Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Bayesian approach to modelling plant pest introduction

Jonathan Yuen

Swedish University of Agricultural Sciences, SE 75007 Uppsala, Sweden

Bayesian analysis

Thomas Bayes (1702-1761)

Tomato Disease Example

- An MSc student was sent to study other diseases on virus-resistant tomatoes in Nicaragua
- Was told by the scientists in Nicaragua that the diseases were early blight and fusarium wilt
- Decision : What agar media do you send with the student
- Depends on the person making the decision and what information they might have about tomato diseases in the tropics

Non-Bayesian approach (Frequentist)

- Basic principle:

  parameters as fixed but unknown quantities

  probability as long-run frequency

- The true mean of the population exists and it is a number
- If you sample again and again you will find something that is based on it!
- CI_{95} contains true value 95 out of 100 times

Non-Bayesian approach

Assumptions and estimation for **ANOVA**

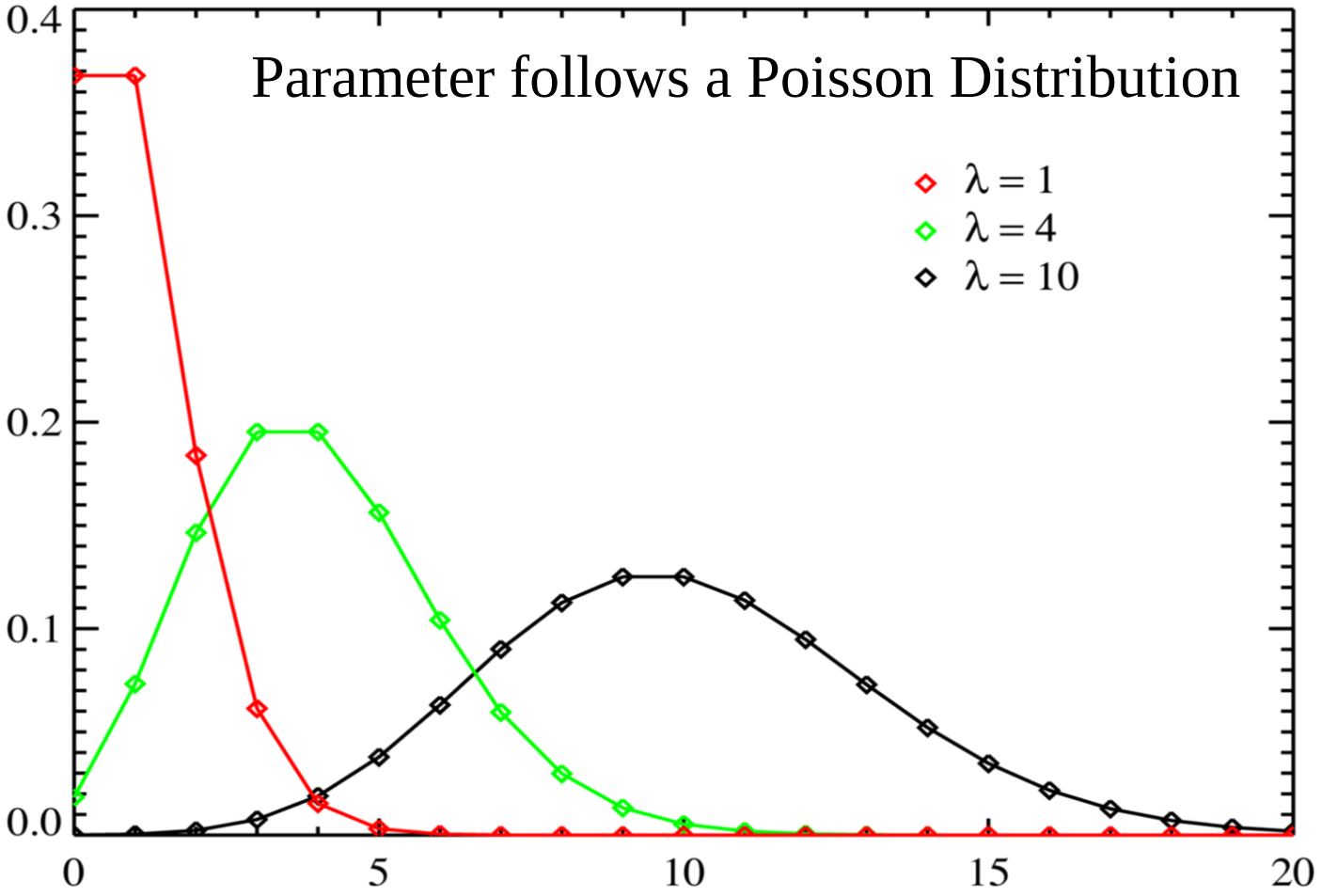
- Unknown but fixed effects contribute to the underlying value for each of the cells
- Normal distribution of deviations from these unknown underlying values and the observed values
- ML estimate is the same as a least-squares estimate

Bayesian inference: Key ideas

- parameters as random variables
- probability used to describe uncertainty about unknowns
- combine information (prior distributions and current data) to draw conclusions

Bayesian methods

Describe uncertainty with probability distributions



Some important terms
(or how to talk like a Bayesian)

