Modeling and Monitoring Crop Disease in Developing Countries

John Quinn[1], Kevin Leyton-Brown[2], Ernest Mwebaze[1]

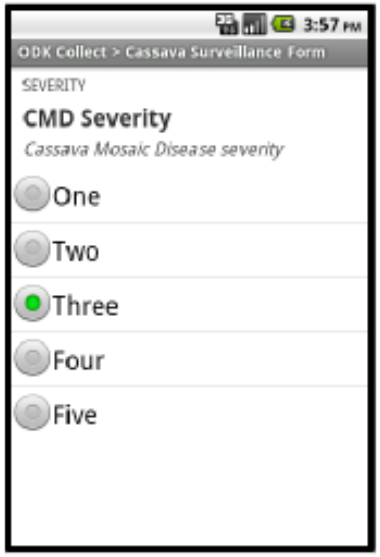
[1]Department of Computer Science
Makerere University, Uganda

[2]Department of Computer Science
University of British Columbia

AAAI 2011

Information about infectious crop diseases is vital in developing countries. However, current survey methods are expensive, limited, and slow: 2 months survey time + 3 months of data entry.

Ubiquity of mobile telephony in the developing world has created opportunities for better systems...


3:57 PM
ODK Collect > Cassava Surveillance Form
SEVERITY
CMD Severity
Cassava Mosaic Disease severity
One
Two
Three
Four
Five


3:50 PM
ODK Collect > Cassava Surveillance Form
PICTURE
Main Picture
Please take a snap shot of the cassava leaf
Take Picture


3:58 PM
ODK Collect > Cassava Surveillance Form
GEO Location
gps
Geographical Position System
Record Location


Date:Sunday March 27 2011 6:34:46 PM UTC
Disease Rate:CMD:4,CGM:5,CBS:1
GPS Accurracy:6.0

Outline of talk

Three ways in which the utility of the survey can be further improved with a fixed budget:

- Improvements in spatial disease estimation given observations.
- Adapting the locations to which surveyors should be sent, given real-time survey updates.
- Using computer vision to automatically diagnosis diseases given camera-phone input.

Density modelling problem

- In a crop disease survey, each plant is assigned a disease level $y_i \in \{d_1, ..., d_D\}$.
- The observed data is of the form $\mathscr{D} = \{x_i, y_i | i = 1, \ldots, N\}$ where $x_i \in \mathbb{R}^2$ is a spatial location.
- We would like to infer $P(y^* | x^*, \mathscr{D})$.
  - Can then calculate things like *incidence*, the proportion of infected plants: $P(y^* > d_1 | x^*)$.
- We should deal with the observations y_i as ordinal categories, $d_1 < d_2 < \cdots < d_D$, and not code them as numbers: we do not have a principled way of mapping these categories to the real line, for example.
- If possible, we should deal with observations in continuous space, rather than losing information through aggregation.

Survey utility

Our goal is to come up with accurate estimates of the areas we have not sampled directly. We have three principles, elicited from our collaborators:

1. Small errors are better than big errors.
2. It is more important to assess whether an area is diseased or not, than to accurately guess the specific severity.
3. Being wrongly optimistic—under-predicting the incidence of disease—is about twice as bad as being wrongly pessimistic.

Survey utility

Our utility function, for guessing $\hat{y}$ given truth y is

$$u_y(\hat{y}) = -\mathit{OptimismFactor}(y,\hat{y}) \cdot (\mathit{Error}(y,\hat{y}) + \mathit{DiseasePenalty}(y,\hat{y}))$$

where

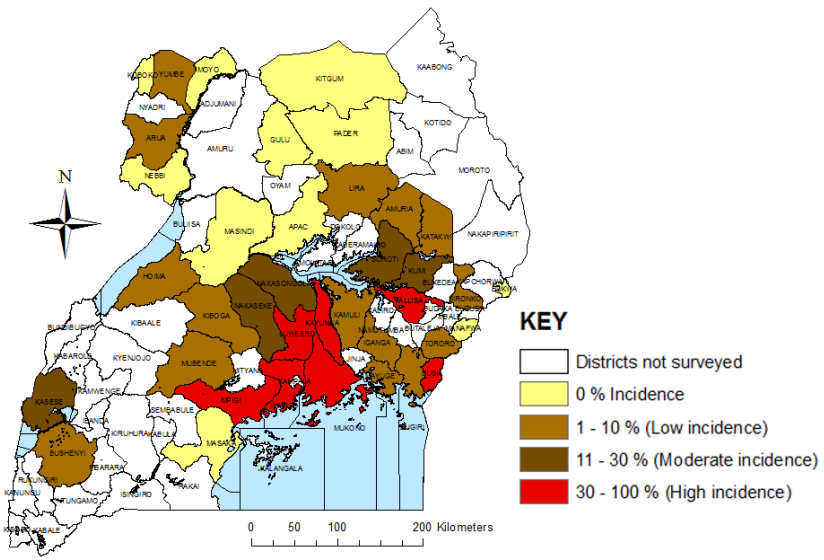
$$\mathit{Error}(y,\hat{y}) = \frac{|y-\hat{y}|}{D-1};$$

$$\mathit{DiseasePenalty}(y,\hat{y}) = \begin{cases} \alpha & y = 1 \text{ XOR } \hat{y} = 1 \\ 0 & \text{otherwise;} \end{cases}$$

$$\mathit{OptimismFactor}(y,\hat{y}) = \begin{cases} \beta & y > \hat{y} \\ 1 & \text{otherwise.} \end{cases}$$

Current spatial density model

Example of current type of map generated by survey teams (cassava brown streak disease, 2009):



Densities are aggregated at district level. Two statistics are calculated: *incidence* and *severity*.

Other approaches in the literature: kriging (non-ordinal), MRFs (discretized space).

