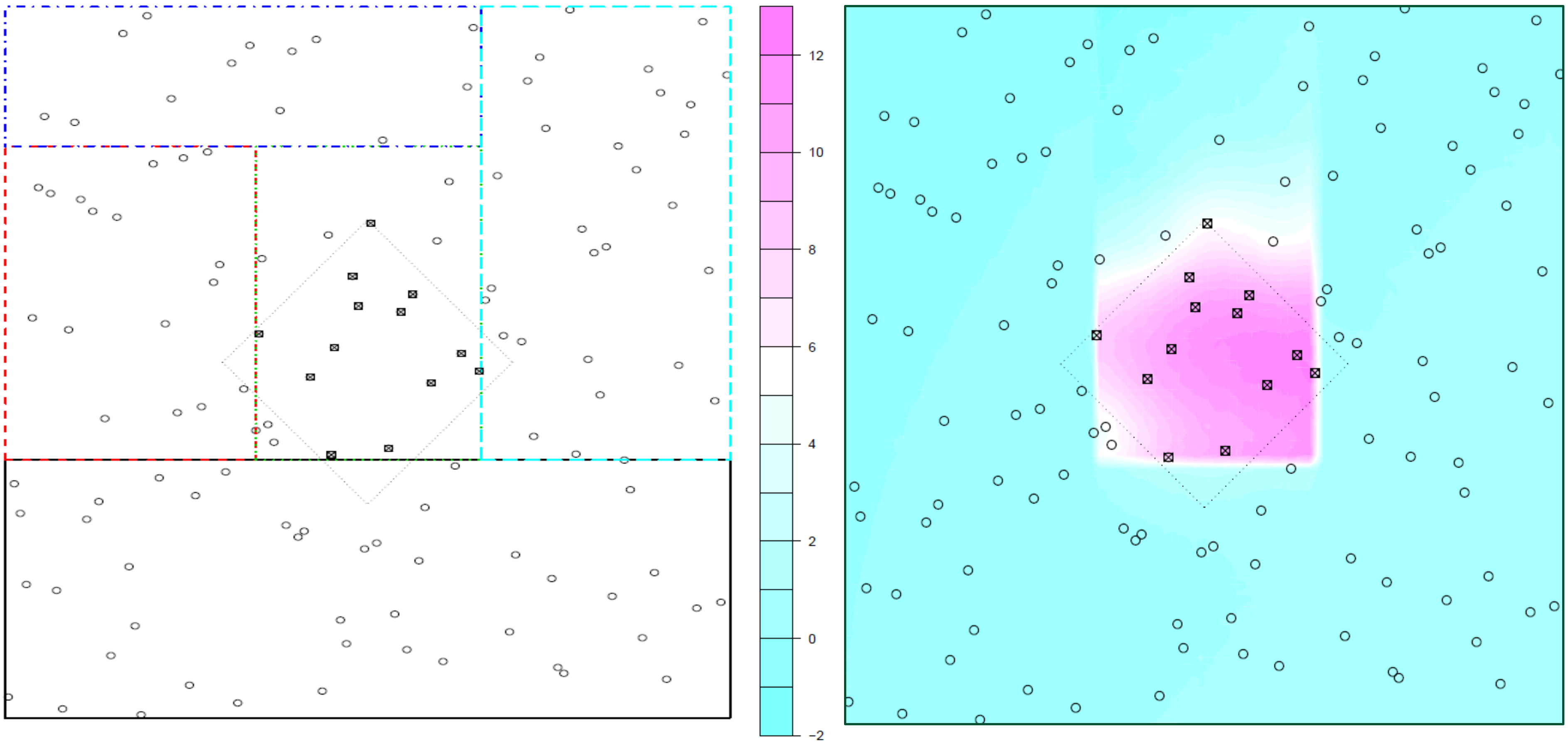


Voronoi GP model **Treed GP model**

Toy example

UNIVERSITY OF LEEDS

MAP division | Predictive mean

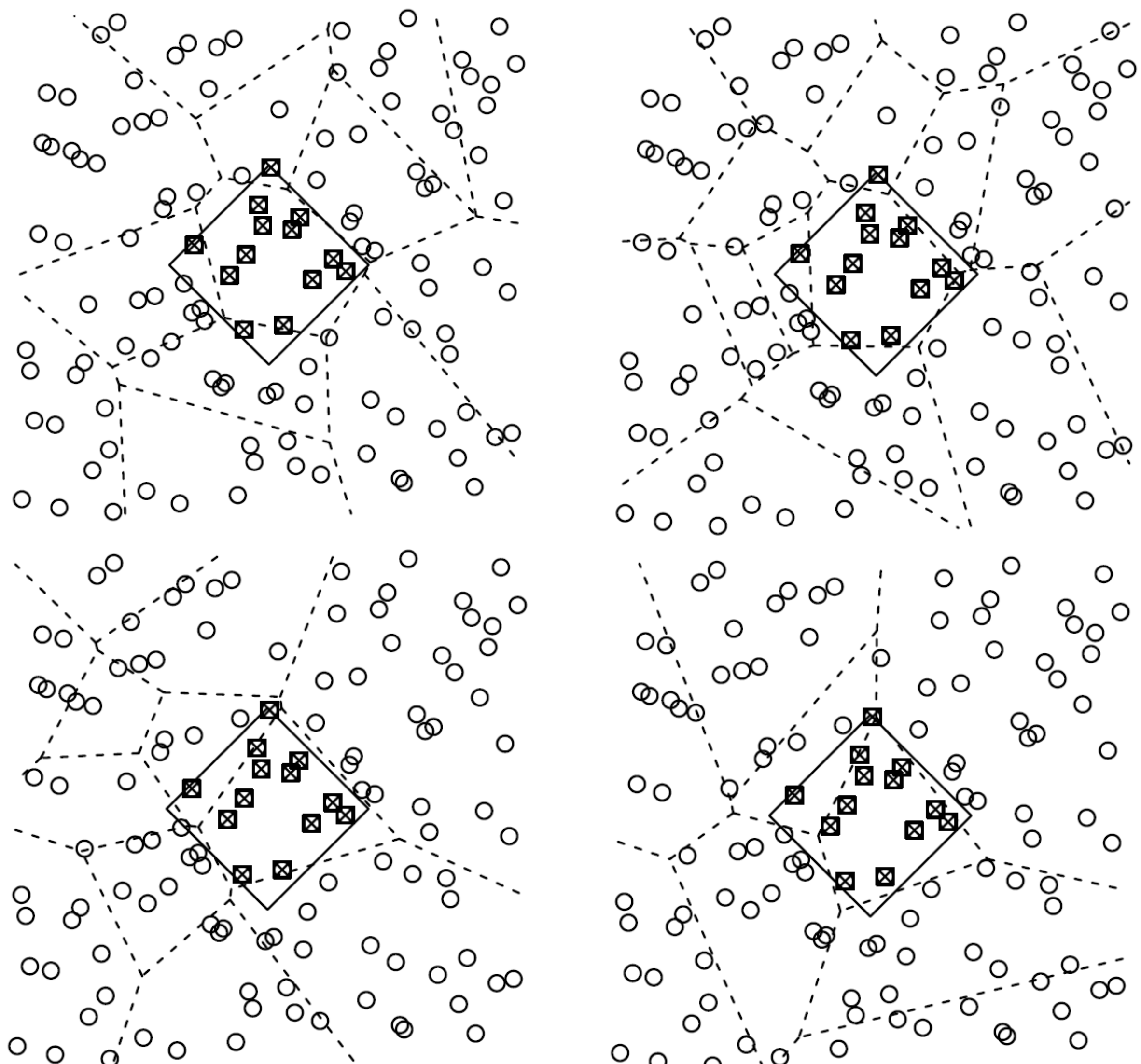


Treed GP model

Toy example

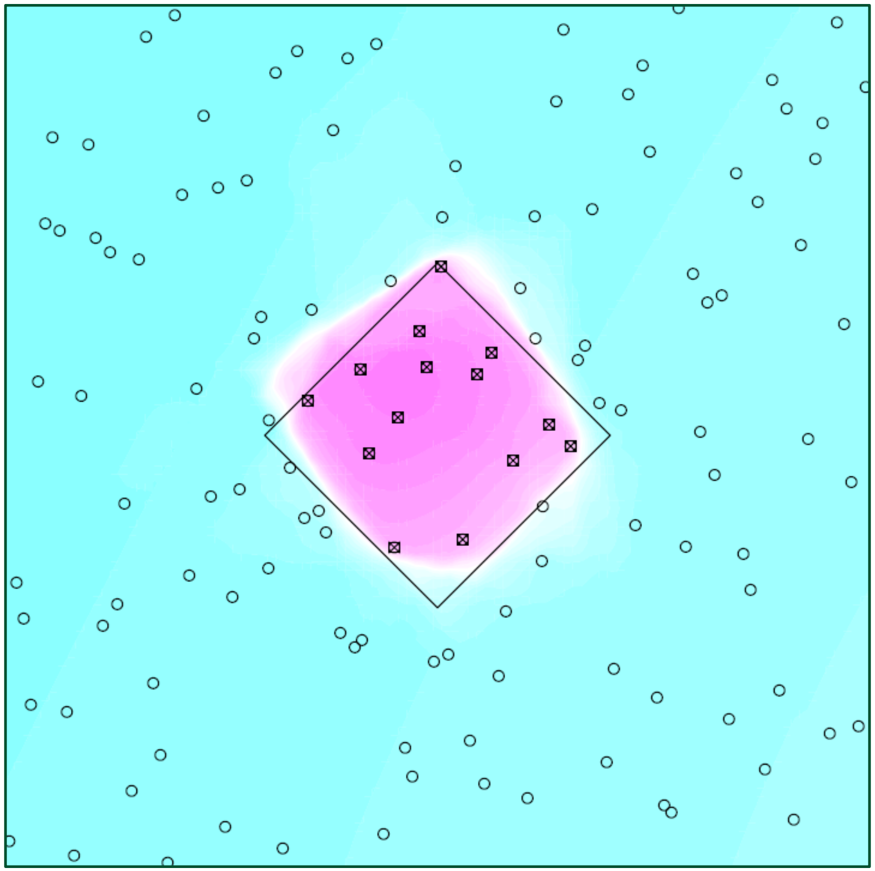
UNIVERSITY OF LEEDS

From our posterior, we can get various plausible tessellations.

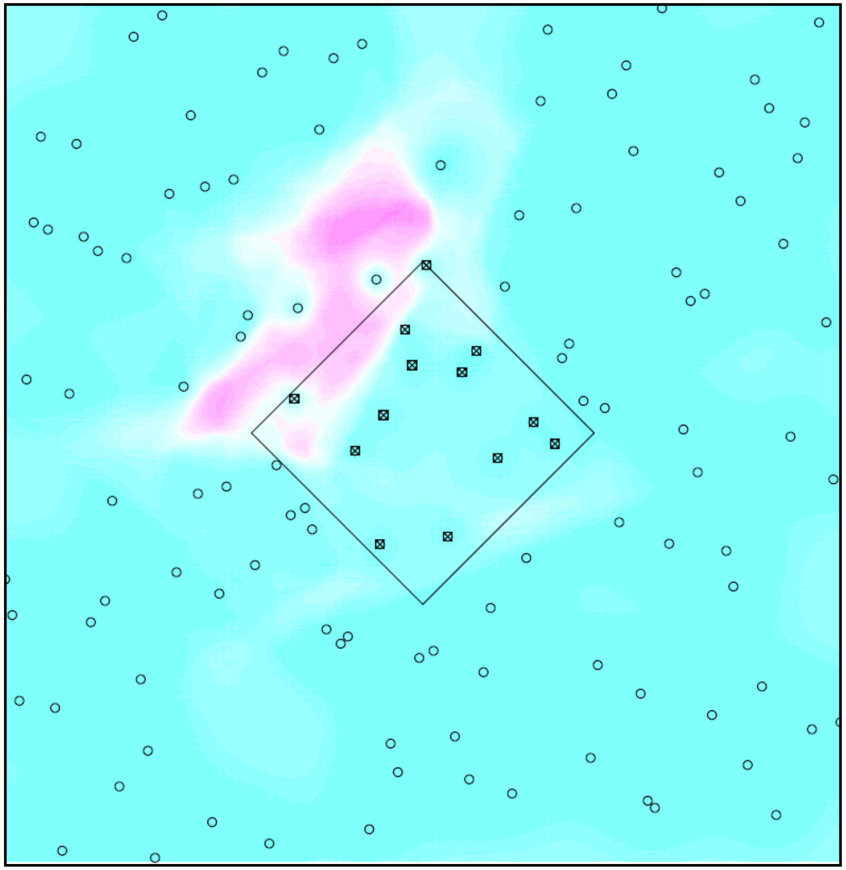


Toy example

Predictive mean

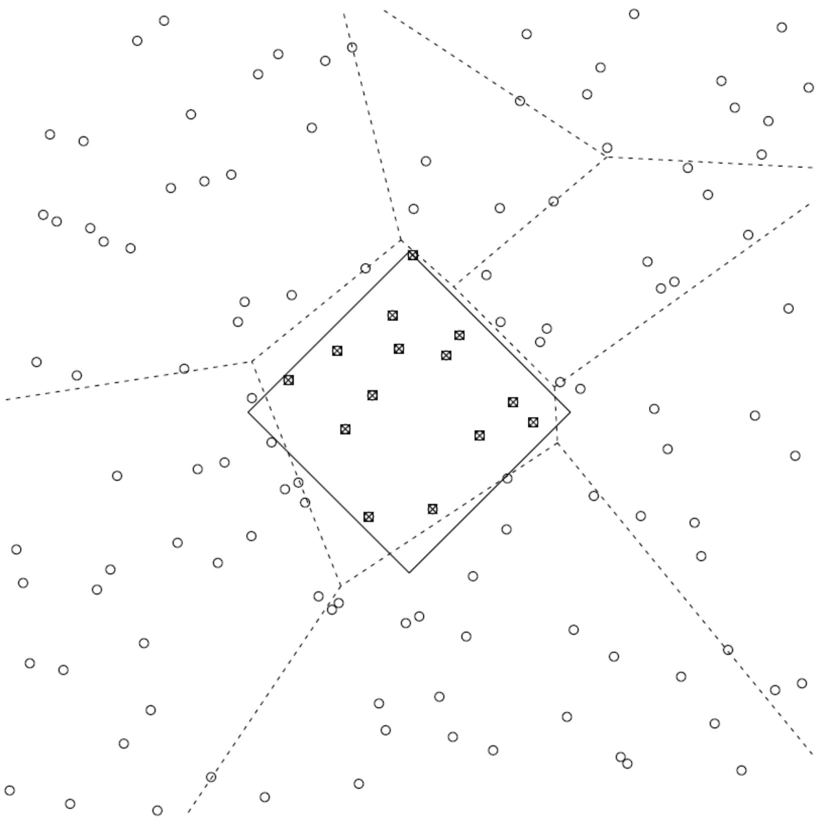


Predictive std dev.