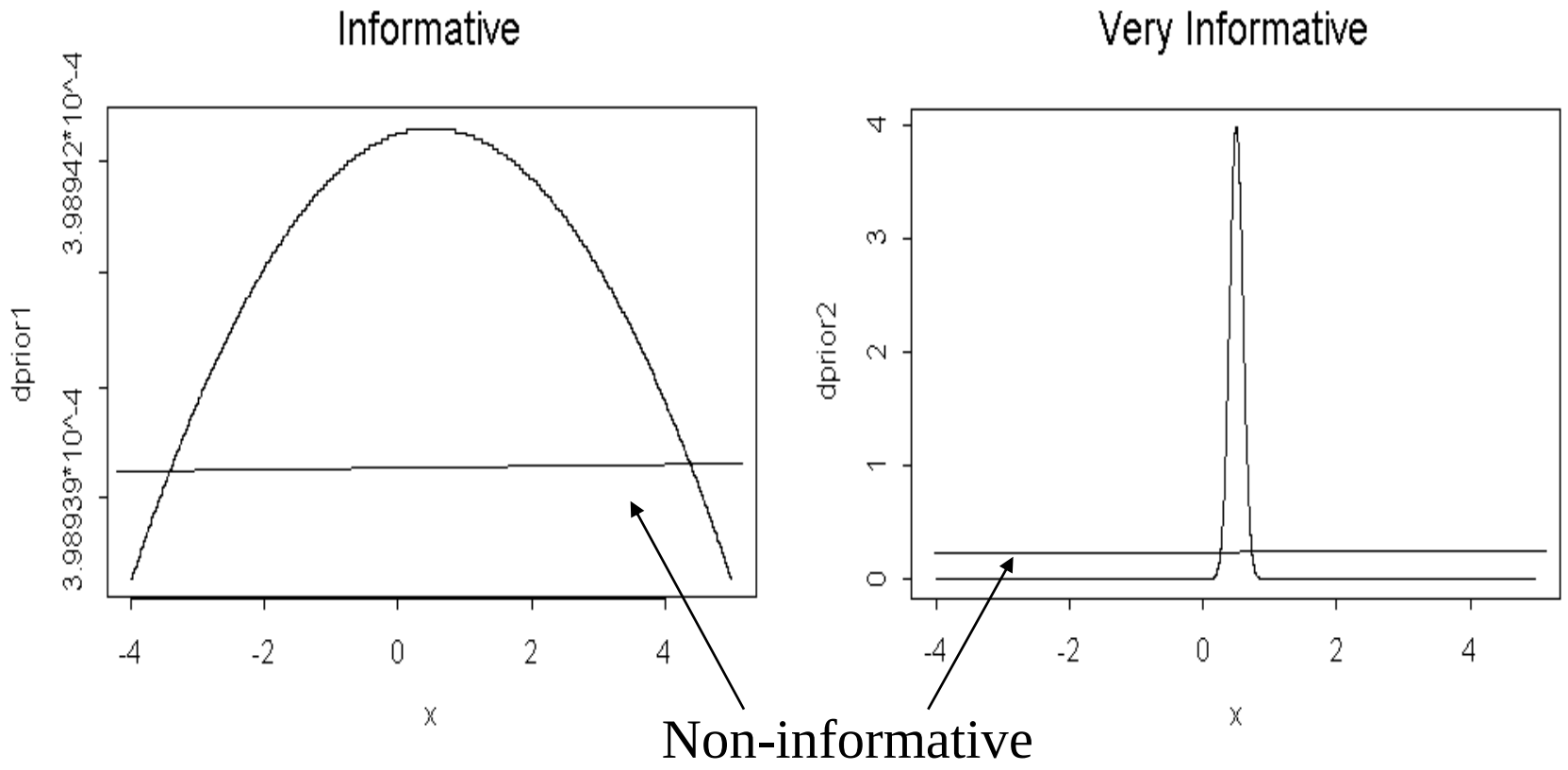
- Prior – idea before collection new information
- New Data – likelihood
- Updated knowledge – Also called posterior
- Credibillity interval– two limits derived from the posterior distribution, and calculated such that a pre-determined area under the distribution (95% for example) is between the two limits. Sometimes called a Bayesian confidence interval, but should not be confused with a frequentist confidence interval.

Tomato Example

- Prior – varies from person to person.
    - What are the common diseases of tomatoes in the tropics?
- Current data – being told that the diseases were fusarium wilt and early blight
- Updated knowledge from combining data and prior knowledge
- Student's prior was relatively uninformed and PDA was sufficient
- Other priors lead to both PDA and TZC
    - My prior is that bacterial wilt (caused by *Ralstonia solanacearum*) is common in the tropics. I sent a bottle of TZC with the student (and she used it to confirm the presence of *Ralstonia solanacearum.*

Bayesian methods

Distributions to describe prior knowledge/belief



Bayes's Theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})}$$

Bayesian inference

- Posterior inference

  - Bayes' theorem to derive posterior distribution

$$p(\theta \mid y) \propto p(y \mid \theta)\ p(\theta)$$

Posterior

Distribution

(Updated knowledge)

Likelihood

(Collected data)

Prior distribution

(Prior knowledge/belief)

ELISA and HIV

a diagnostic example

- 10,000 HIV positive individuals (tested with a Western blot, the gold standard) were tested with a new ELISA test. 9990 were positive with the ELISA test, 10 were negative
- Sensitivity is 9990/10,000 or 99.9%
- 10,000 nuns who denied risk factors for HIV were also tested with the ELISA test. 9990 were negative with the ELISA test, 10 were positive
- Specificity is 9990/10,000 or 99.9%
- LR+ = 99.9/.1 = 999
- LR- = 0.1/99.9 = 0.001

Likelihood ratios

- LR+ = 99.9/.1 = 999
- LR- = 0.1/99.9 = 0.001
- Prior odds are 1/99 = 0.010101 (general population)
- 1/999 = 0.001001 (blood donors)
- 1/9 = 0.111111 (drug abusers)

Prior and posterior odds using likelihood ratios

Prior Odds		LR+ (999)	LR- (0.001001)
1 in 99	0.010101	10.1	1 x E-5
1 in 999	0.001001	1	1 x E-6
1 in 9	0.111111	111	1 x E-4

Most problems cannot be solved in a simple manner and require other methods

Numerical Methods (Software)

- Gibbs Sampling
- BUGS (=Bayesian Using Gibbs Sampling) software WinBugs
- Openbugs
- Approximate Bayesian Computation (ABC)
- Useful when calculating the probability of the data, given the parameters, is difficult or impossible.
- See Makowski et al. (2011) for an application of ABC.
- Integrated Nested Laplace Approximations (INLA)
- See Martínez-Minaya (2015) for an example

Gibbs Sampling

- **Gibbs sampling** is an **algorithm** to generate a sequence of samples from the **joint probability distribution** of two or more **random variables**.
- Gibbs sampling is an example of a **Markov chain Monte Carlo** algorithm.