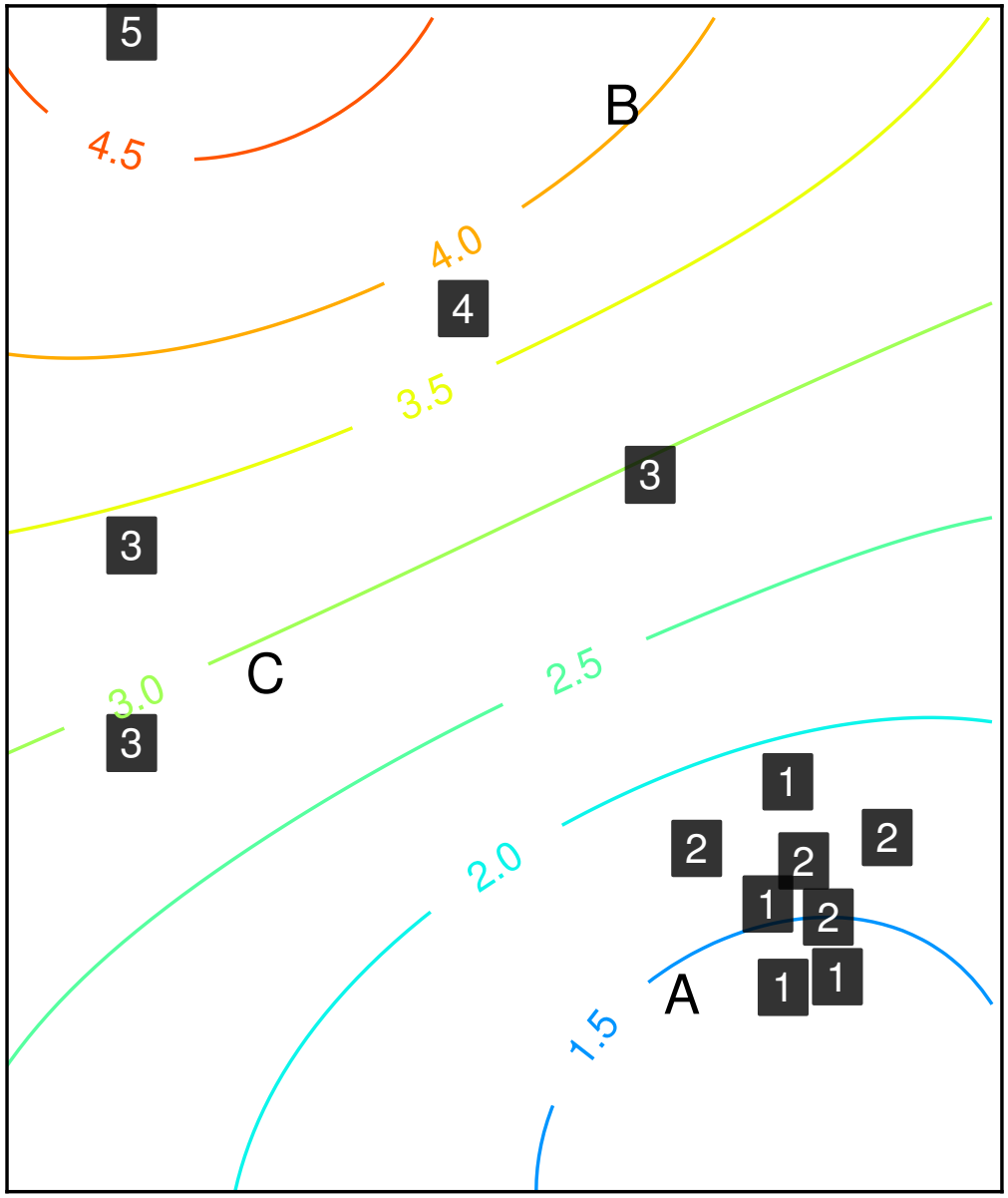
Gaussian process ordinal regression

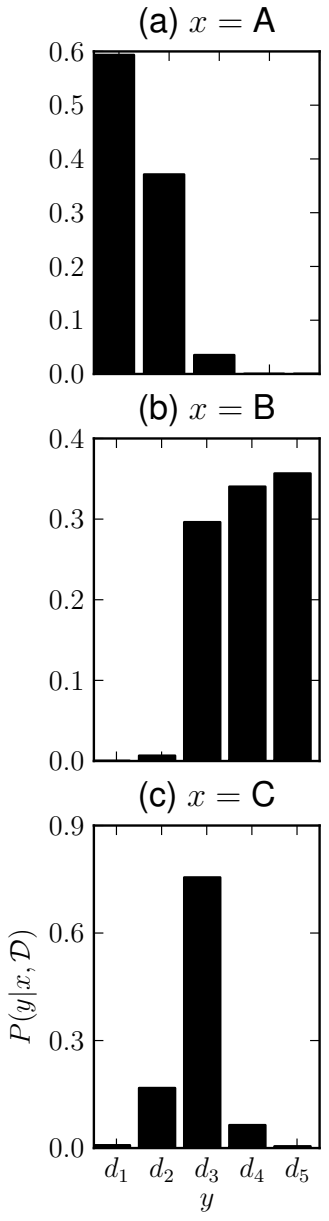
- We can use GP regression with ordinal observed data to model density in continuous space.
- First, define a hidden function of the location $f(x_i)$, and specify a covariance between the function at different positions:

$$K(x_i, x_j) = \exp -\frac{\kappa}{2}\left(\left(x_{i,1} - x_{j,1}\right)^2 + \left(x_{i,2} - x_{j,2}\right)^2\right)$$

- Then relate the observations with this hidden function with a likelihood term: $p(y_i|f(x_i))$.
- We then integrate out $f(x_i)$ to obtain a distribution across ordinal categories at each location.

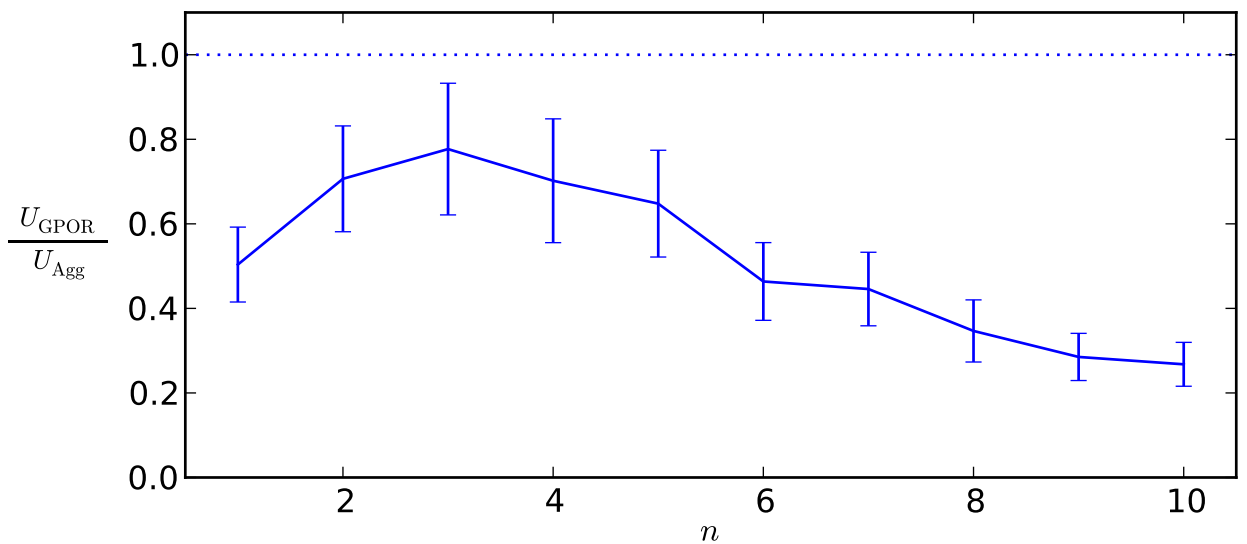
GPOR inference example





Comparison of GPOR with aggregation

A comparison of the utility of GPOR and aggregate density models at different levels of aggregation (an $n \times n$ grid):