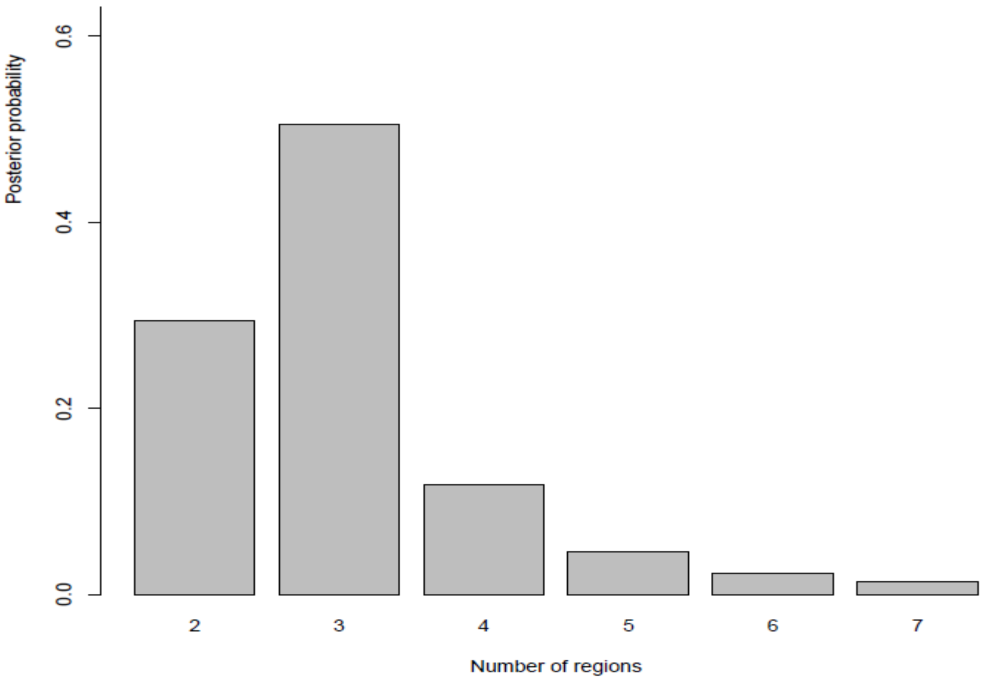


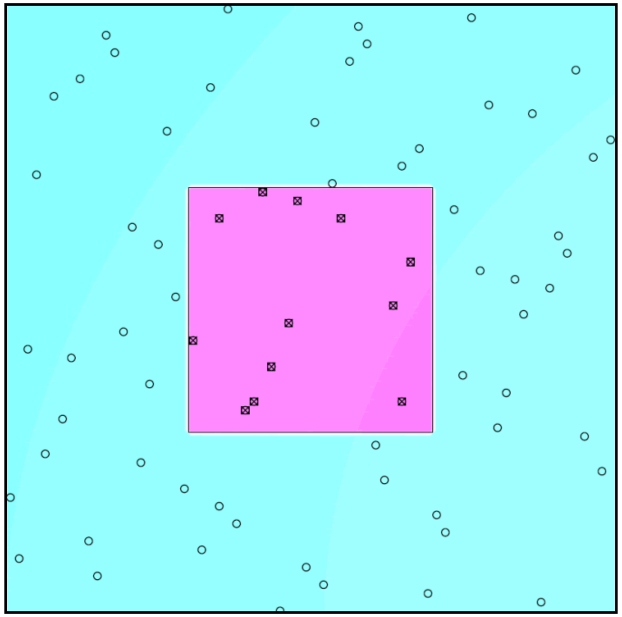
Voronoi GP model

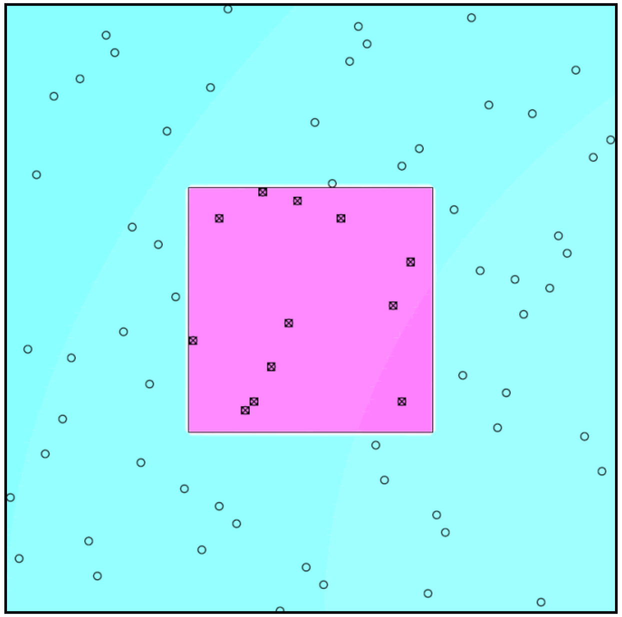
Toy example

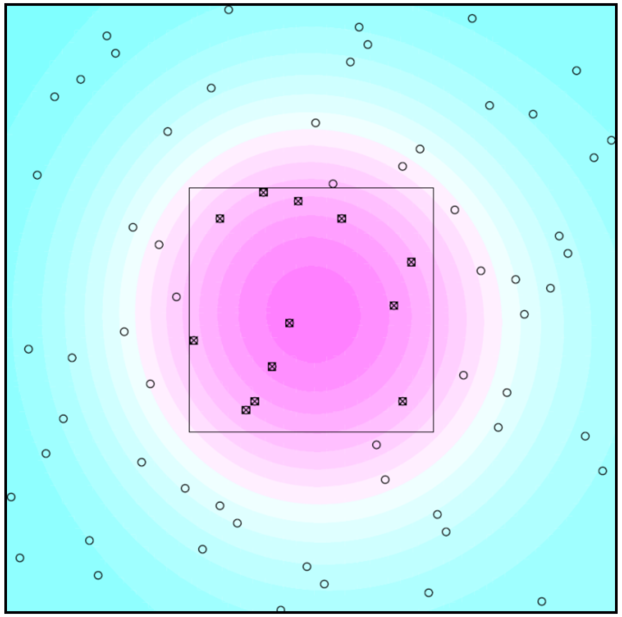


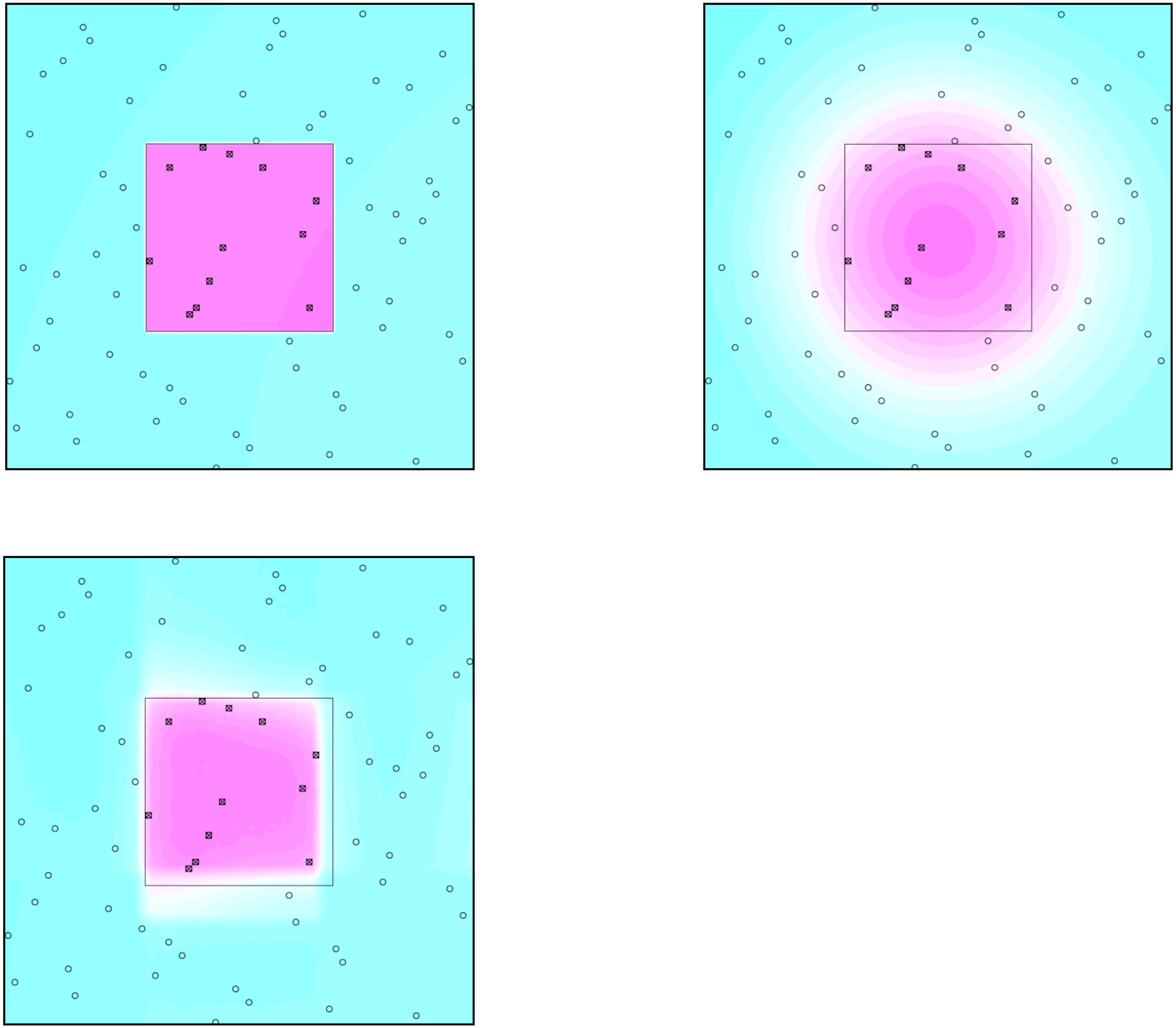
We can also look at our MAP tessellation and the probabilities of getting different numbers of regions.