Gibbs Sampler

Posterior distributions typically are hard to obtain analytically

Gibbs sampler: iterative, numerical approach that produces a Markov chain

….that eventually converges and can give information about a posterior distribution

Gibbs Sampling

- Purpose: *Draw from a Joint Distribution*

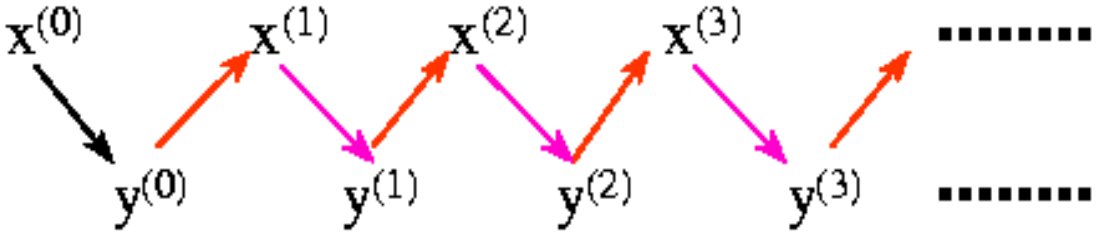
$$x = (x_1, \ldots, x_n); \quad \text{target } \pi(x)$$

- Method: *Iterative Conditional Sampling*

$$\forall\, i, \ \text{Draw} \quad x_i \sim \pi(x_i | x_{[-i]})$$

Why does it work?

- It is a *Markov Chain*!!



If $X^{(0)} = x_0$, then distribution of $X^{(t)}$ is $N(\rho^{2t} x_0, 1-\rho^{4t})$ which "converges" to $N(0,1)$ as $t \to \infty$.

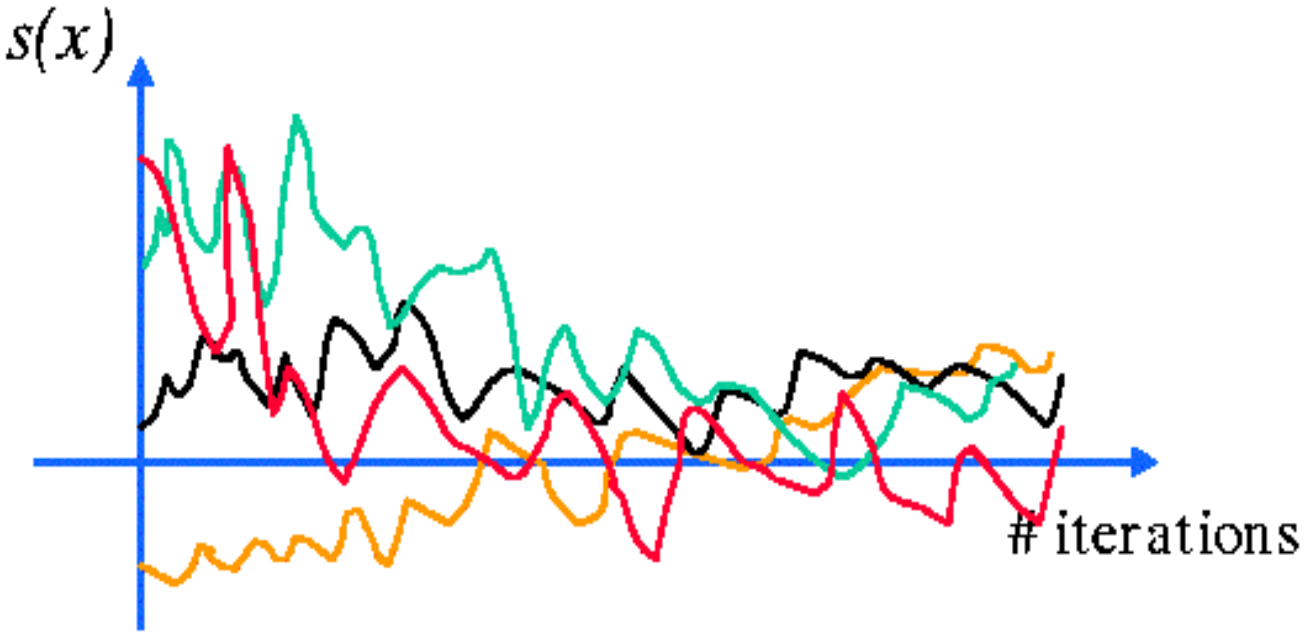
Joint distribution of $(X^{(t)}, Y^{(t)})$?

$$N\left(\begin{pmatrix}\rho^{2t}x_0\\ \rho^{2t+1}x_0\end{pmatrix},\begin{pmatrix}1-\rho^{4t} & \rho(1-\rho^{4t})\\ \rho(1-\rho^{4t}) & 1-\rho^{4t+2}\end{pmatrix}\right) \xrightarrow{t\to\infty} N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

Convergence (con't)