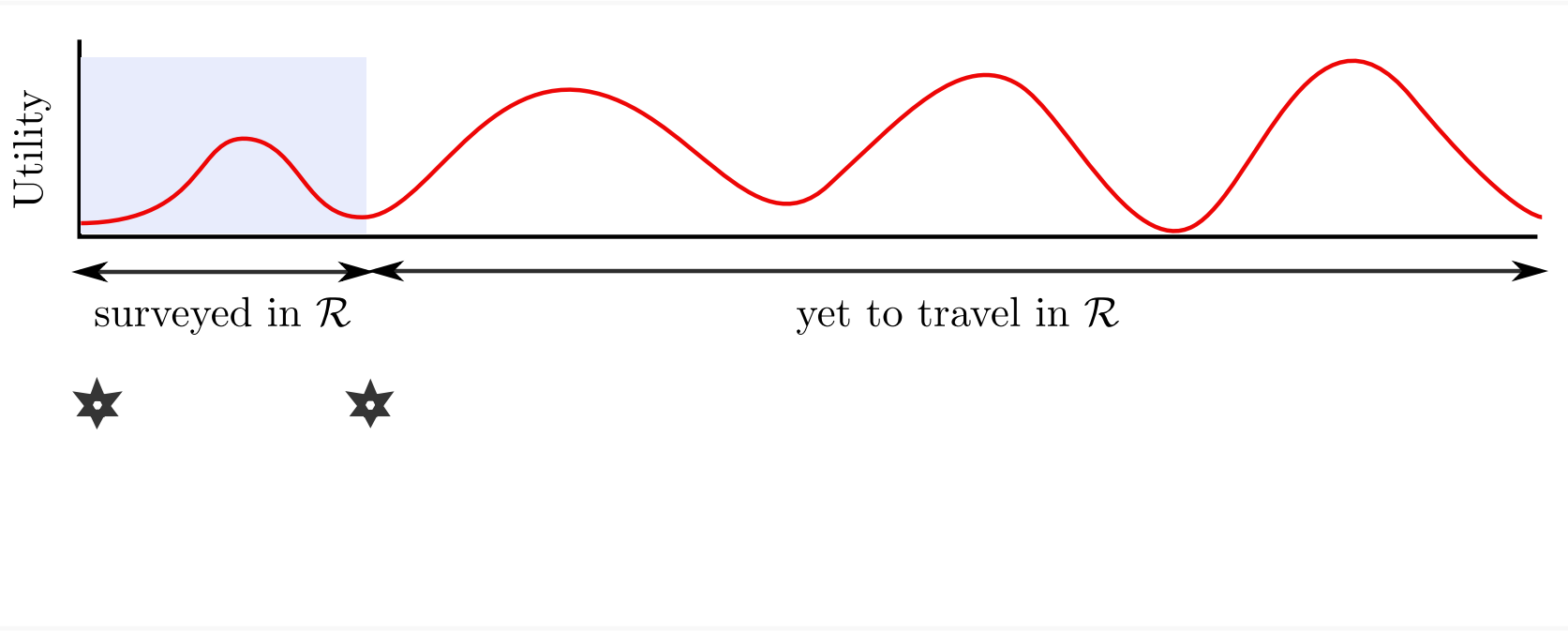
A loss ratio of less than 1 means that the GP model outperformed the aggregation model.

Survey site optimization

- Given real time updates using mobile phones as data collection tools (more about this later), we can dynamically change the survey plan according to which regions are most interesting.
- In the usual active learning setting, this is straightforward: go to the places where the estimate is most uncertain.
- We have an online variation of this problem: surveyors are on a road circuit (represented as a 1D manifold $\mathscr{R}$), and we keep moving in the same direction.
- Hence we will often prefer to sample a nearby point before sampling a distant point, even if the latter is expected to be more informative.
- Currently, the survey plan is to sample at roughly equal intervals along $\mathscr{R}$.

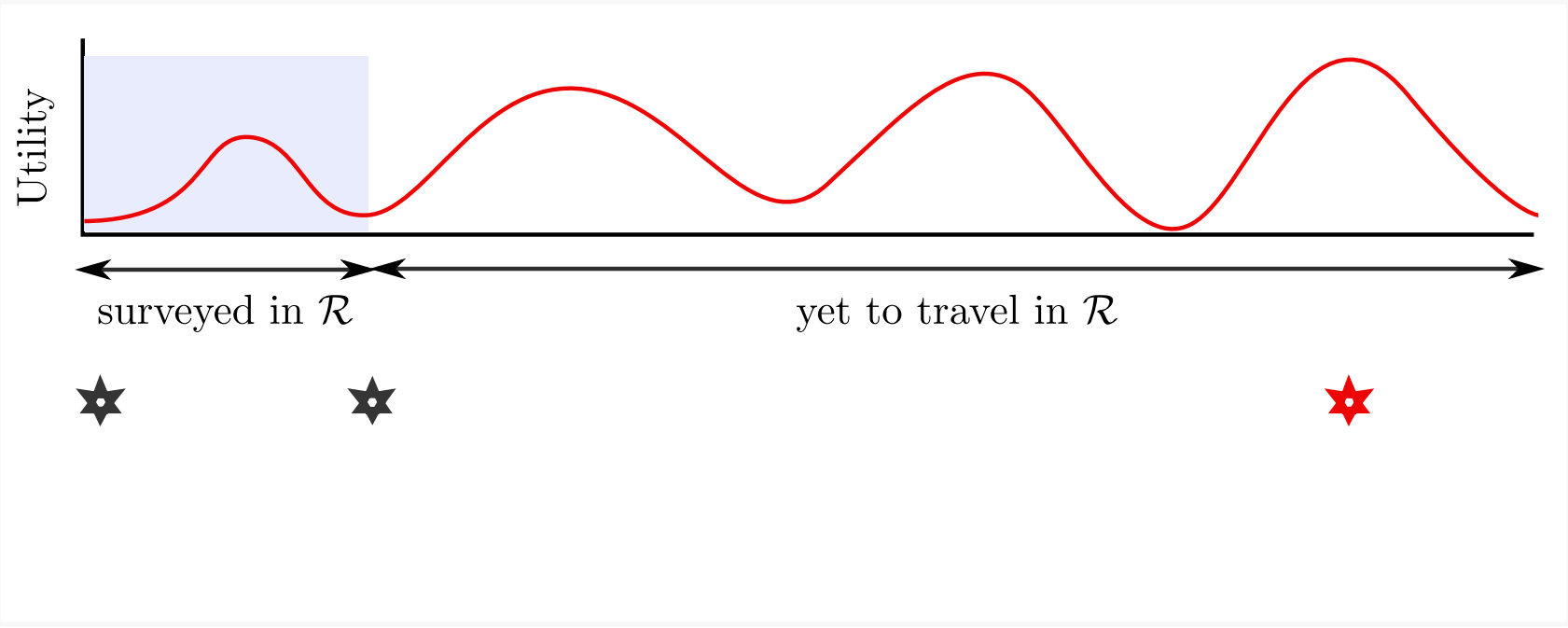
Optimization method

For example, start with two samples having been made along the route. Assume we have $k = 3$ stops still to make and want to calculate the next stop.