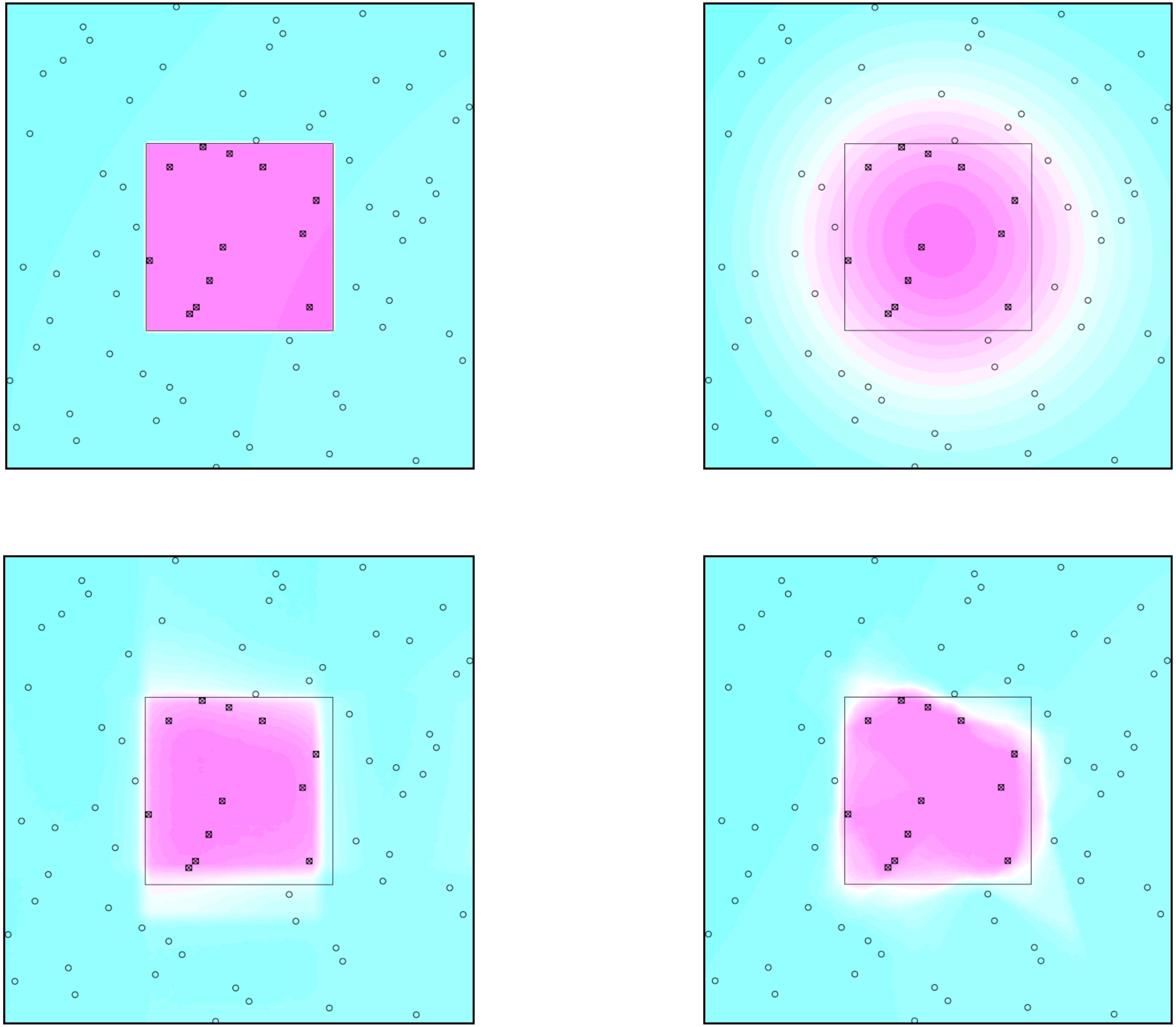


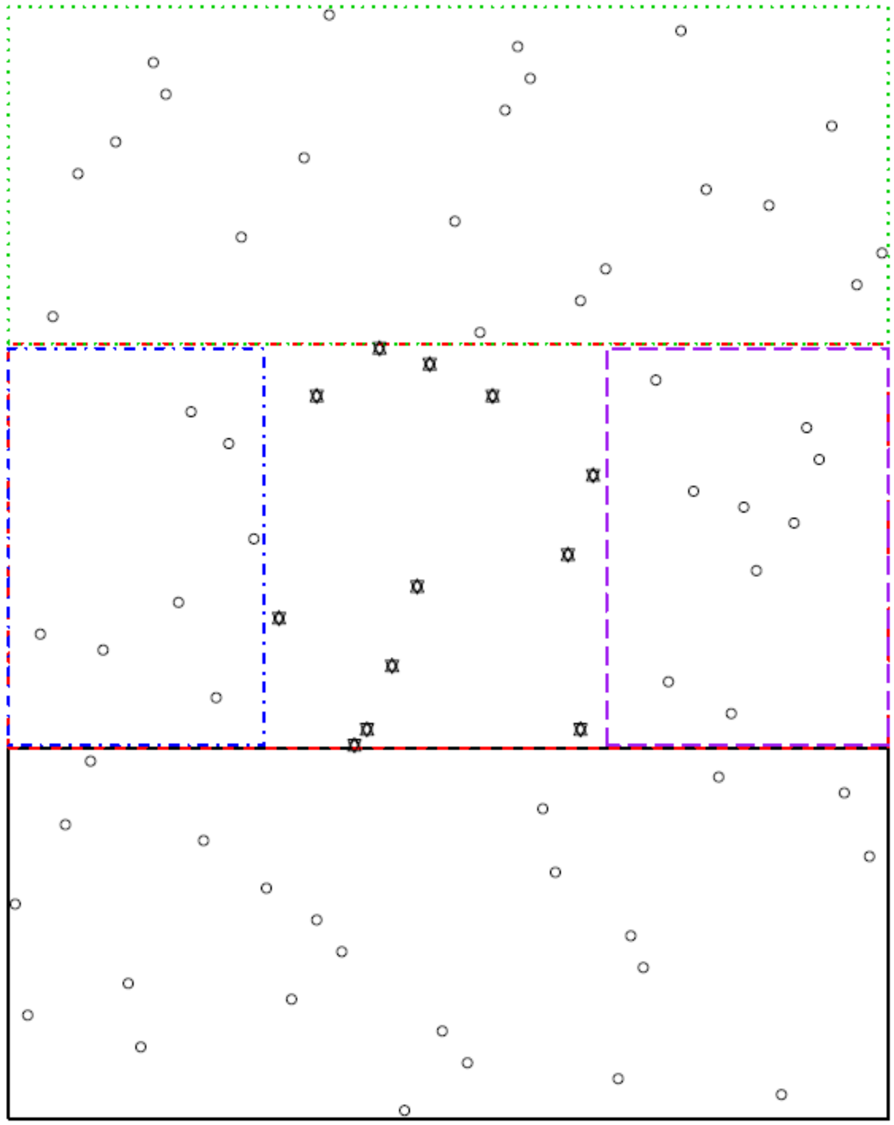


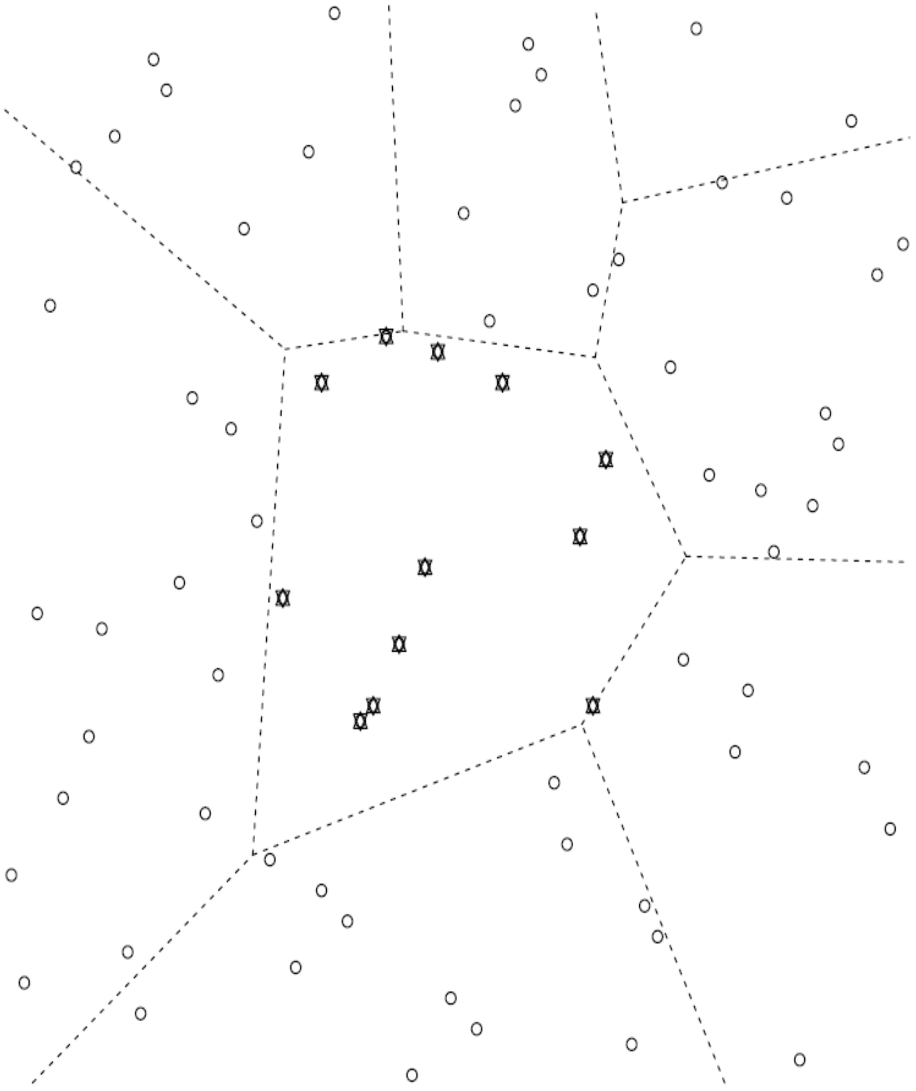










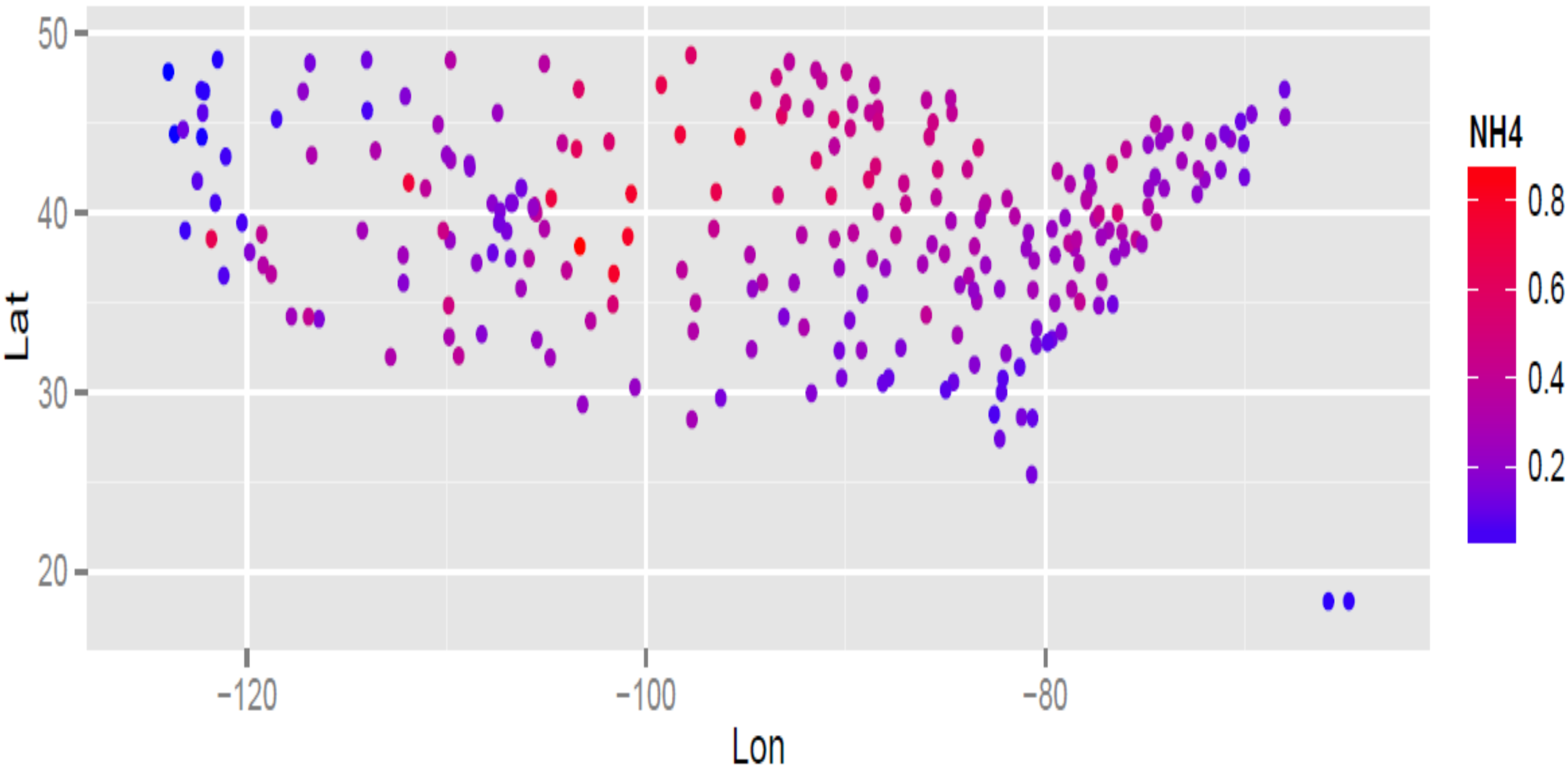


MSE: 1.98 **MSE: 1.84**

Spatial statistics example

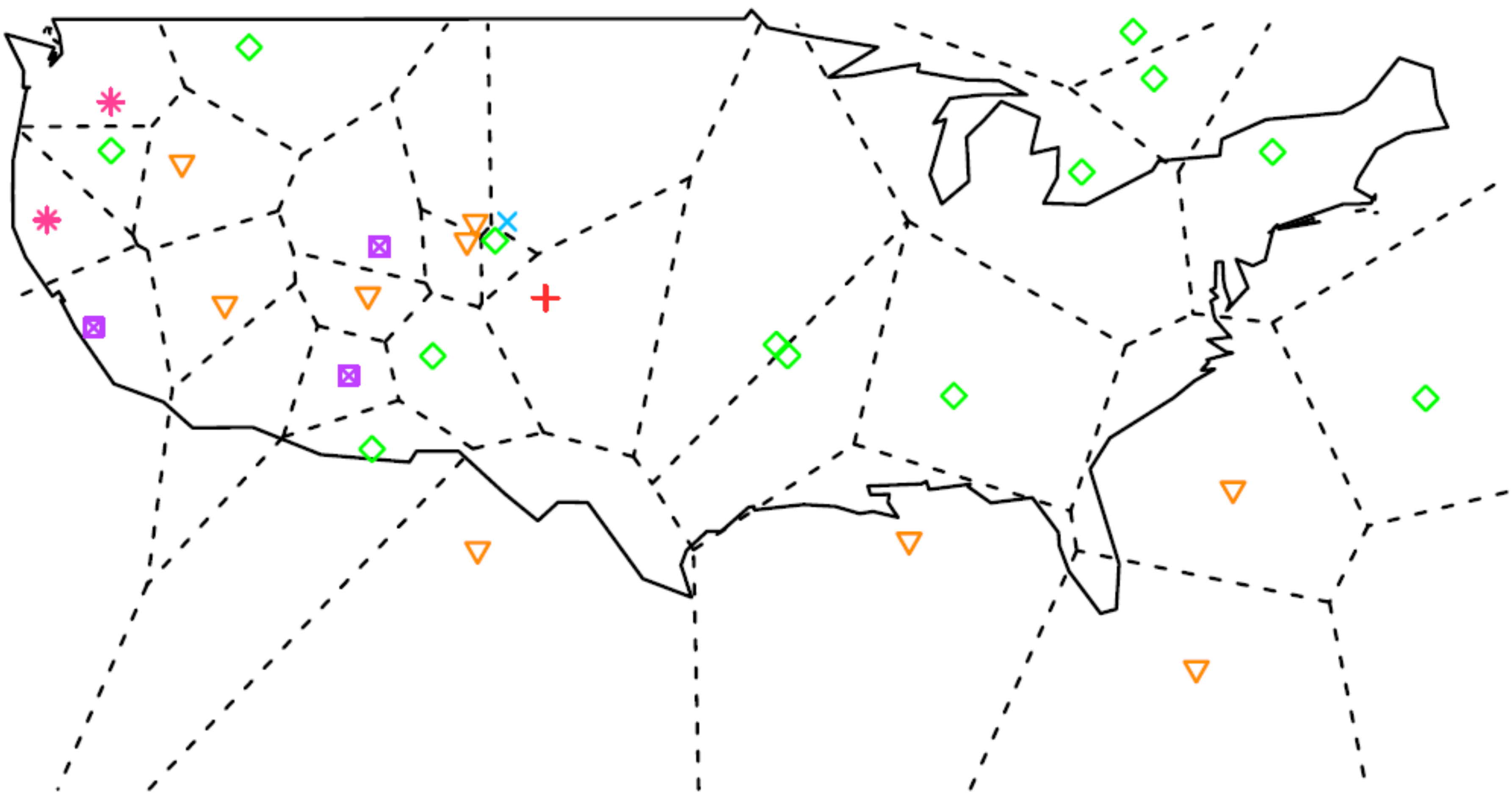
UNIVERSITY OF LEEDS

In the context of traditional kriging, we considered ammonia concentration at ground level across 250 US sites (2007).