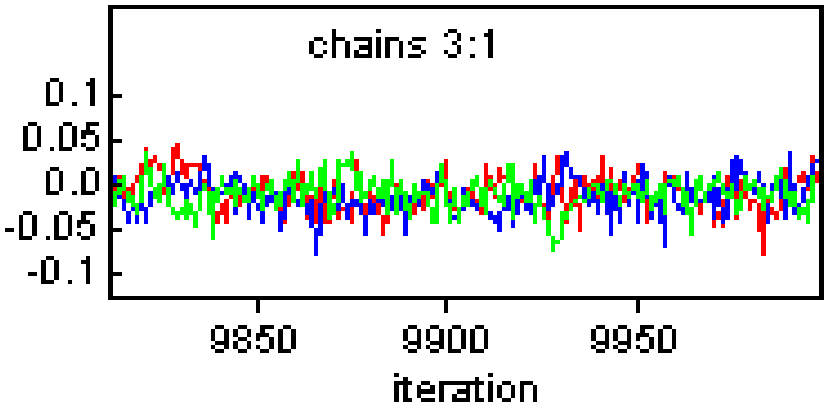
- Gelman and Rubin (1992, *Statist. Sci.*)

Multiple chains approach: Compare within-chain variation and between-chain variation.

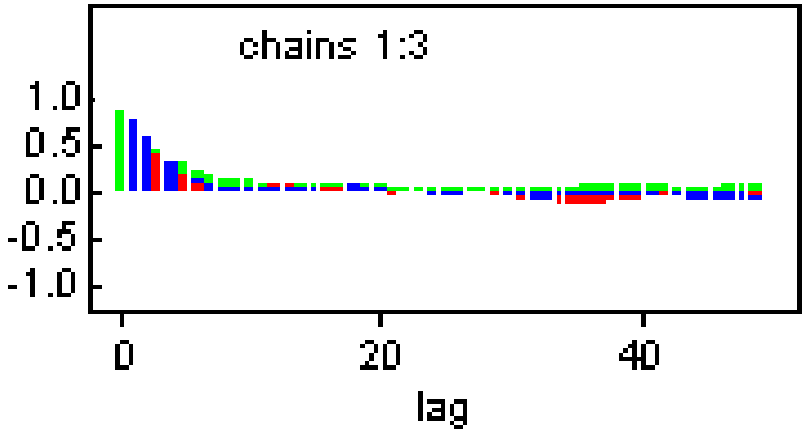


Convergence

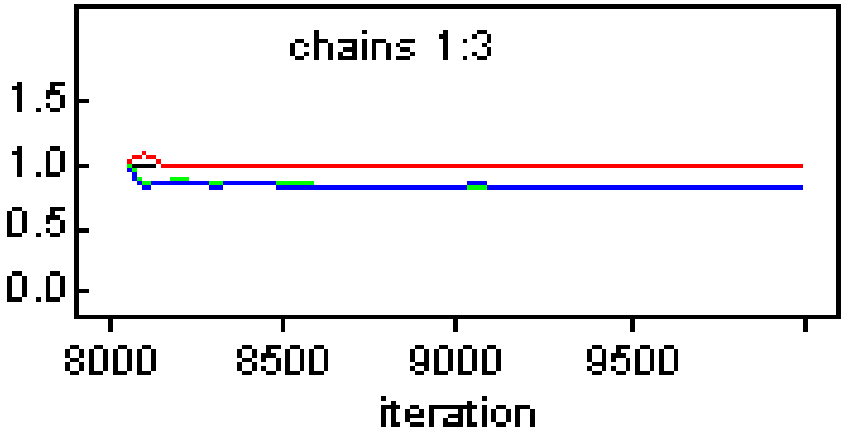
A. Trace plots



B. Autocorrelations

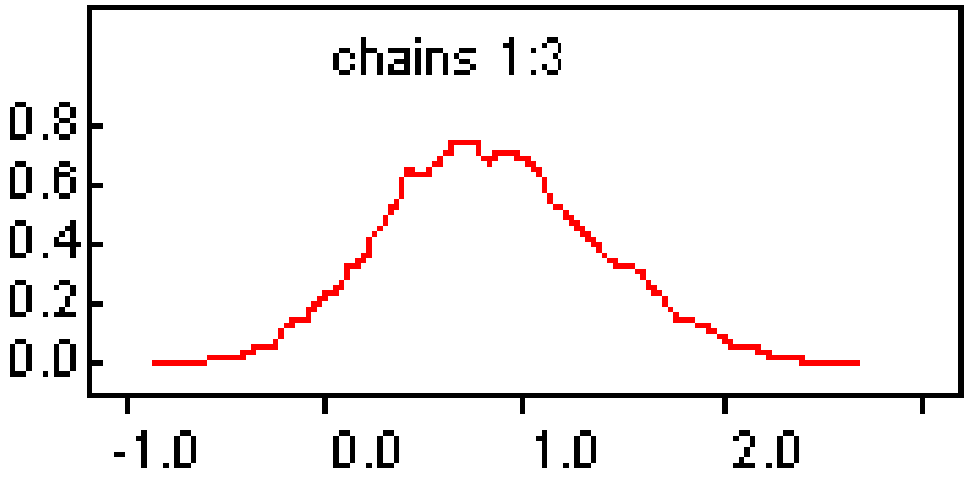


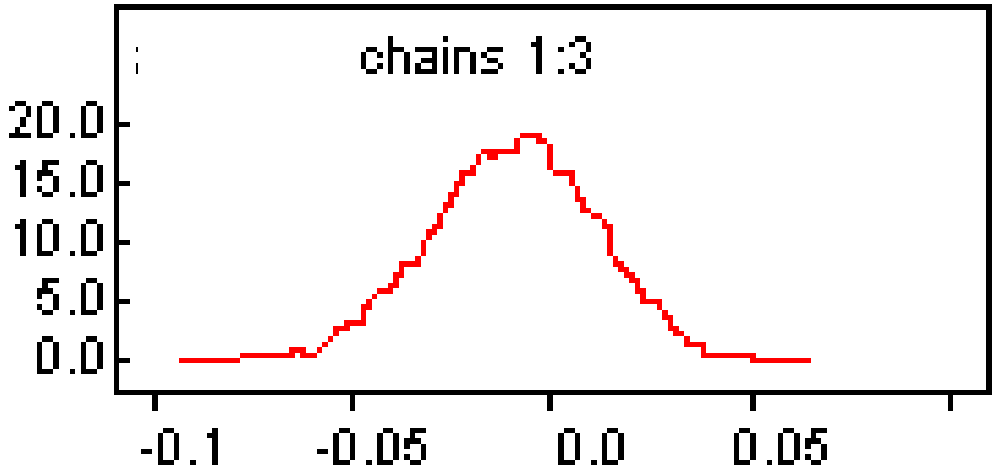
C. Gelman-Rubin (GR)