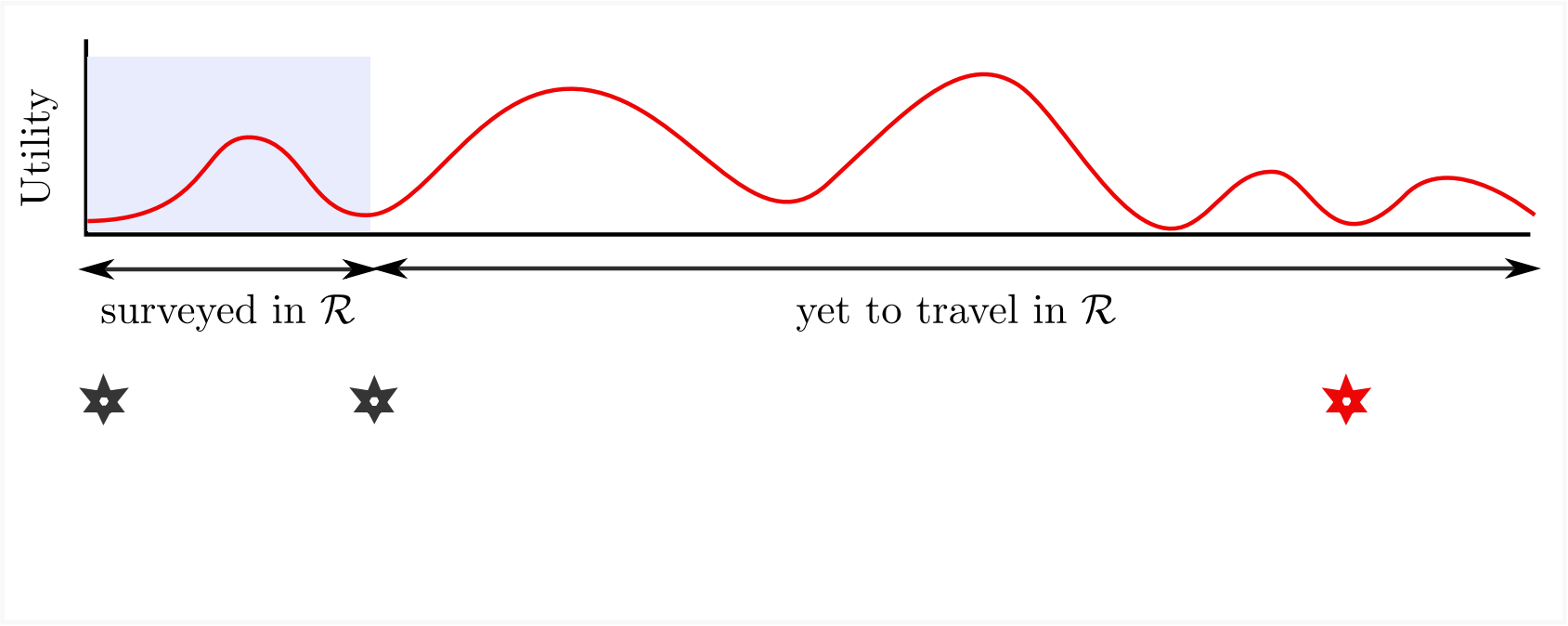
Optimization method

Greedily select the point along the remaining route with the largest expected utility