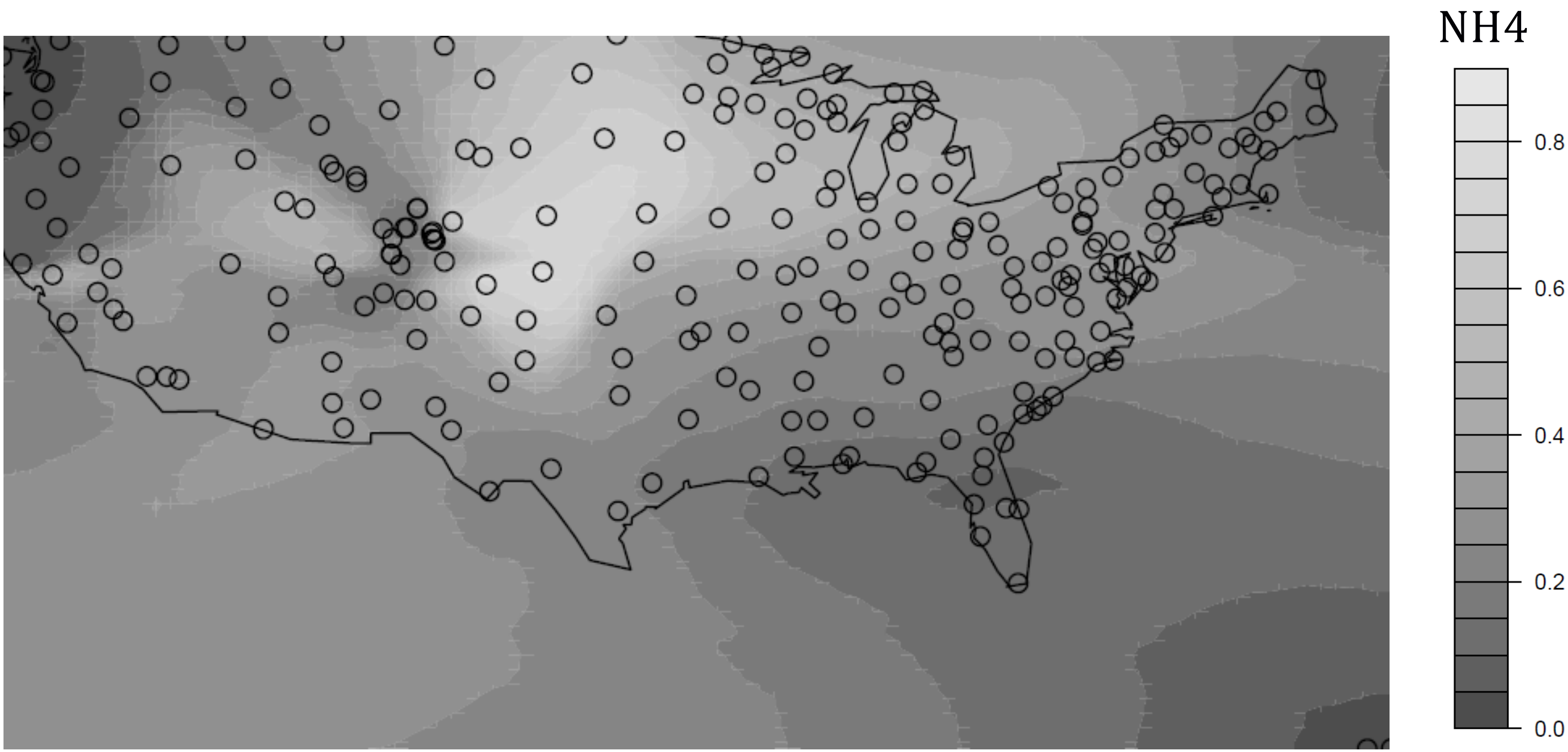
Spatial statistics example

UNIVERSITY OF LEEDS



Spatial statistics example

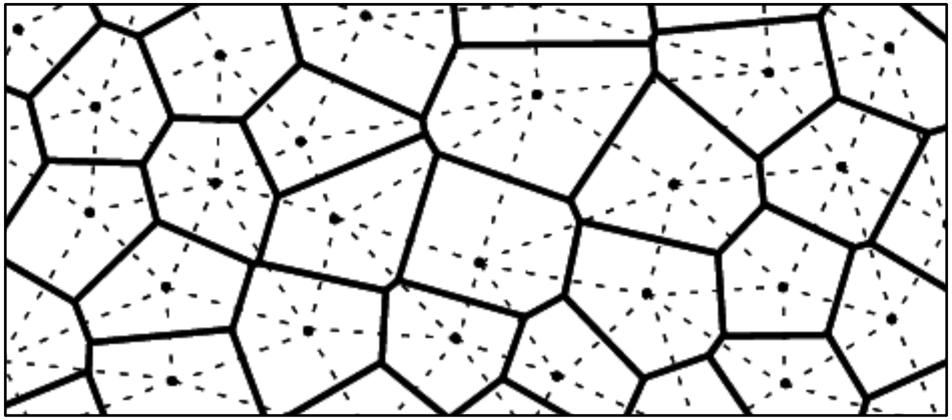
UNIVERSITY OF LEEDS



Proposing new points

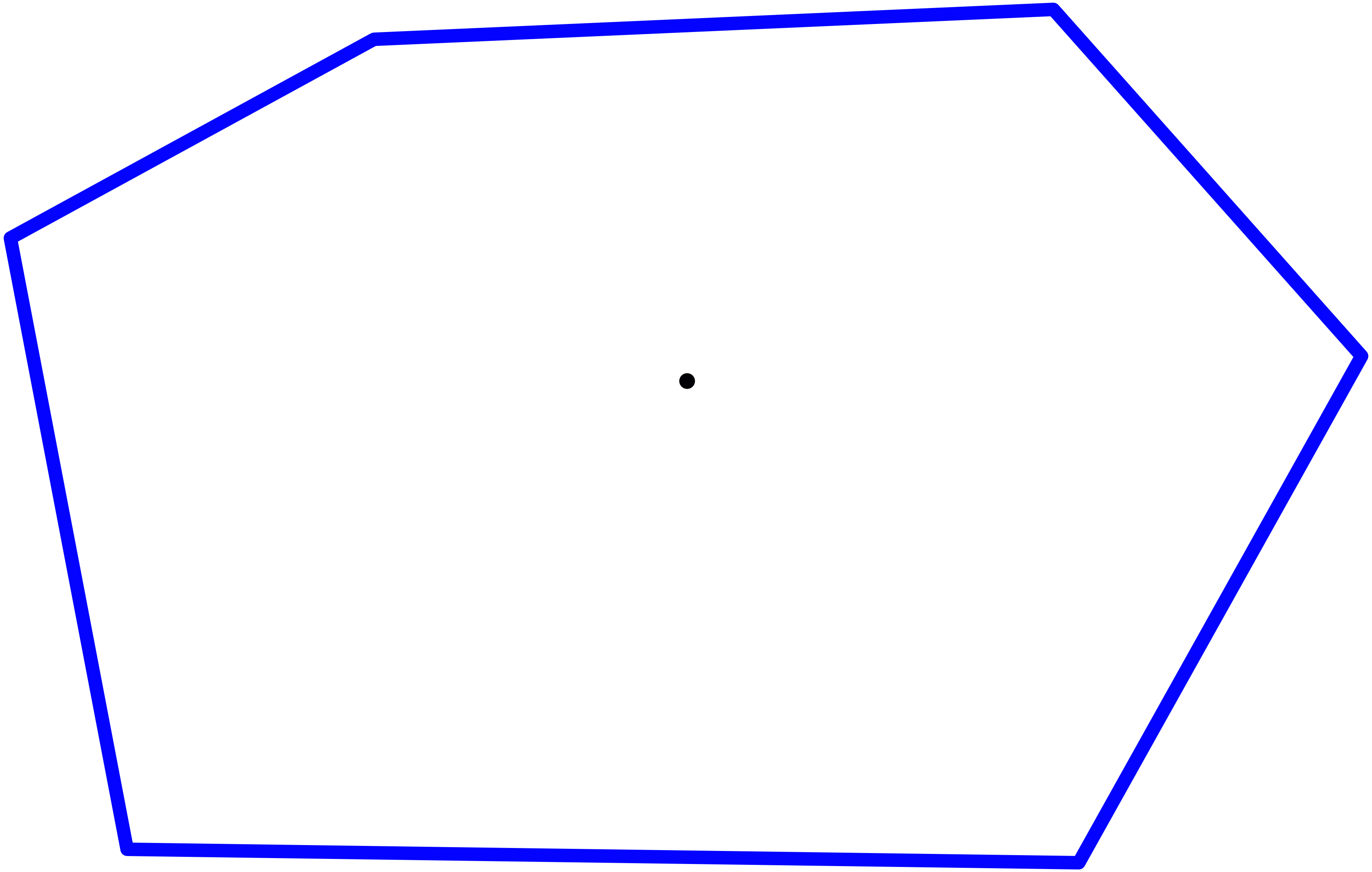
To improve estimation, we could

1) Target areas with high uncertainty;
2) Just continue with a space filling theme;
3) Try to improve our estimation of the region boundaries.