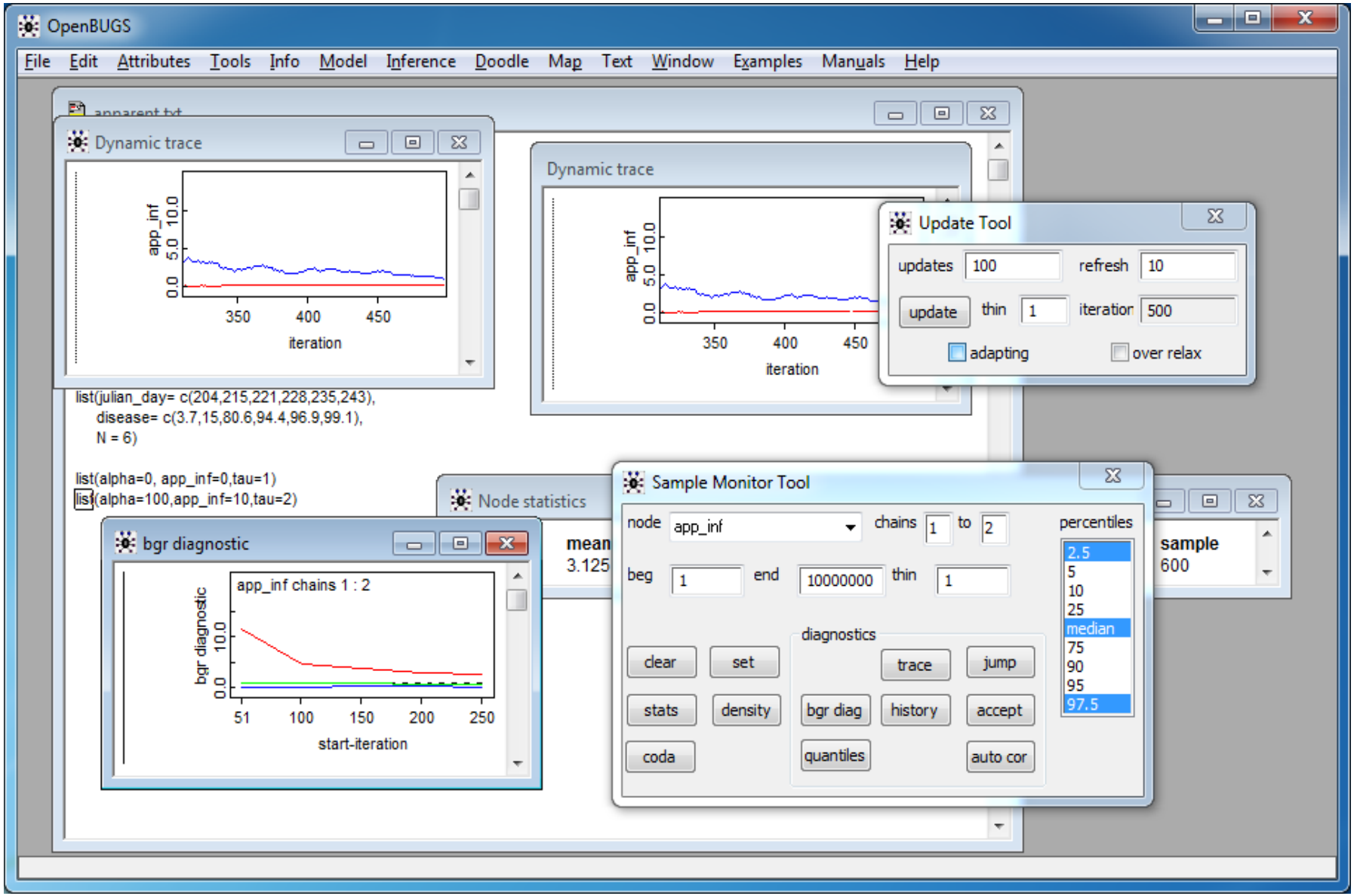


Convergence

D. Kernel density

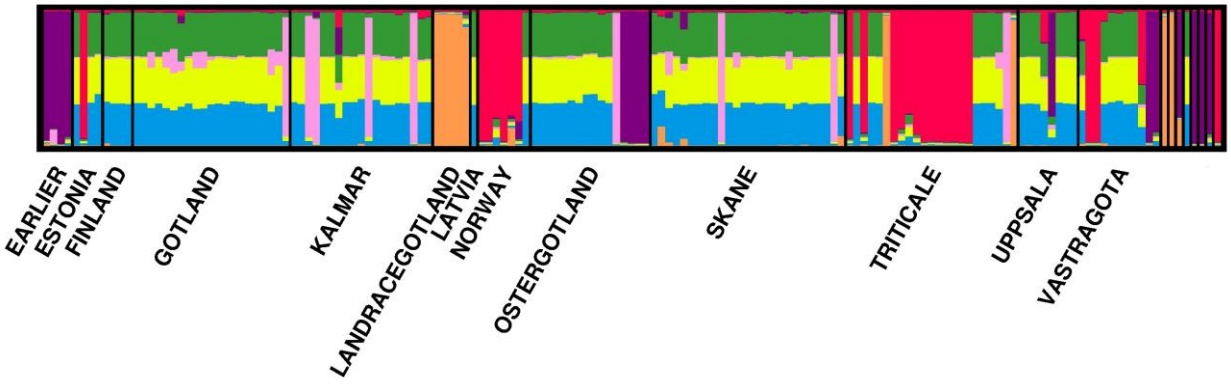


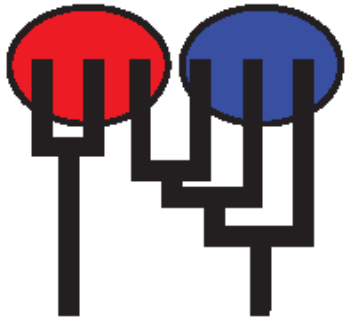



OpenBUGS
File Edit Attributes Tools Info Model Inference Doodle Map Text Window Examples Manuals Help
Dynamic trace
app_inf
0.0 5.0 10.0
350 400 450
iteration
Dynamic trace
app_inf
0.0 5.0 10.0
350 400 450
iteration
Update Tool
updates 100 refresh 10
update thin 1 iteration 500
adapting over relax
list(julian_day= c(204,215,221,228,235,243),
disease= c(3.7,15,80.6,94.4,96.9,99.1),
N = 6)
list(alpha=0, app_inf=0,tau=1)
list(alpha=100,app_inf=10,tau=2)
Node statistics