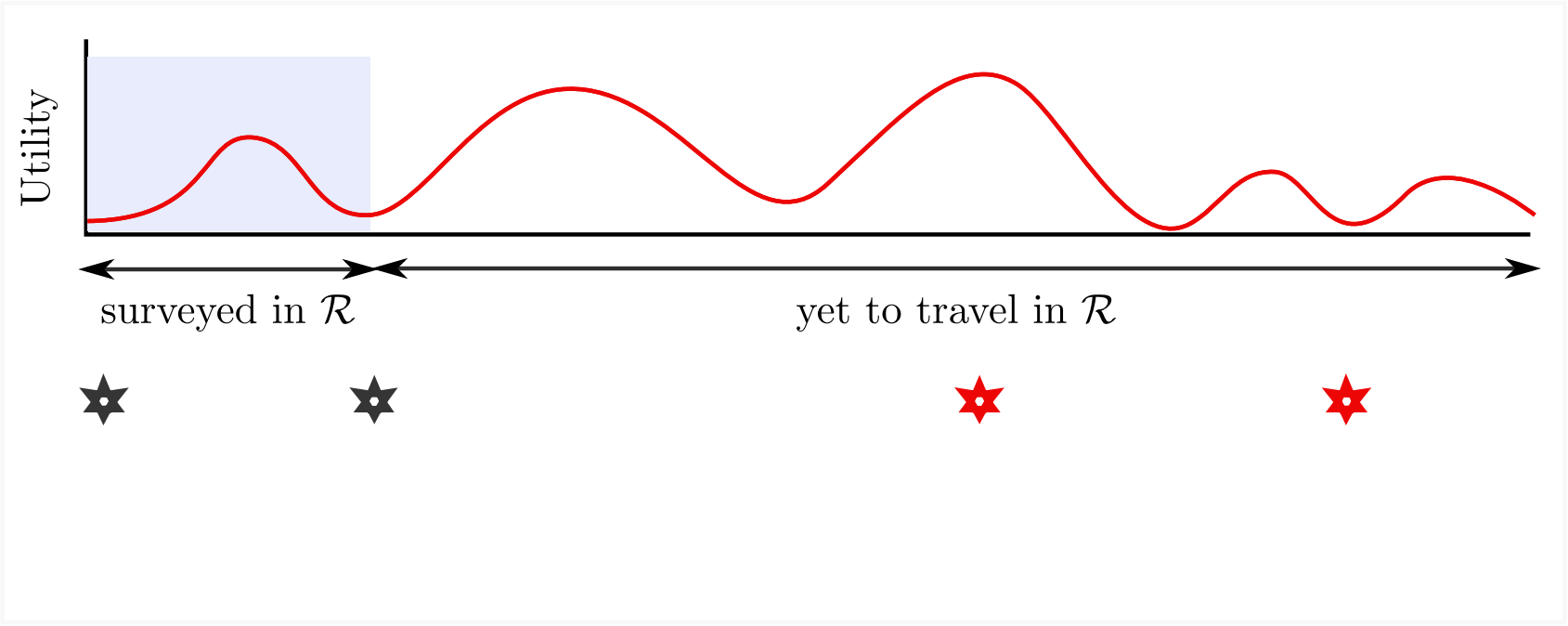
Optimization method

Draw samples from the model. Update the utility at each point on $\mathscr{R}$



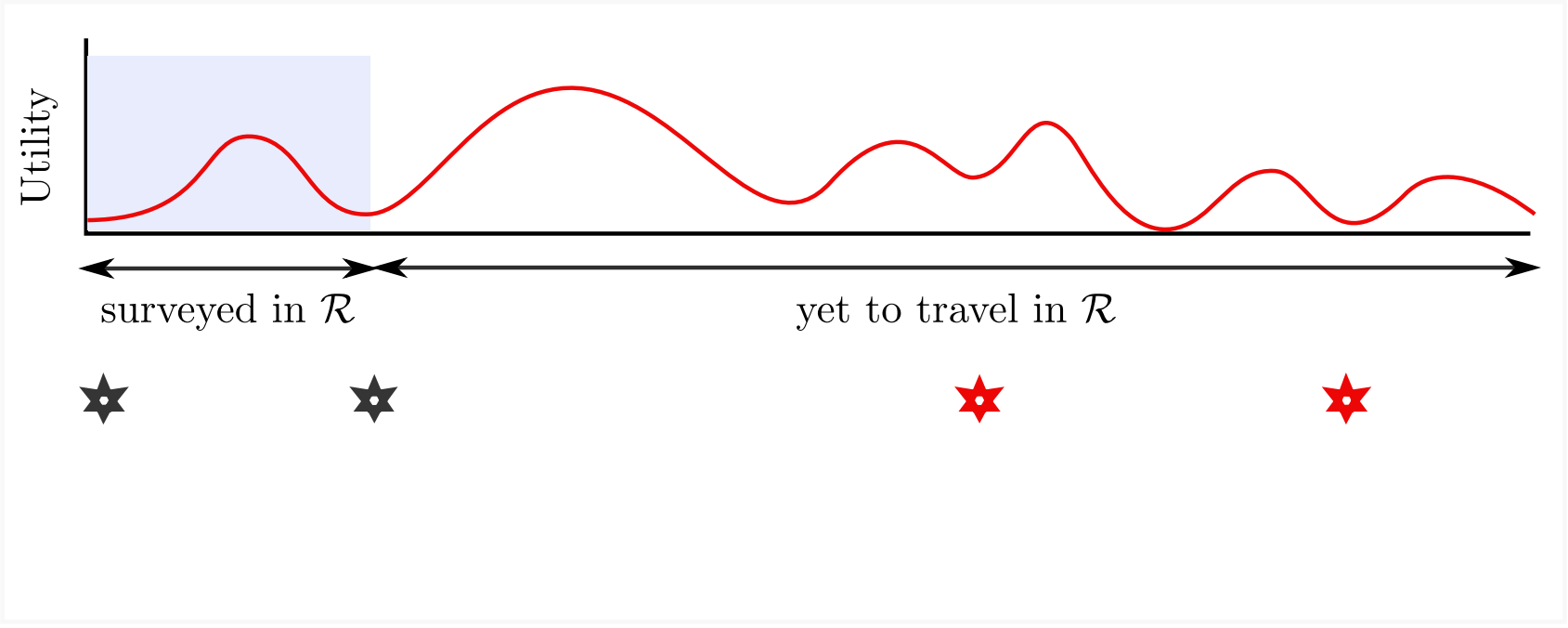
Optimization method

Greedily select the next point



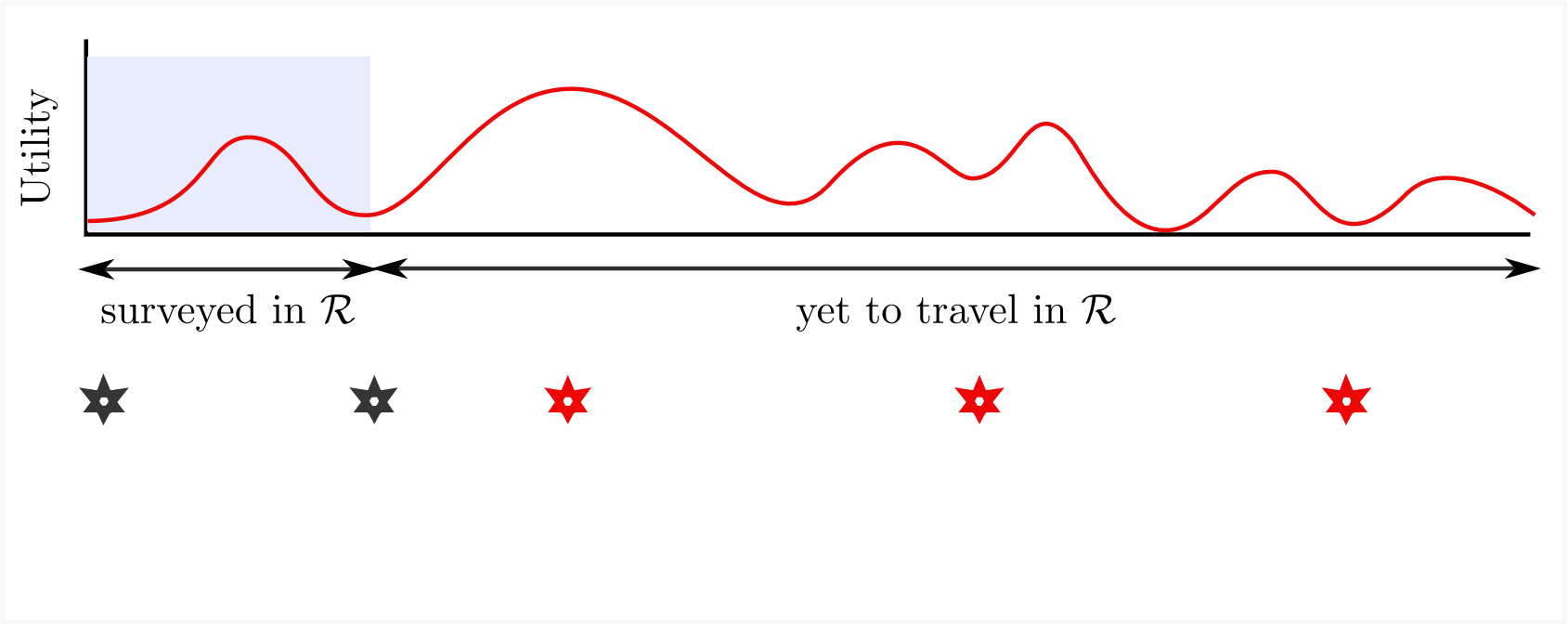
Optimization method

Draw samples from the model. Update the utility at each point on $\mathscr{R}$