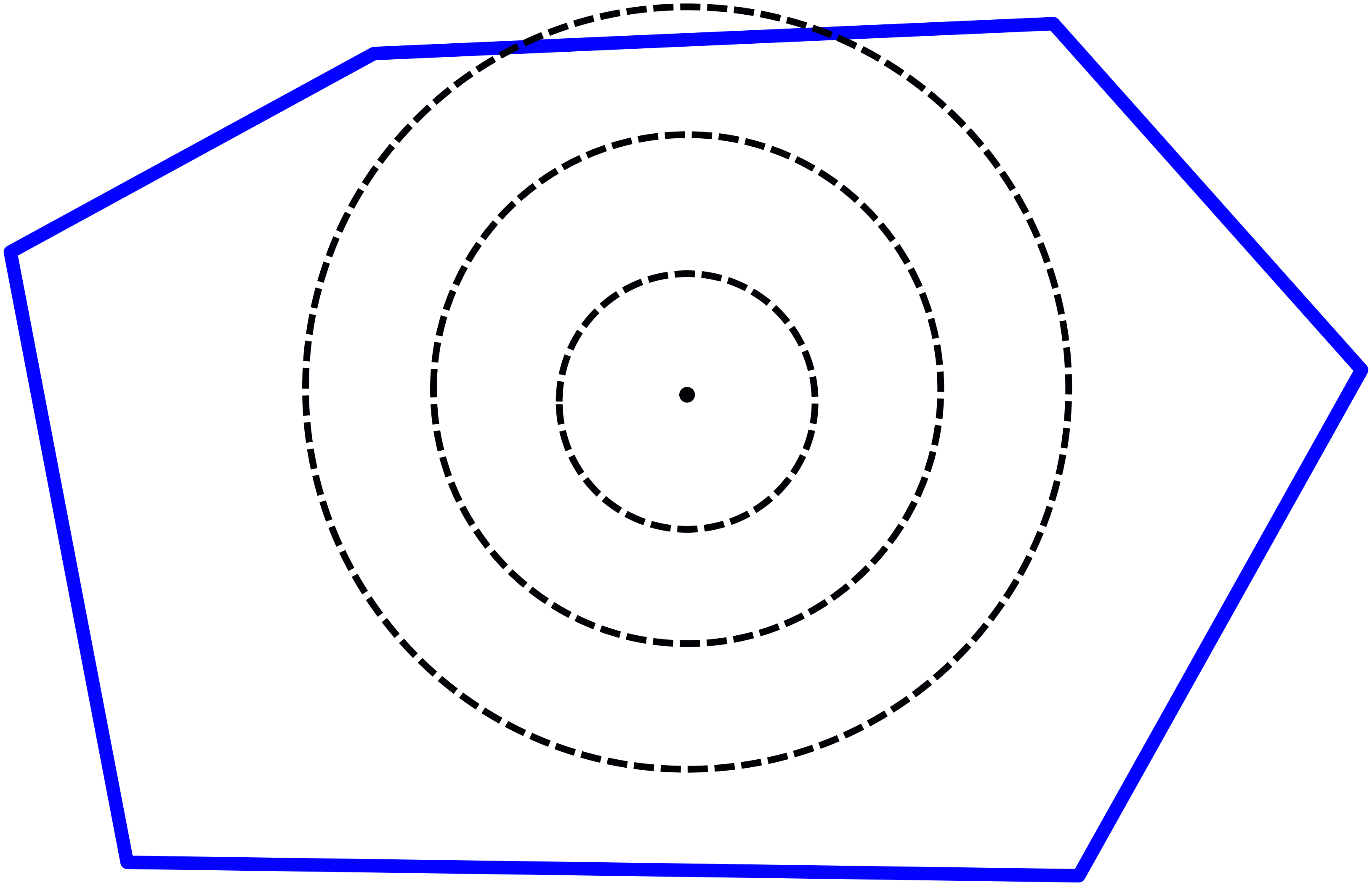


Finding points that lie on a boundary in 2d is relatively simple.