mean
3.125
sample
600
bgr diagnostic
app_inf chains 1 : 2
bgr diagnostic
0.0 10.0
51 100 150 200 250
start-iteration
Sample Monitor Tool
node app_inf chains 1 to 2
beg 1 end 10000000 thin 1
percentiles
2.5
5
10
25
median
75
90
95
97.5
diagnostics
clear set trace jump
stats density bgr diag history accept
coda quantiles auto cor

Bayesian methods are also a part of other programs

- Analysis of population groupings/origins with ’structure’
- Analysis of populations sizes and migrations patterns with ’migrate





Nordic migration of *P. infestans*

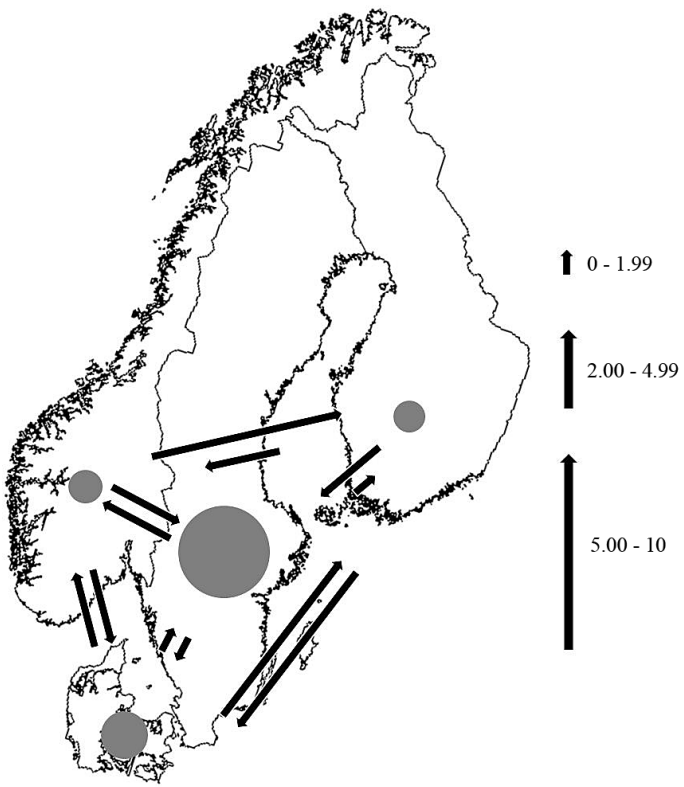
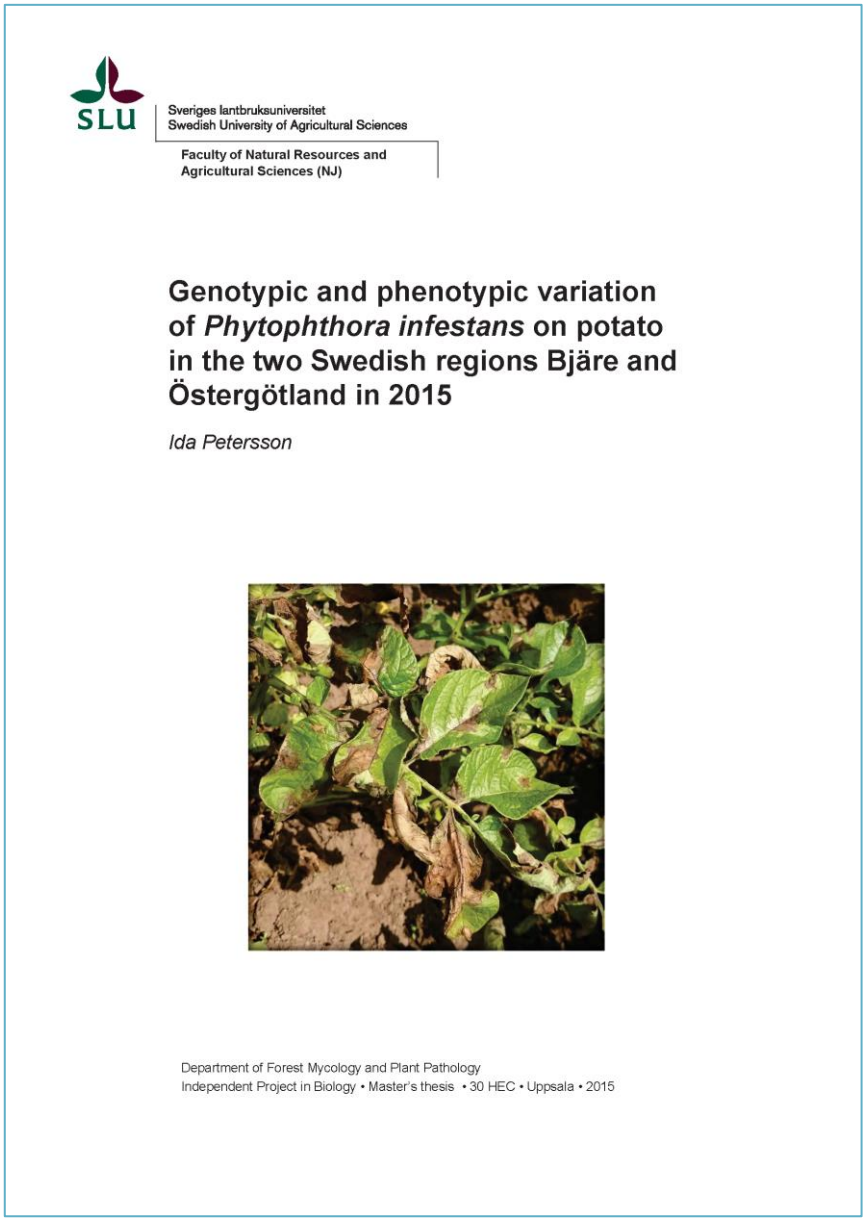