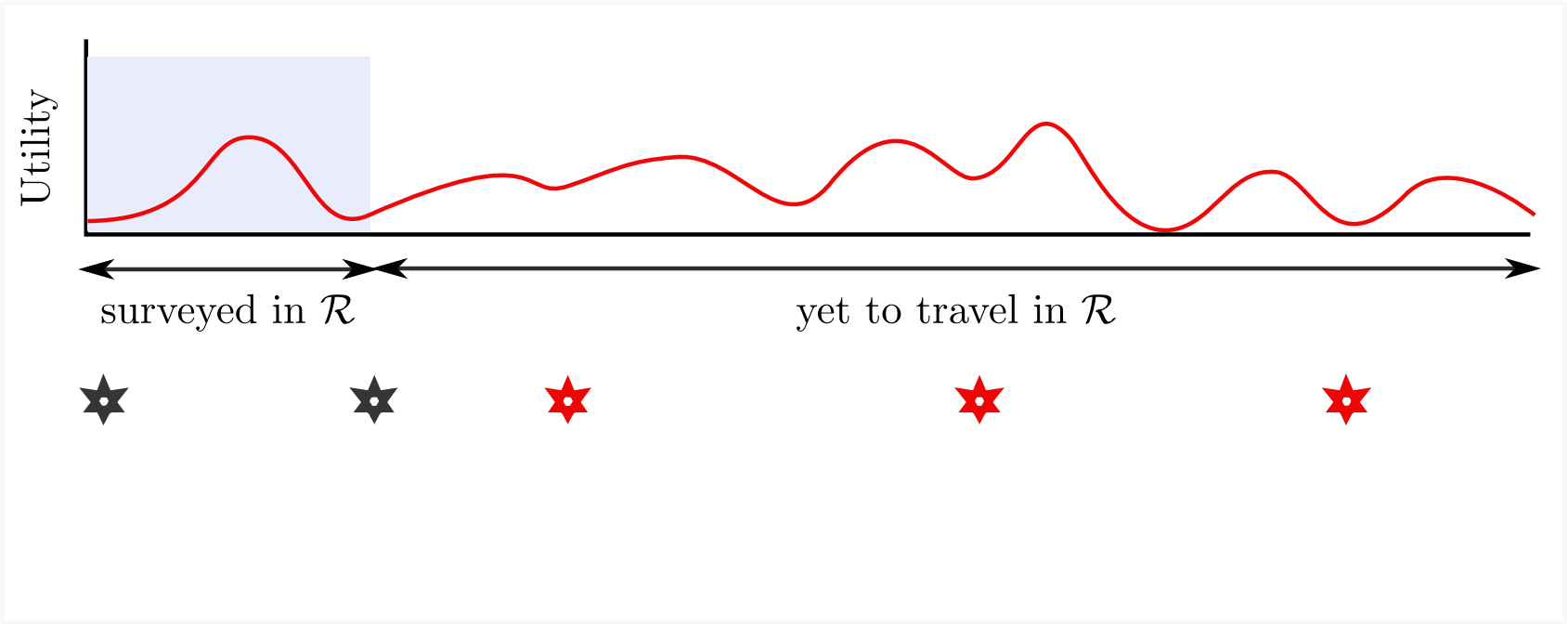
Optimization method

Greedily select the next point (last of the k points).



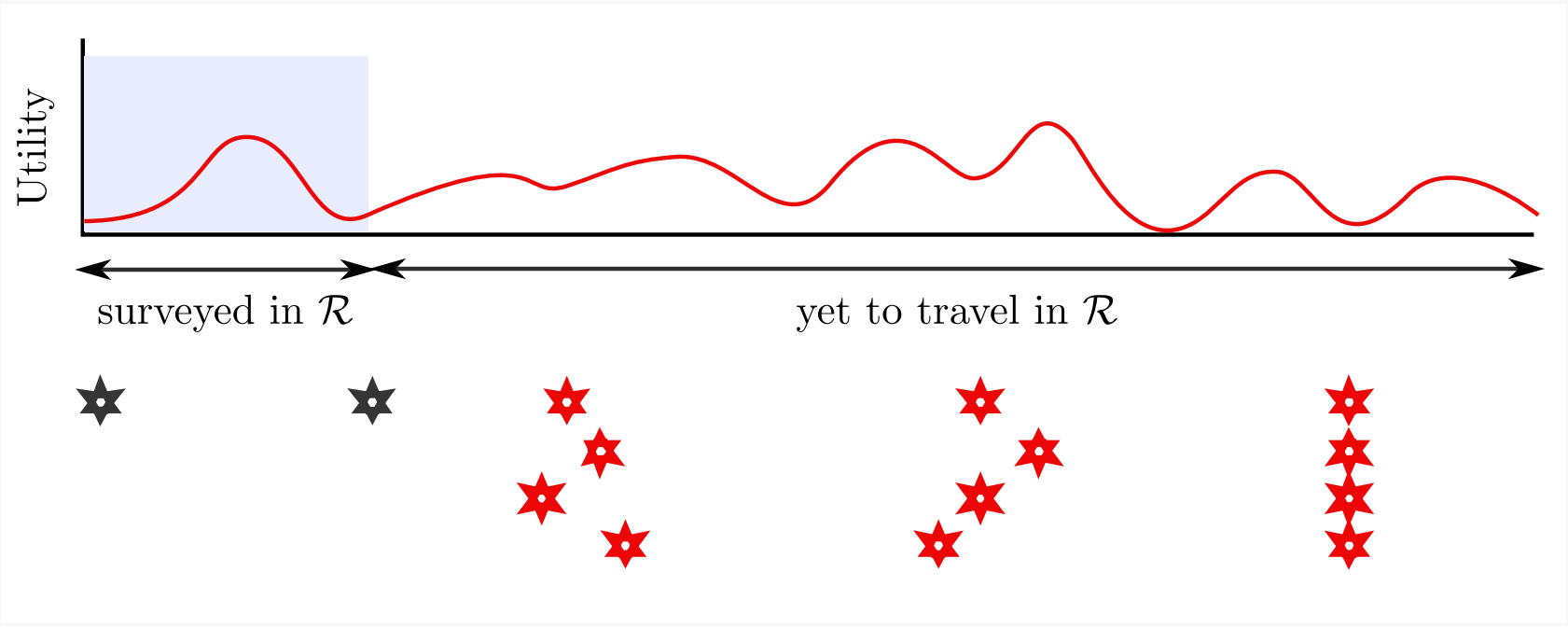
Optimization method

Draw samples from the model. Update the utility at each point on $\mathscr{R}$



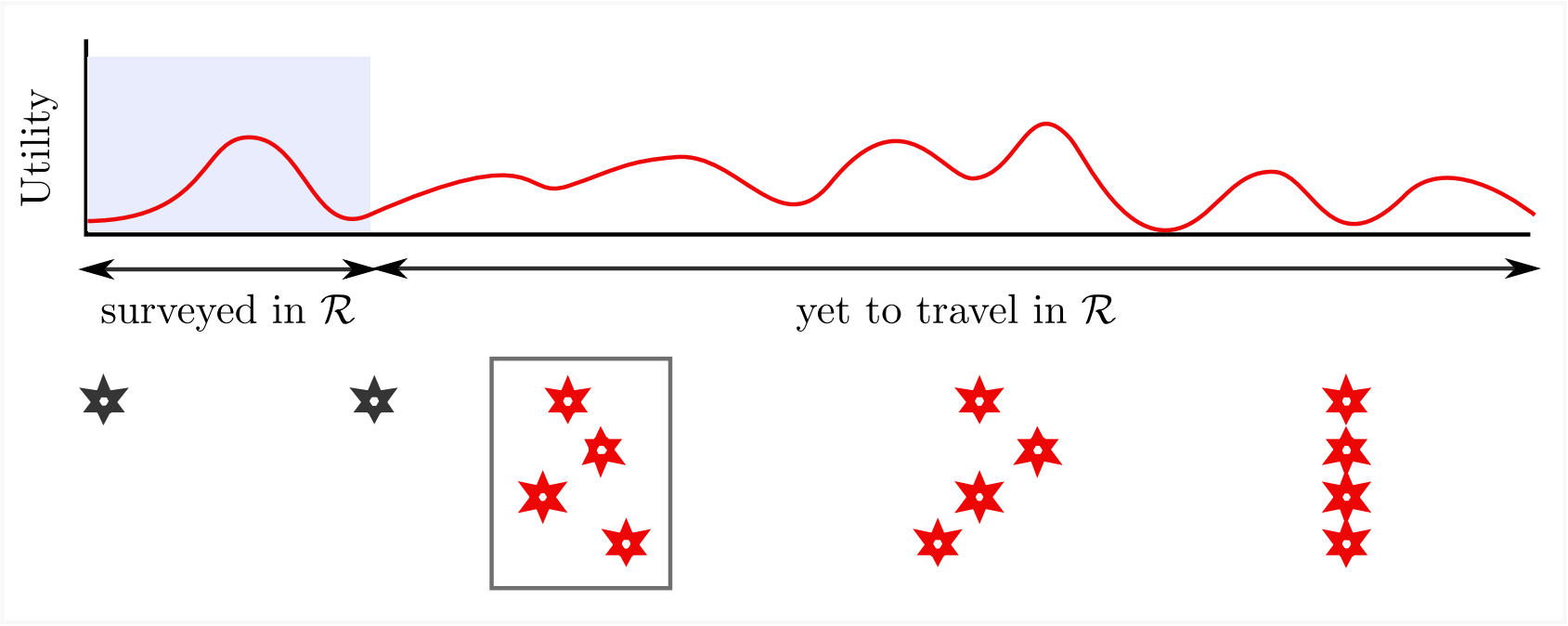
Optimization method

Find multiple sets of survey stops in this way.