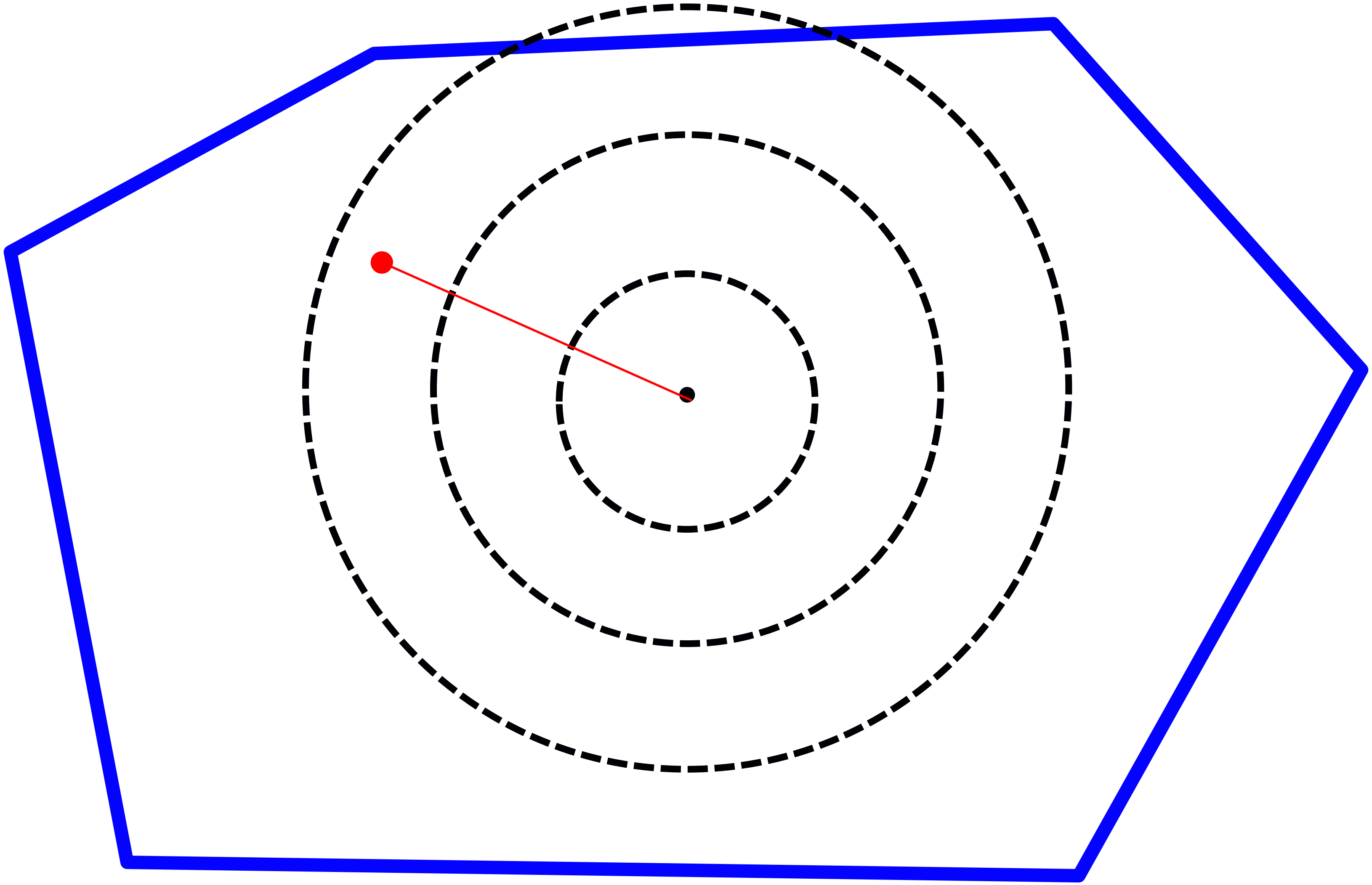
Finding the boundary



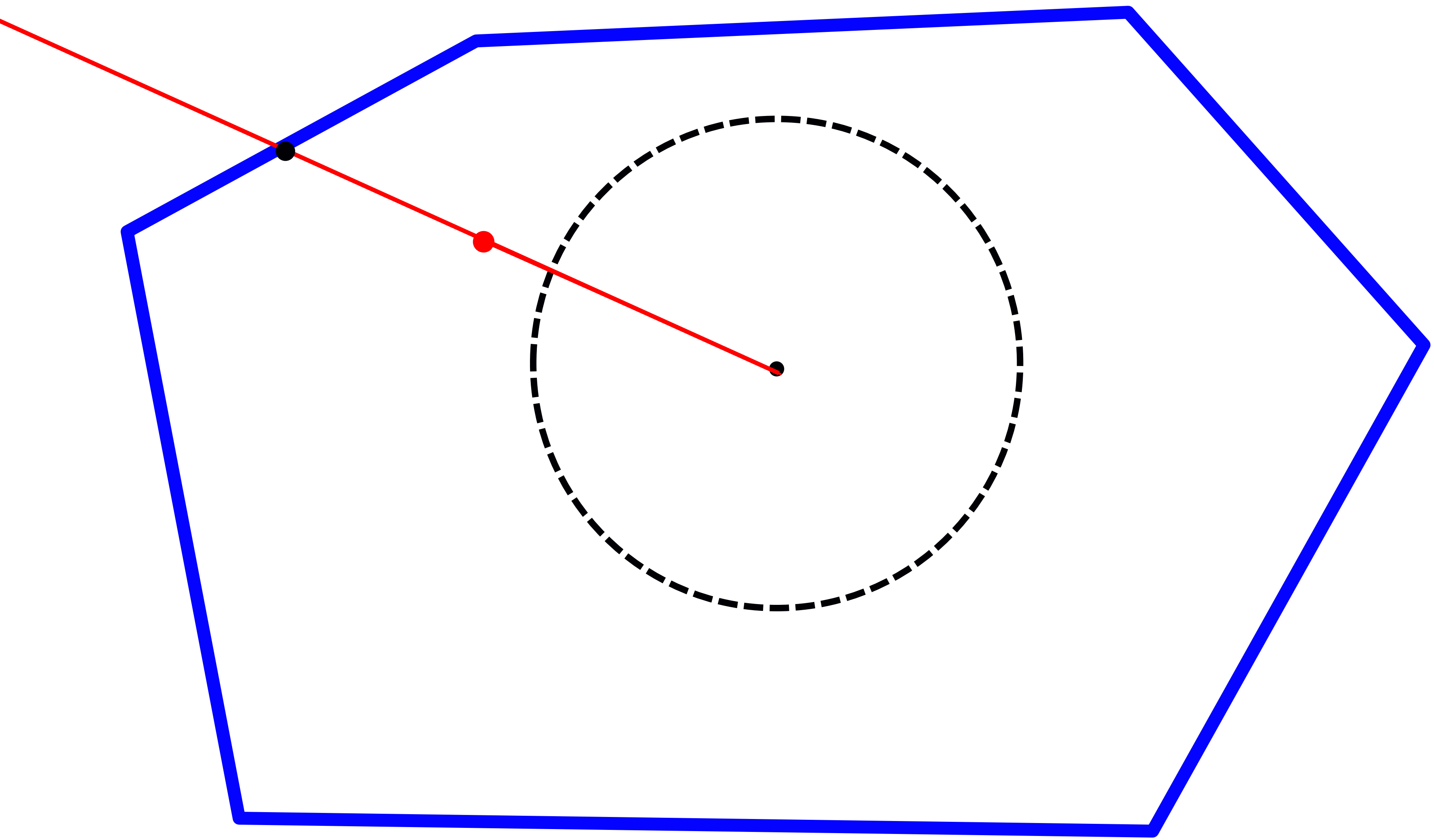
Finding the boundary



Finding the boundary

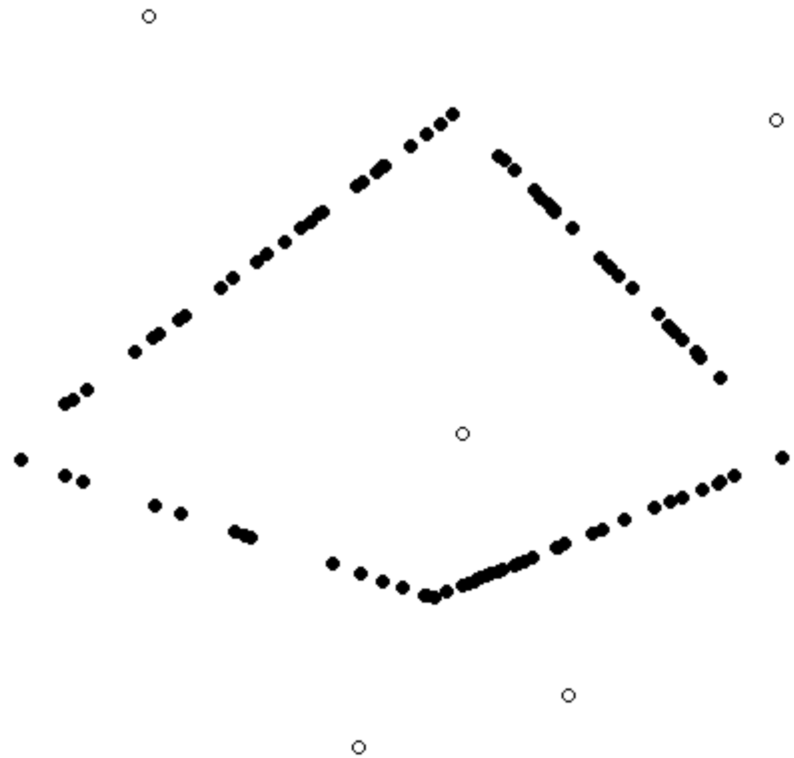


Finding the boundary

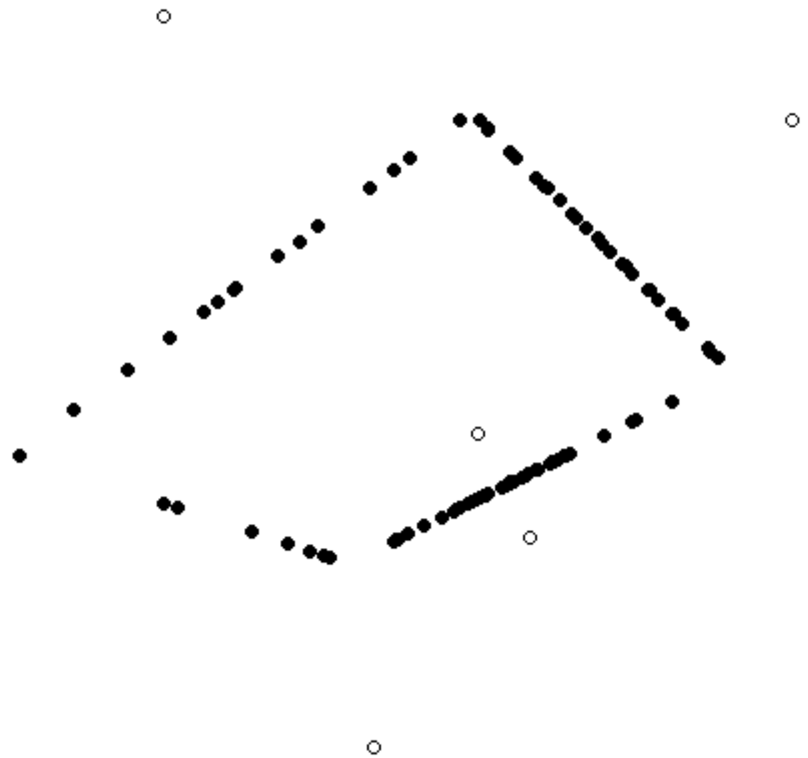


Finding the boundary

UNIVERSITY OF LEEDS

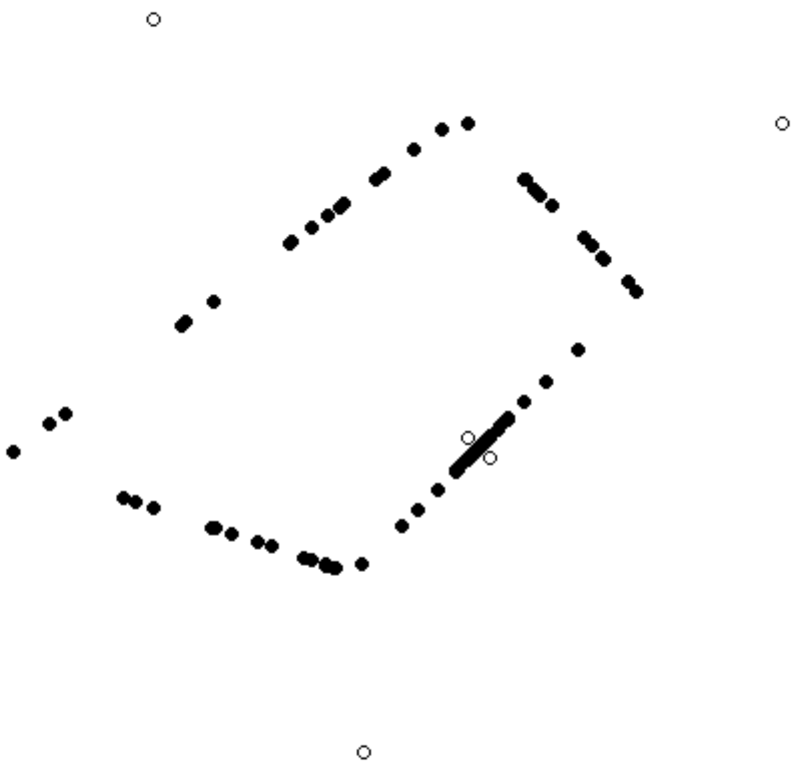


Finding the boundary

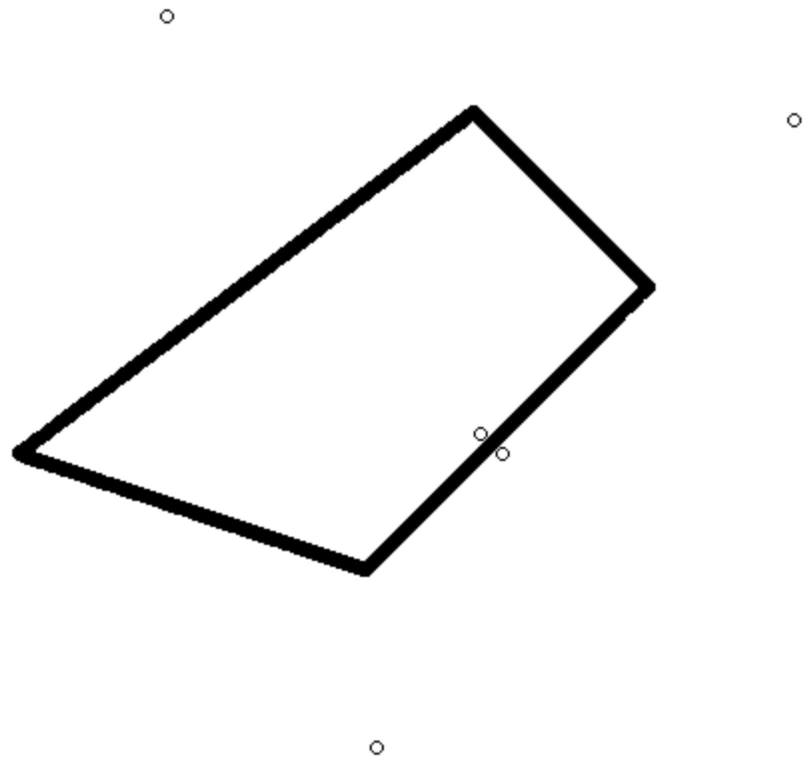


Finding the boundary

UNIVERSITY OF LEEDS



Finding the boundary



Proposing new points

Algorithm that lacks subtlety:

1) Randomly choose a centre within region of interest.
2) Randomly choose a point on the boundary from that centre's Voronoi tile.
3) Check if the point is on edge of region, and keep if it is.
4) Repeat 1-3 many times to get candidate set.

Proposing new points

UNIVERSITY OF LEEDS

Algorithm that lacks subtlety:

1) Randomly choose a centre within region of interest.
2) Randomly choose a point on the boundary from that centre's Voronoi tile.
3) Check if the point is on edge of region, and keep if it is.
4) Repeat 1-3 many times to get candidate set.

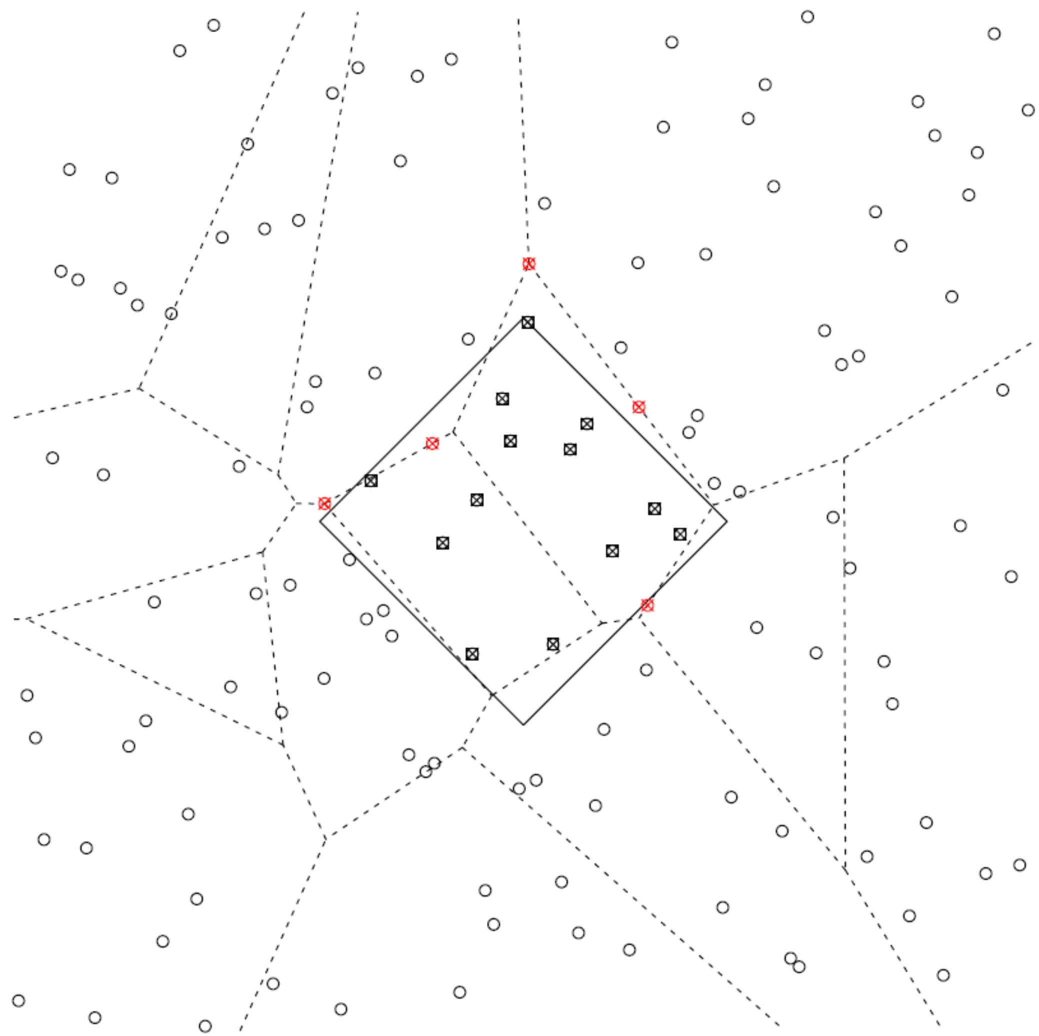
Then attempt to maintain space-filling property:

1) Find point in candidate set that is furthest from the training points.
2) Add that point to set of training points.
3) Find point in remaining candidate set that is furthest from the training points and the added point.
4) Add that point to set of training points.
5) Continue repeating process until enough points are found.

Toy example revisited

UNIVERSITY OF LEEDS

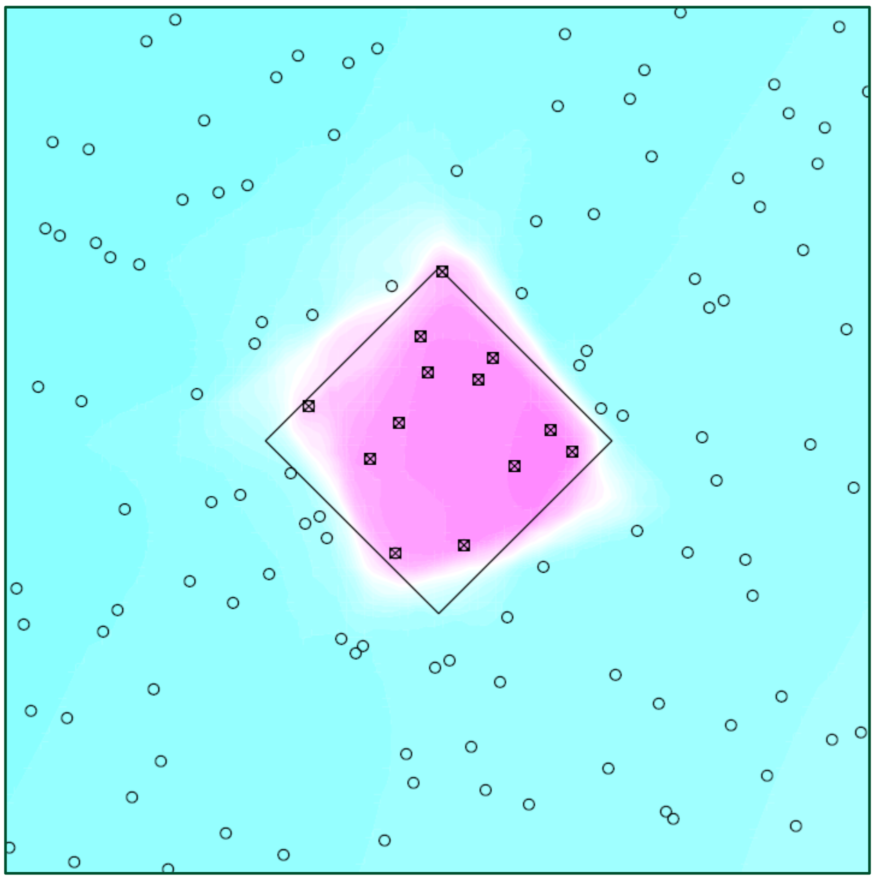
We decide that we can afford to sample at five extra points.