Table 4. Marginal likelihood, Bayes factor, and model probability comparing one, two, three and four subpopulation.

	Bezier approximation of marginal likelihood	Ln of (bayes factor)	Model probability
4 populations	-2333.54	0	1
3 populations	-2386.73	-53.19	0
2 populations	-2736.30	-402.76	0
1 population	-4060.35	-1726.81	0



A comparison between samples from two different regions in Sweden

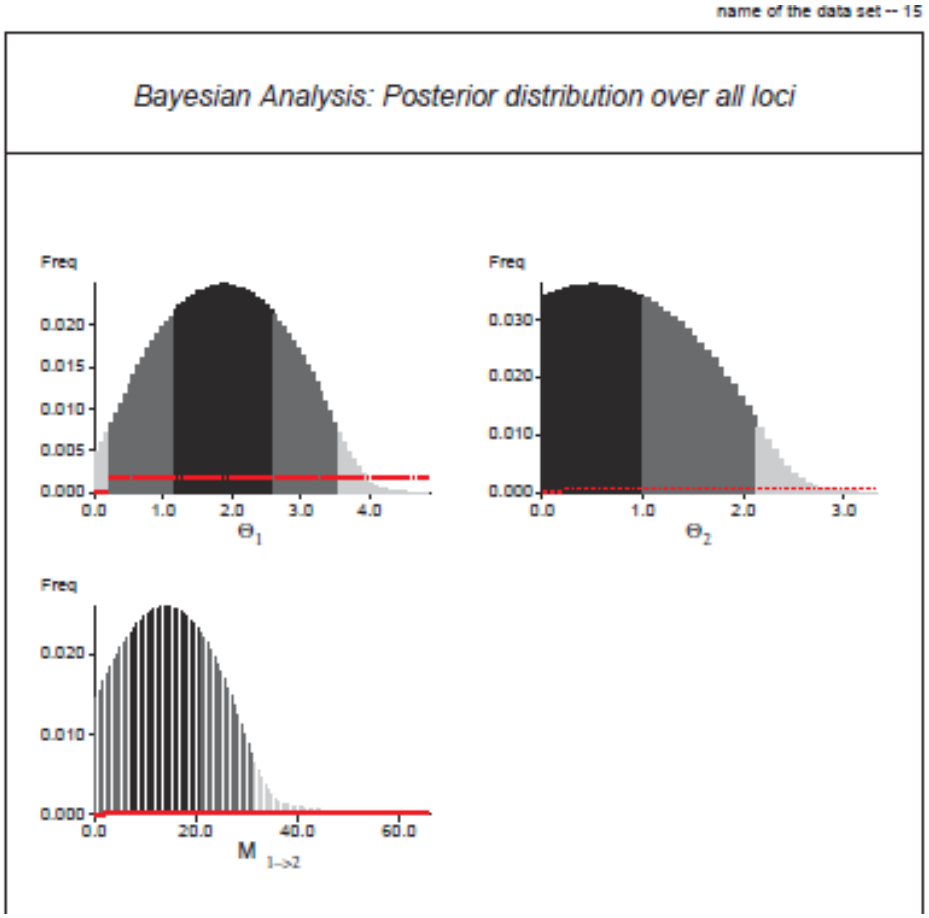
- Samples taken from Bjäre and Östergötland
- Two fields in each region
- Two foci in each field