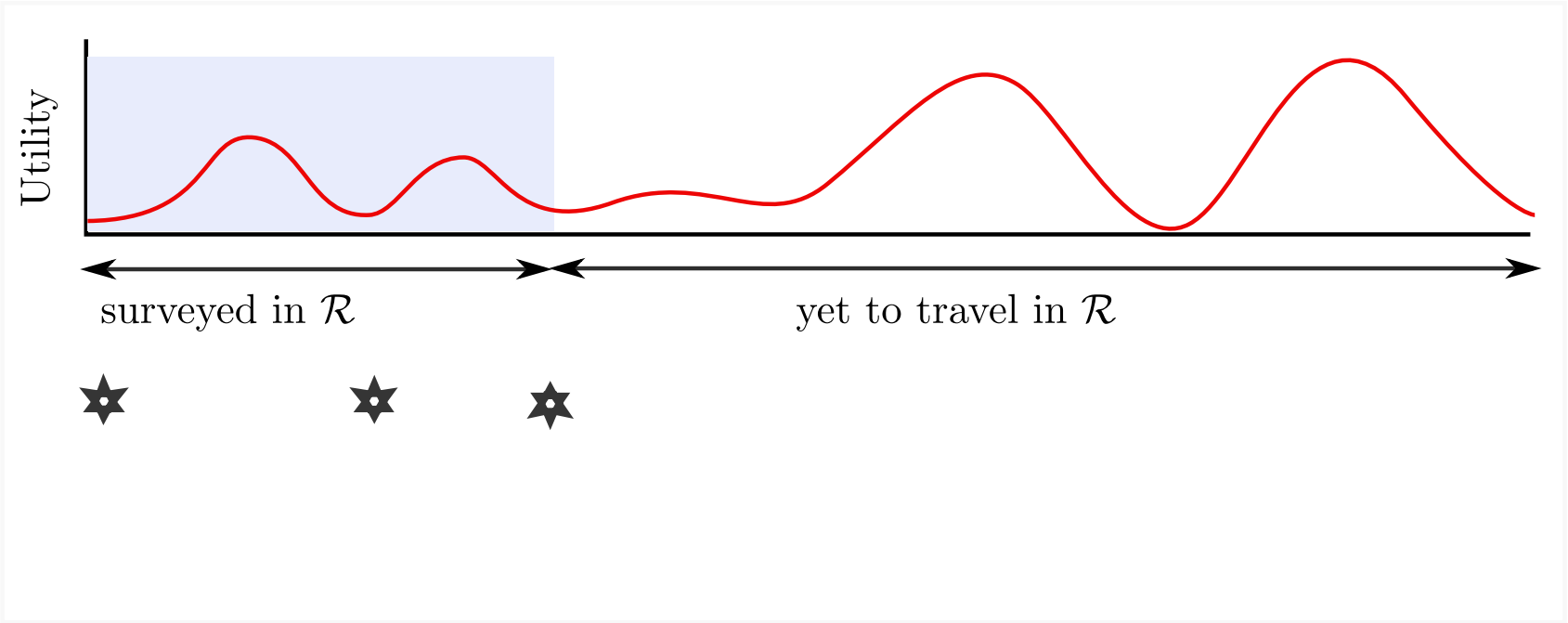
Optimization method

Consider the set of nearest survey locations



Optimization method

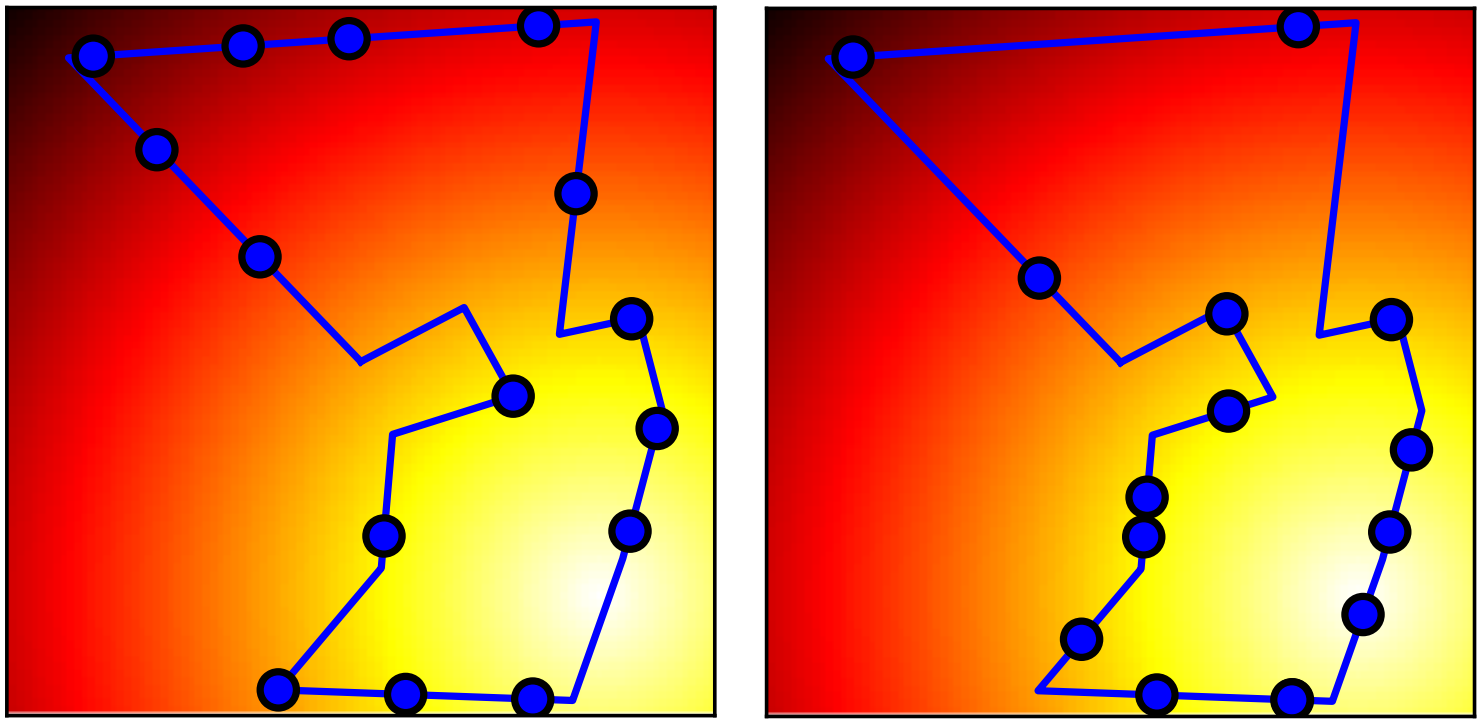
The first order statistic becomes our next survey stop.