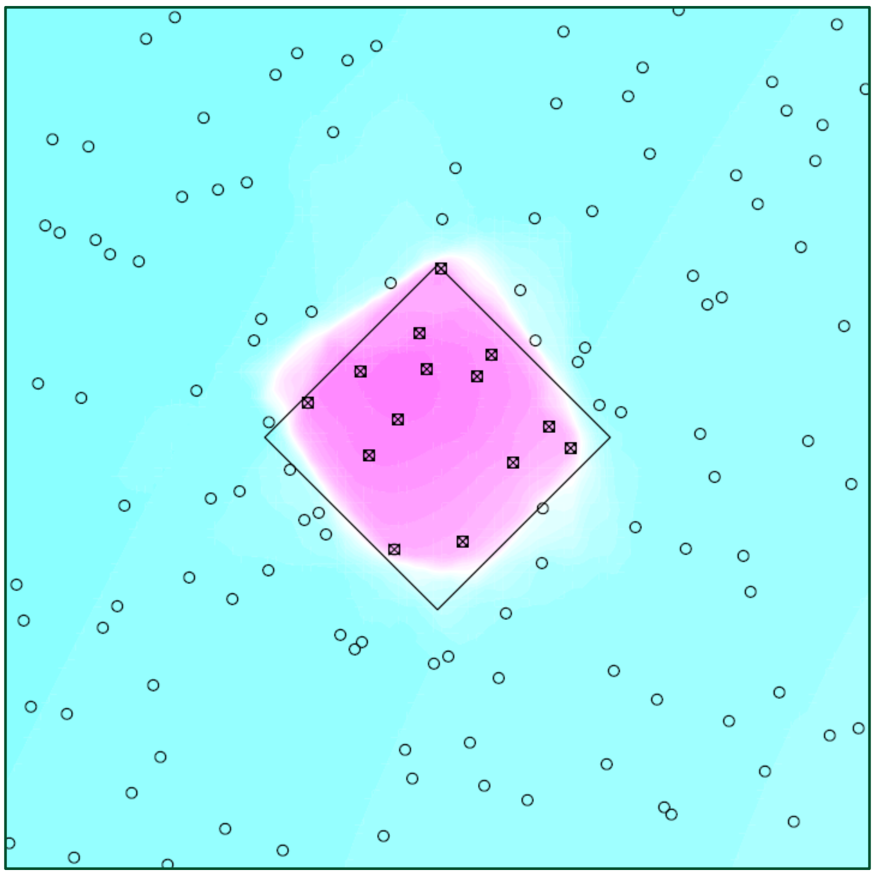
Toy example revisited

UNIVERSITY OF LEEDS

First data set



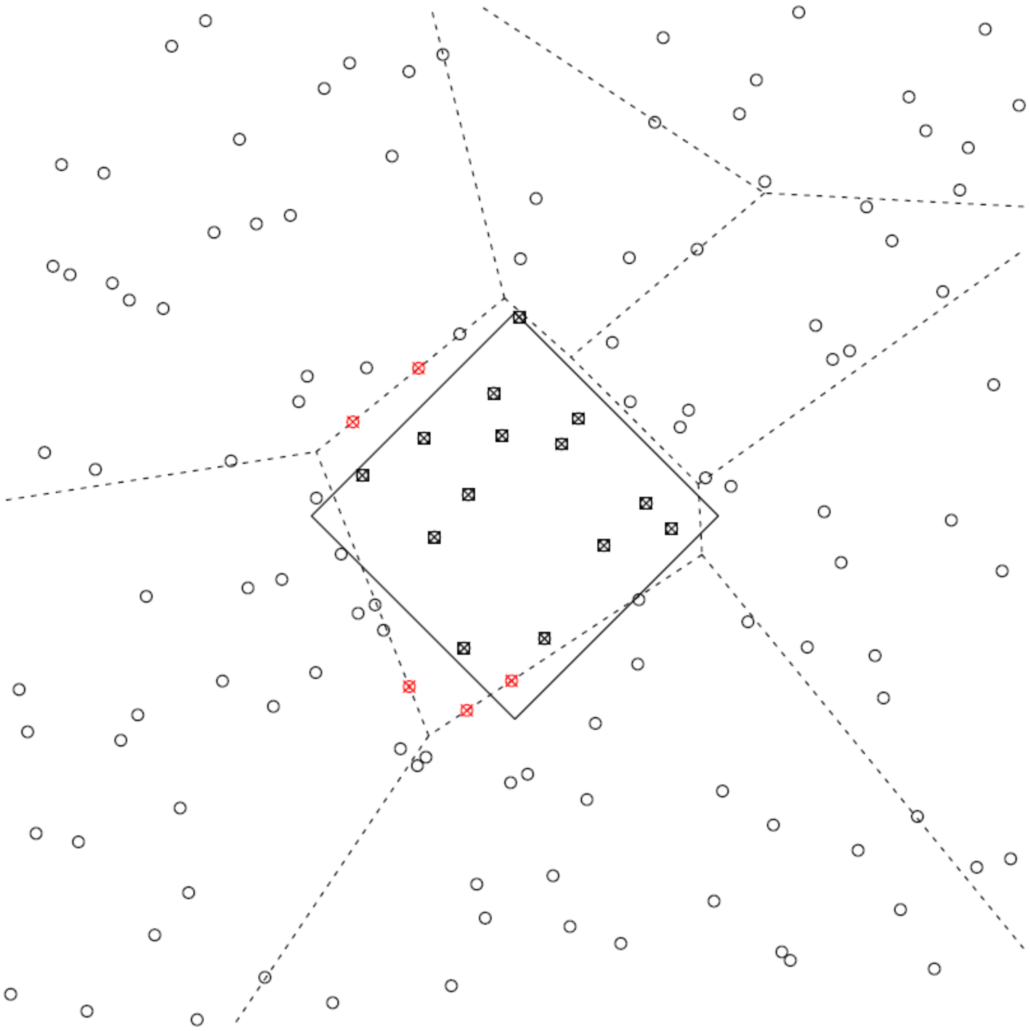
Second data set



Toy example revisited

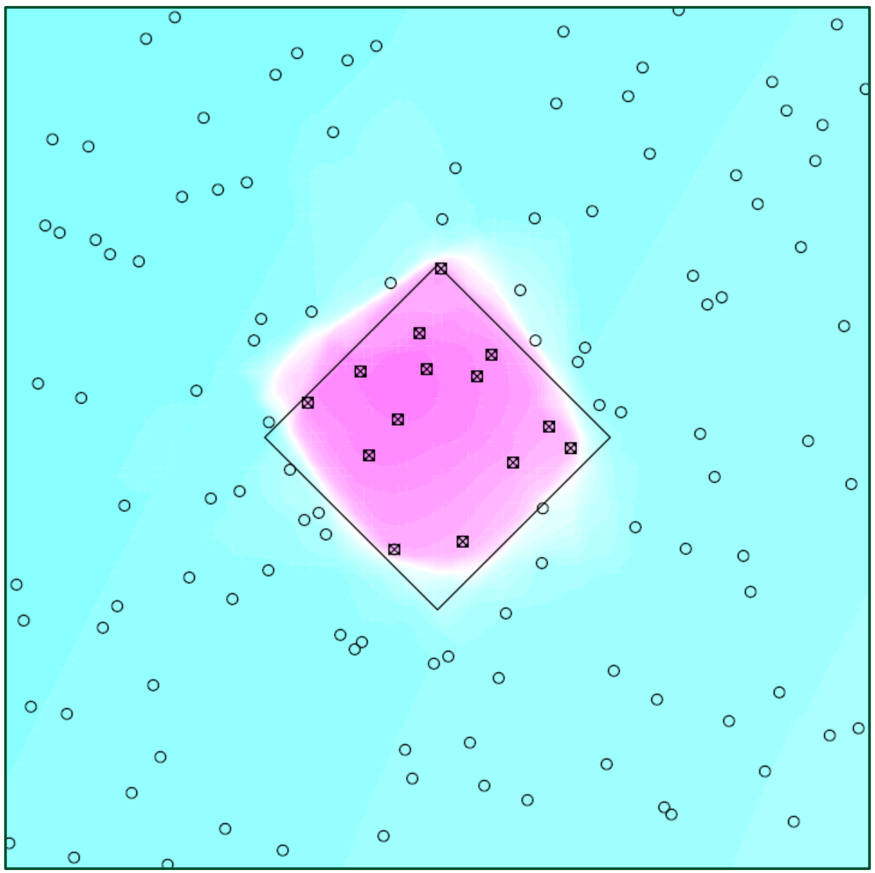
UNIVERSITY OF LEEDS

We decide that we can afford to sample again at five extra points.

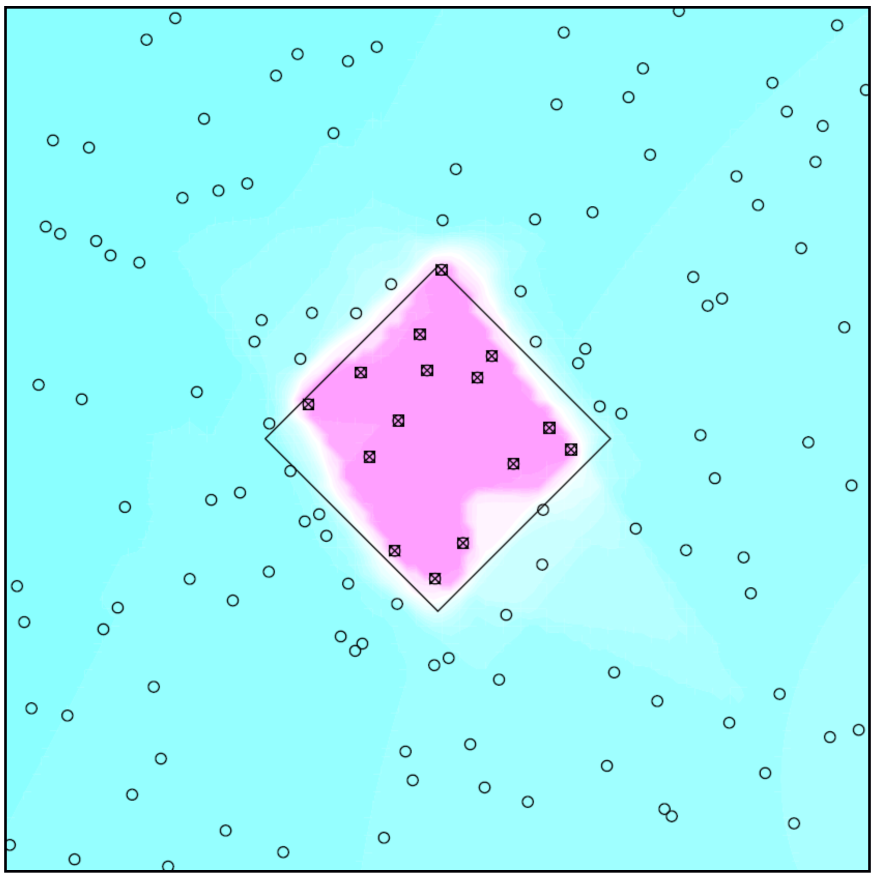


Toy example revisited

Second data set



Third data set



Modelling a Cloud Field

Cloud fields are a prime example of non-stationary behaviour in the natural world.

Modelling a Cloud Field

UNIVERSITY OF LEEDS

Exploring the sensitivity of cloud fraction to uncertainty in aerosol concentration.

Eddy/cloud resolving model (System for Atmospheric Modelling)

Grid mesh: $Dx = Dy = 200$ m, $Dz = 10$ m,

$Dt = 2$ s;

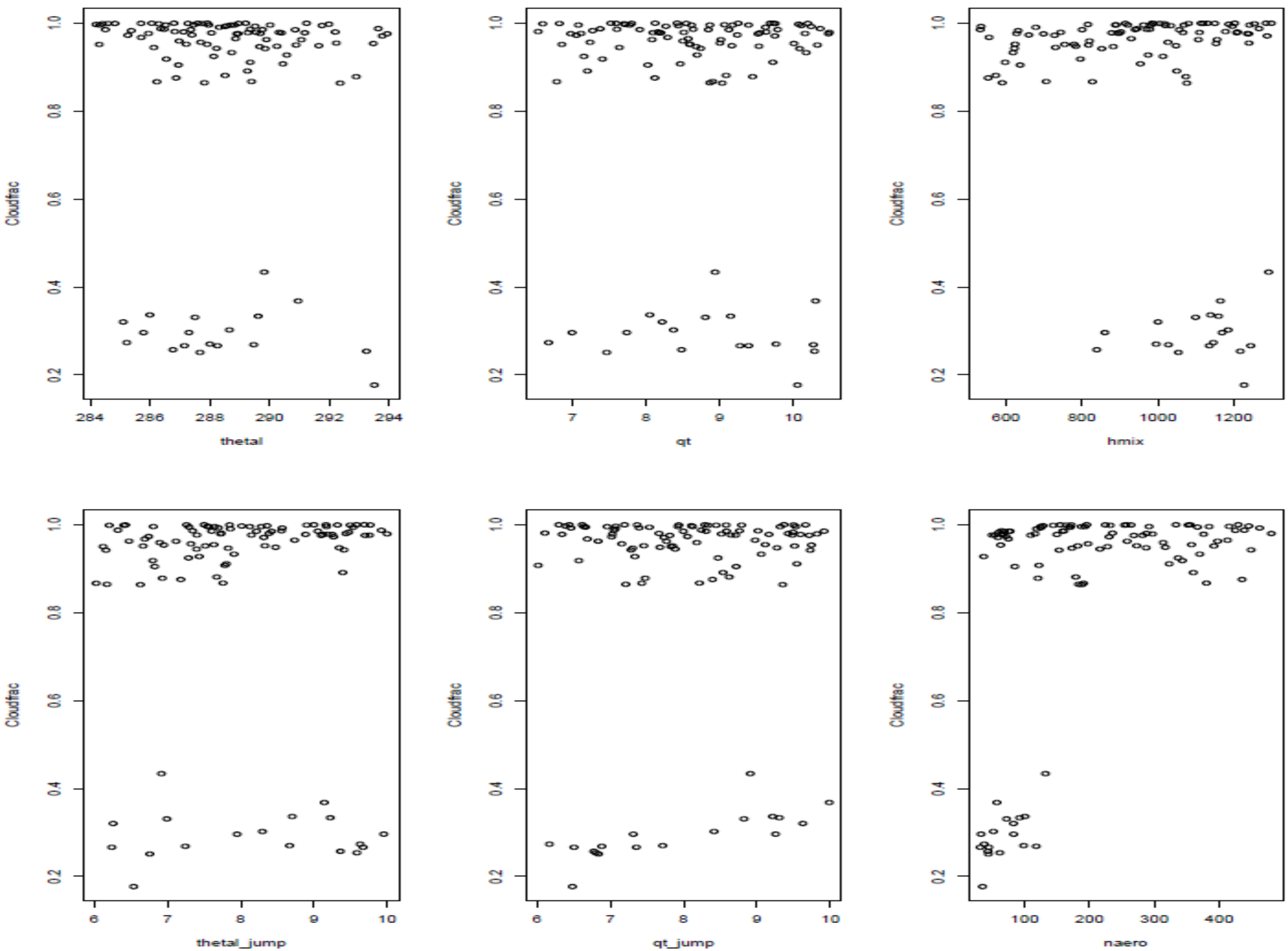
Domain size: 40 km x 40 km x 1.5 km.