Comparison with Bayes's Factor

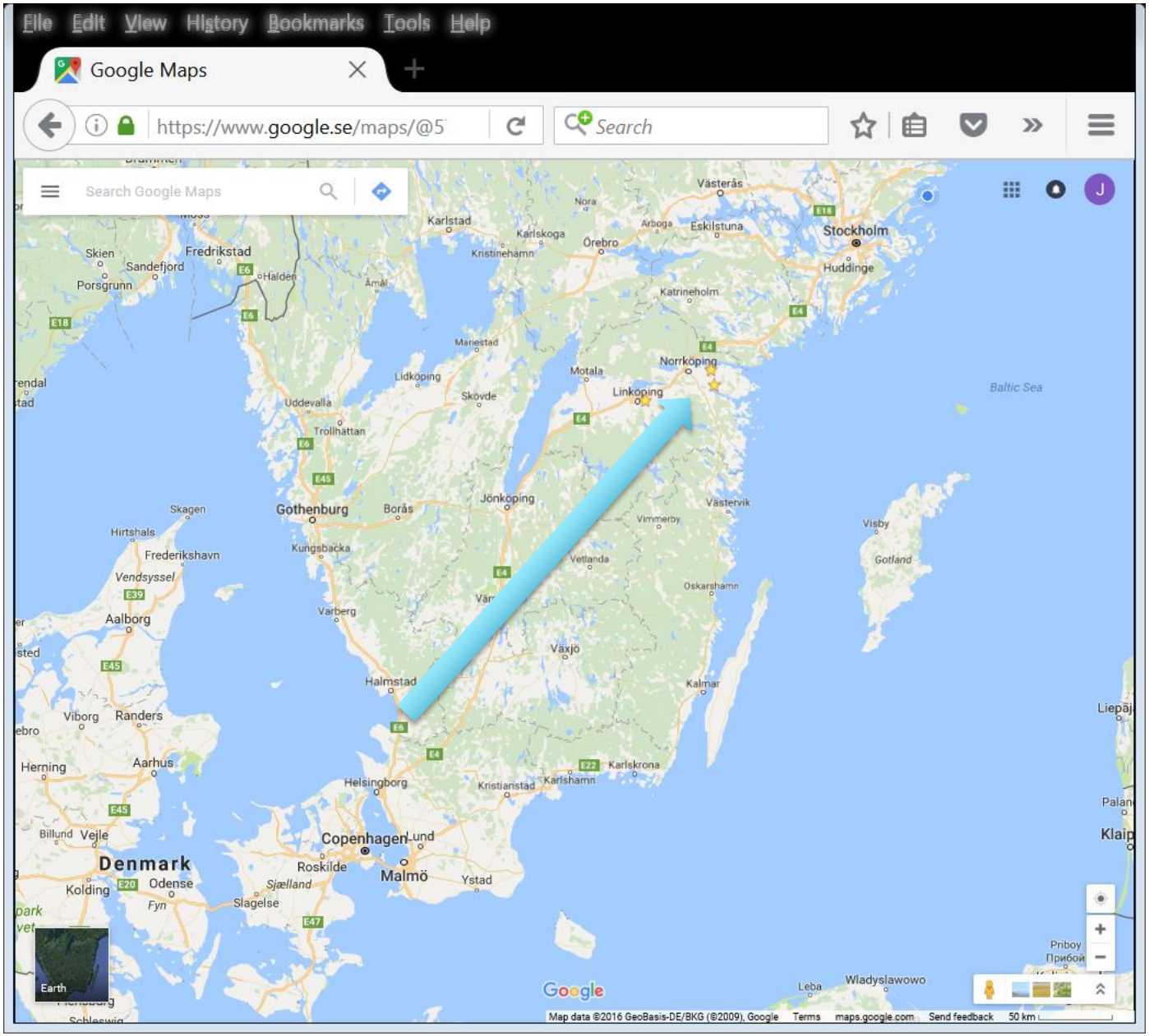
	Marginal Likelihoods				
Model	Run 1	Run 2	Run 3	Run 4	Average
1 single population	-1341.36	-1291.8	-1333.39	-1312.27	-1319.705
2 populations, 2-way migration	-1316.6	-1407.32	-1352.26	-1321.58	-1349.44
2 populations, Bjäre to Östergötland only	-1178.38	-1188.34	-1179.57	-1179.64	-1181.4825
2 populations, Östergötland to Bjäre only	-1169.57	-1209.05	-1201.77	-1199.04	-1194.8575

Larger value Marginal Likelihood → better model. Best model here is 2 populations with migration only from Bjäre to Östergötland

Estimates of population sizes and migration



Bjäre population is larger than the one in Östergötland


File Edit View History Bookmarks Tools Help
Google Maps
https://www.google.se/maps/@5
Search
Search Google Maps
Stockholm
Huddinge
Västerås
Eskilstuna
Örebro
Karlskoga
Karlstad
Nora
Arboga
Kristinehamn
Katrineholm
Norrköping
Linköping
Motala
Mariestad
Lidköping
Skövde
Uddevalla
Trollhättan
Baltic Sea
Gothenburg
Borås
Jönköping
Kungsbacka
Varberg
Halmstad
Växjö
Kalmar
Oskarshamn
Vetlanda
Vimmerby
Västervik
Visby
Gotland
Skagen
Hirtshals
Frederikshavn
Vendsyssel
Aalborg
Fredrikstad
Sandefjord
Skien
Porsgrunn
Halden
Åmål
Viborg
Randers
Aarhus
Herning
Billund
Vejle
Denmark
Kolding
Odense
Fyn
Slagelse
Sjælland
Roskilde
Copenhagen
Lund
Malmö
Ystad
Helsingborg
Kristianstad
Karlshamn
Karlskrona
Liepāj
Klaip
Palan
Priboy
Прибой
Leba
Wladyslawowo
Earth
Google
Map data ©2016 GeoBasis-DE/BKG (©2009), Google
Terms
maps.google.com
Send feedback
50 km

Where do we stand?

- Bayesian methods are already being used in fields directly related to PRA (parameter estimation)
- Bayesian networks can be used for risk assessment either through commercial programs or open source solutions
- Bayesian methods often are a built-in part of programs used in analysis of population structure of different organisms

How can we optimize acquiring new methodology?

- We should think about how our work is related to neighboring disciplines can be a good start.
- What are our neighboring disciplines?
- Population biology?
- Epidemiology?
- What else?