Adapting survey locations

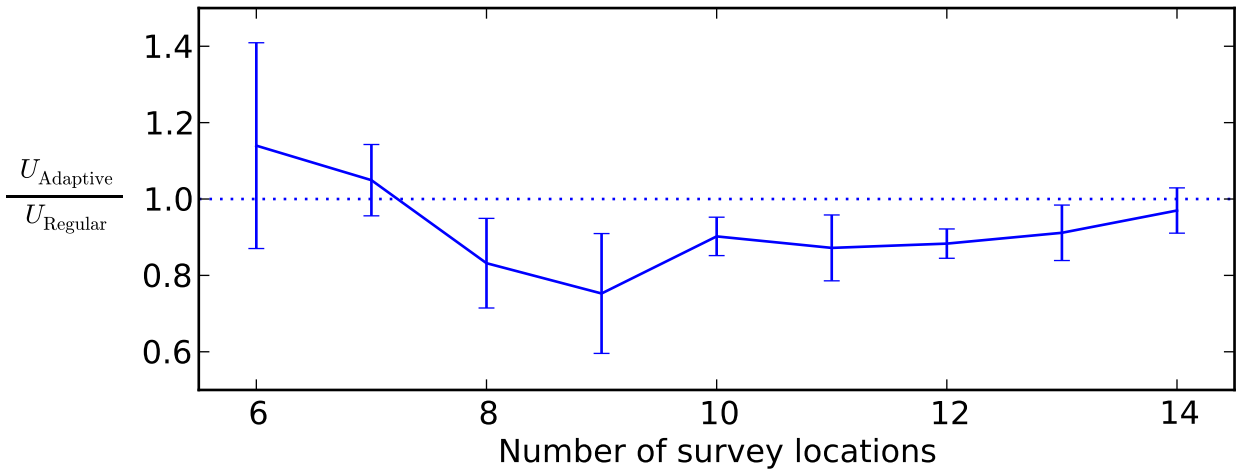
Left: regular sampling (heatmap shows true incidence on both panels).

Right: optimised sampling; survey locations are moved towards regions of higher density.



Comparing regular and adaptive sampling

Comparison of the average utility of simulated survey results, for regular and optimised sampling positions, as a function of the sample budget:



Beyond the smallest number of survey locations (sampling too sparse for adaptive method to be effective), and high number of locations (in which every location is thoroughly sampled) the mean utility ratio $\frac{L_{\text{Adaptive}}}{L_{\text{Regular}}}$ is significantly below 1.

Image-based classification of disease

- One of the constraints on crop disease surveying currently is the scarcity of experts.
- When mobile data collection replaces paper-based surveying, camera-phones can be used to collect image data.
- If we can use these images to perform a diagnosis of disease automatically, then agricultural extension workers with basic training can carry out the survey.

Leaf image examples

Camera phone images of plants which are healthy (left) and infected with cassava mosaic disease (right).

We extract two types of features from these images: hue histograms, and SIFT descriptors.

Classification

We classify using k-NN, where our distance is defined as an average of distances between the two types of features.

For hue histograms, the KL divergence is given by

$$D_{KL}(h_1, h_2) = \sum_i h_{1,i} \log \frac{h_{1,i}}{h_{2,i}}.$$

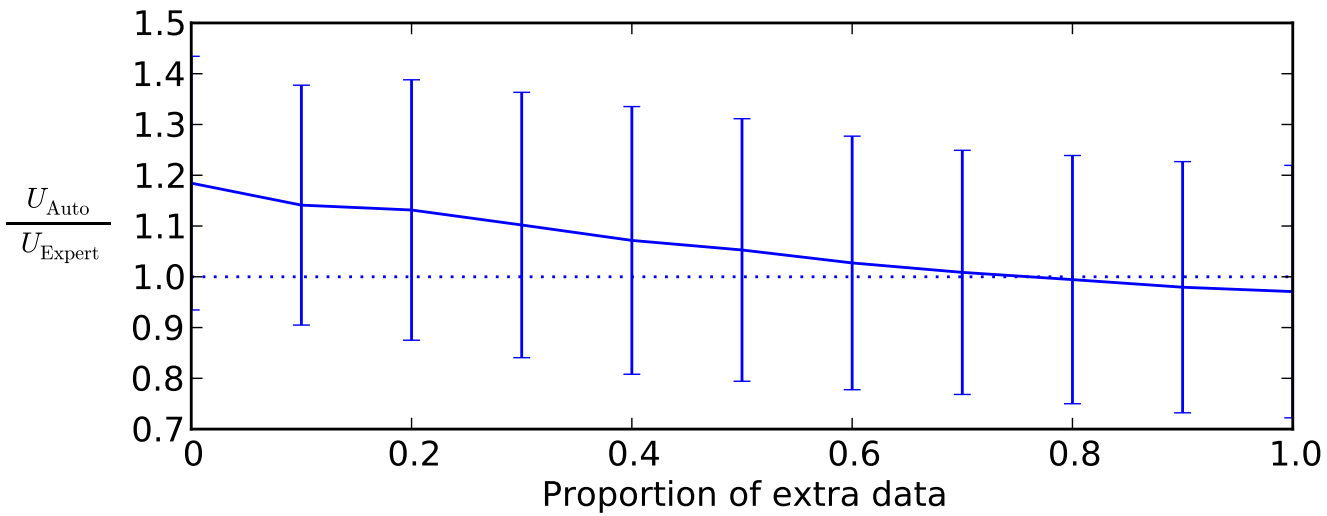
For SIFT features, we calculate a mean μ and covariance Σ for each image. The KL divergence for multivariate Gaussian distributions is

$$D_{KL}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \frac{1}{2}\text{trace}\Big\{\Sigma_1\Sigma_2^{-1} + \Sigma_2\Sigma_1^{-1} - 2I + \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top\Big\}.$$

Utility comparison

Best classifier: AUC=0.961, error rate=0.078.

We compared the utiliy in our simulations with “expert” and “automated” labellings (in the latter, a random 8% of labels are wrong), where different quantities of data are available to the automated system.



The automated scheme has equivalent utility when at least 70% extra data is obtained.

Smartphone implementation

Feature extraction and classification implemented on $100 Android phone:

60,000 of these phones were sold in Kenya this year: significant potential for crowdsourcing.

Summary

- We have introduced a density model, survey optimisation method, and image-based disease classification method.
- These techniques can bring significant savings in both survey cost and time.
- This also increases in the amount of data which can be collected: extension workers can be issued with phones and provide data year round.
- We are planning national survey of cassava disease in Uganda this year.

Maps, data, code:
http://cropmonitoring.appspot.com

Namulonge field station