The model is reasonably expensive: approx. **3 hours per run**, with 240 cores on a 760 Tflop Cray computer cluster.

Modelling a Cloud Field

UNIVERSITY OF LEEDS

We have run an ensemble of simulations according to a Latin hypercube design of size **105** over a **6d** parameter space.


Cloudfac
thetal
qt
hmix
thetal_jump
qt_jump
naero

Emulation results

The MAP model has two regions: one with 87 centres and the other with 18.

We have posterior probabilities of:

Pr(**1 region**) = 0.14, Pr(**2 regions**) = 0.66, Pr(**3 regions**) = 0.20.

We have 30 "test" runs of the cloud model.

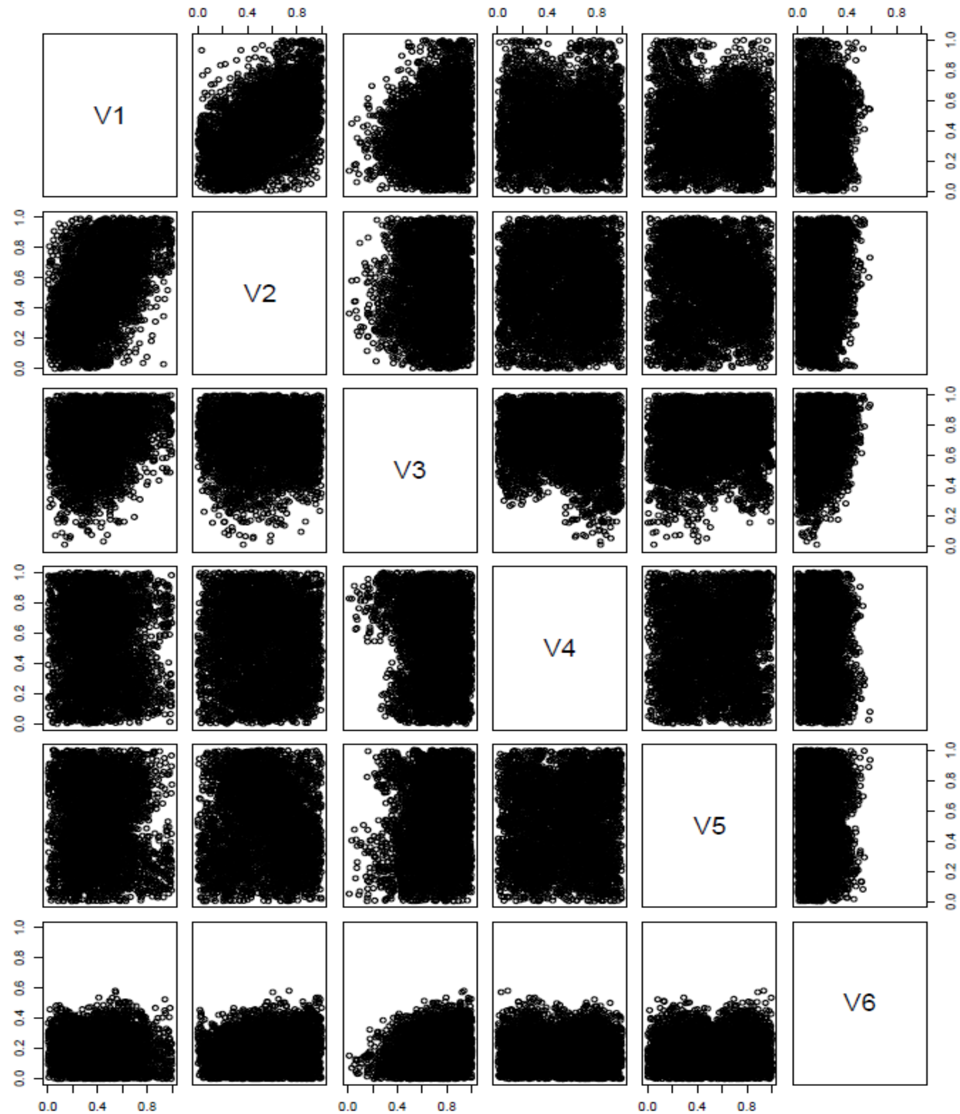
Method	MSE of prediction
Standard GP	0.028
Treed GP	0.032
Voronoi GP	0.016

We can also perform other standard emulator diagnostics.

Visualisation difficulties

UNIVERSITY OF LEEDS

Here are points that **lie on** the boundary between regions (based on the MAP estimate).



Visualisation difficulties

UNIVERSITY OF LEEDS

Each square in the picture gives the proportion of points that fall in region 1 when we consider the 4d grid of points for that particular x_i-x_j combination.

Darker -> higher proportion.

Cloud modelling next steps

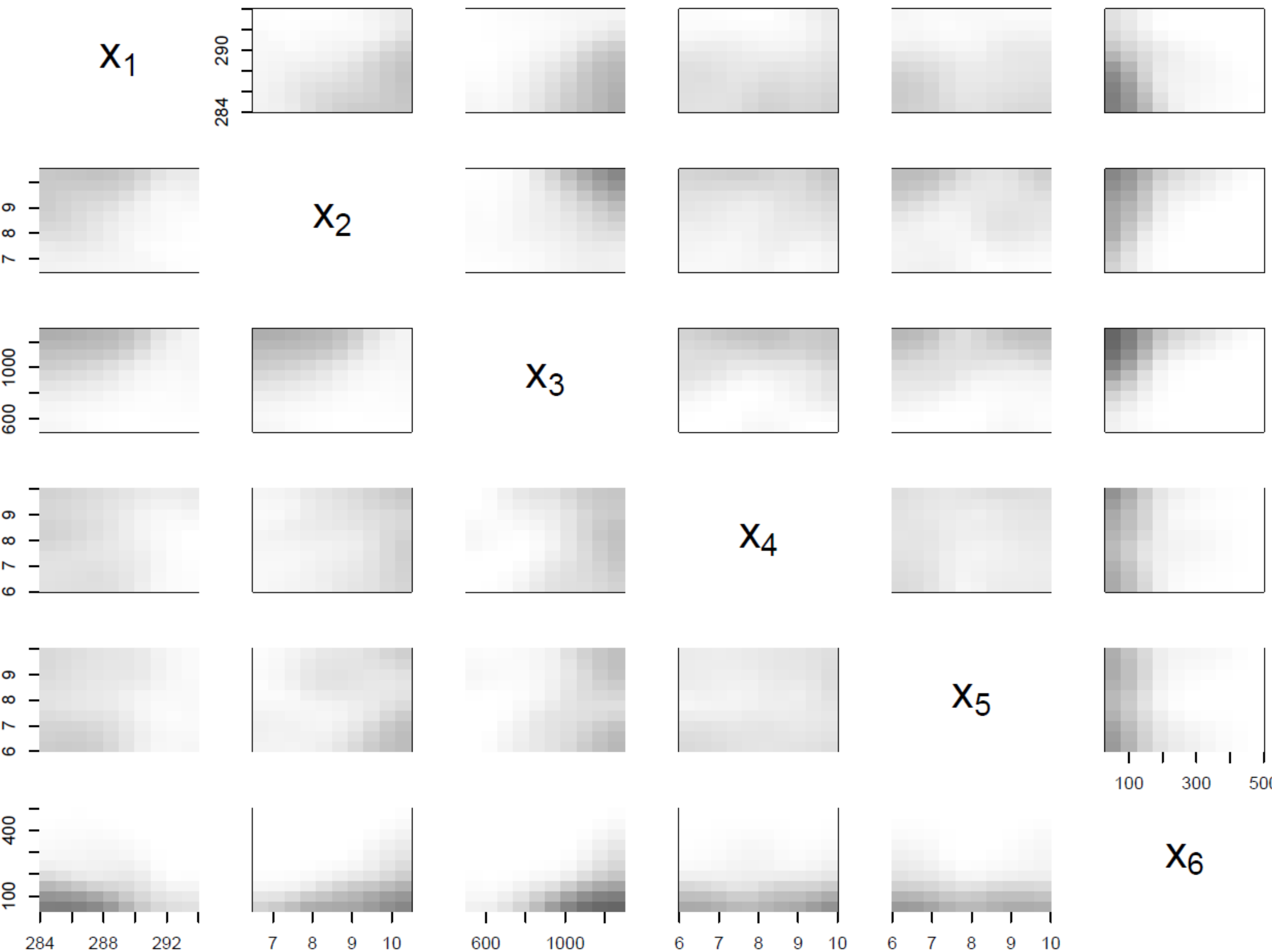
UNIVERSITY OF LEEDS

- We have rerun the analysis including the validation points and found a new MAP boundary.
- We have now passed 30 candidate points to the model owners to help us refine the region boundaries.

We need to think of a sensible way to describe this (potentially) 4d region to them…

Updated results

UNIVERSITY OF LEEDS



Each square in the picture gives the proportion of points that fall in region 1 when we consider the 4d grid of points for that particular x_i-x_j combination.

Darker -> higher proportion.

Possible extensions

UNIVERSITY OF LEEDS

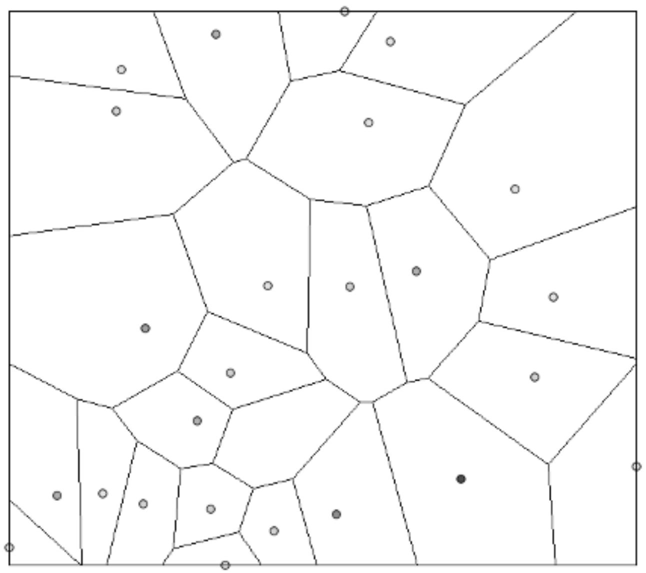
- Why stop at straight lines and convex regions?

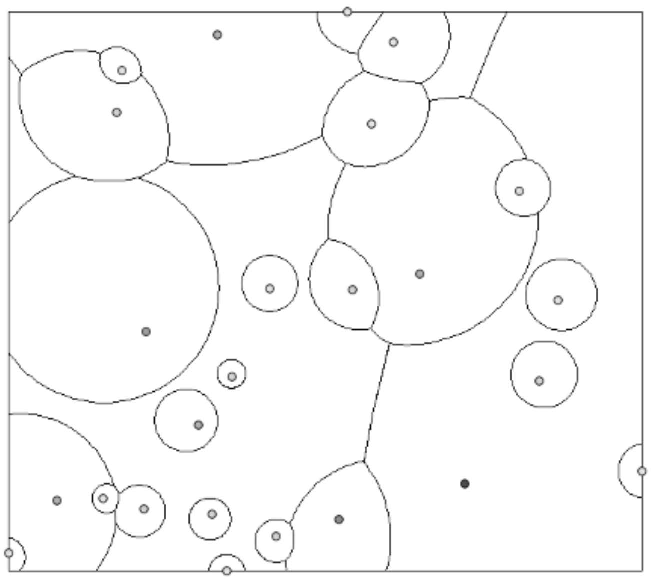
- Why stick to Gaussian processes?
- Our approach is related to k-nearest-neighbour classification and regression – ML methods for computations and visualisations?

Possible extensions

UNIVERSITY OF LEEDS

- Why stop at straight lines and convex regions?





Multiplicatively weighted Voronoi

- Why stick to Gaussian processes?
- Our approach is related to k-nearest-neighbour classification and regression – ML methods for computations and visualisations?

Possible extensions

UNIVERSITY OF LEEDS

- Why stop at straight lines and convex regions?

Standard Voronoi with city-block distance.

- Why stick to Gaussian processes?
- Our approach is related to k-nearest-neighbour classification and regression – ML methods for computations and visualisations?

References

UNIVERSITY OF LEEDS

Gramacy, R. B., & Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, *103*(483), 1119-1130.

Kim, H. M., Mallick, B. K., & Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, *100*(470), 653-668.

Pope, C. *et al.* (2017). Modelling spatial heterogeneity and discontinuities with Voronoi tessellations. *In preparation as it has been since 2007…*