L2. Model parameterization and validation

OVERVIEW

. L1. Approaches to ecological modelling

. L2. Model parameterization and validation

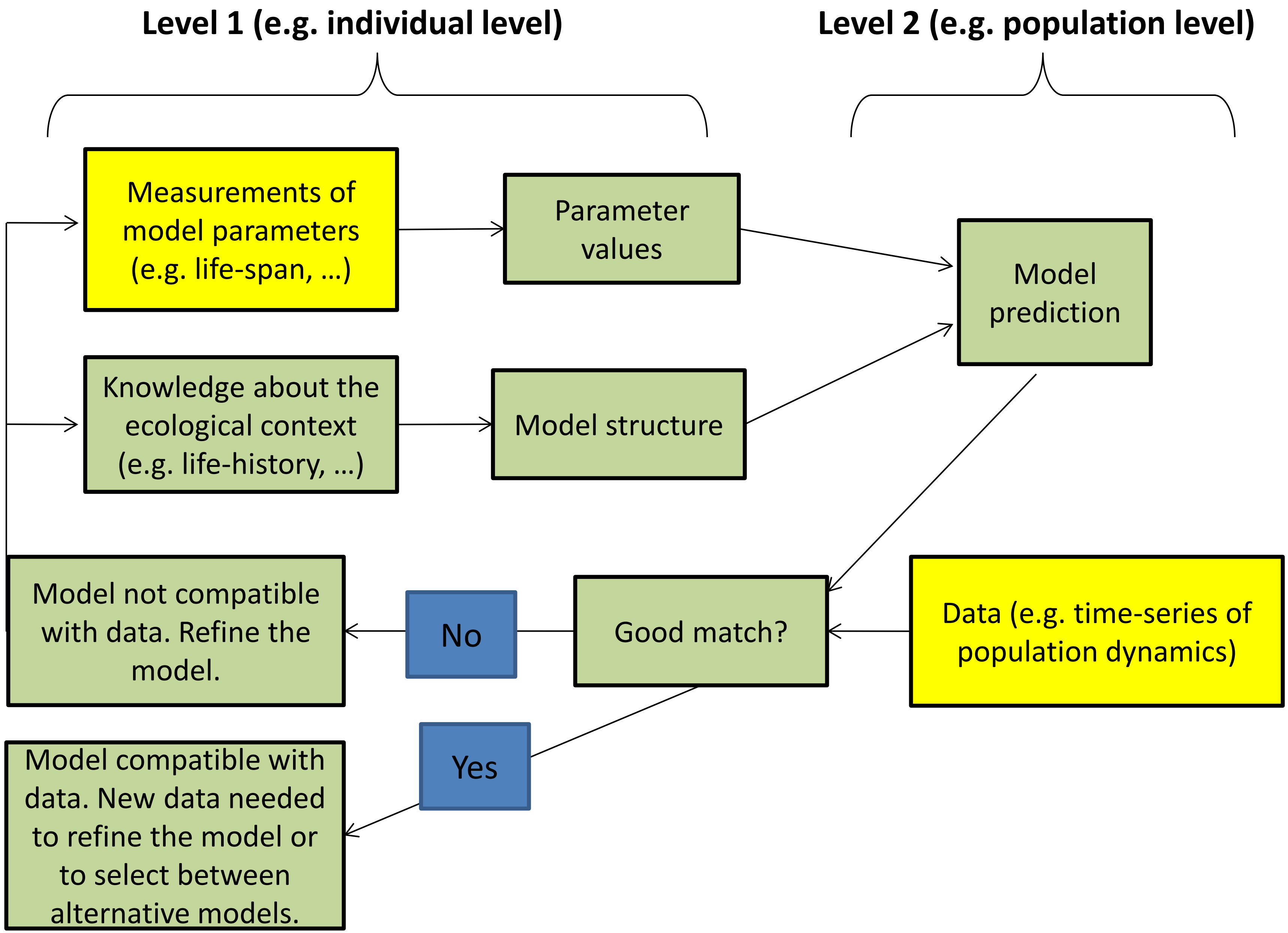
. L3. Stochastic models of population dynamics (math)

. L4. Animal movement (math + stat)

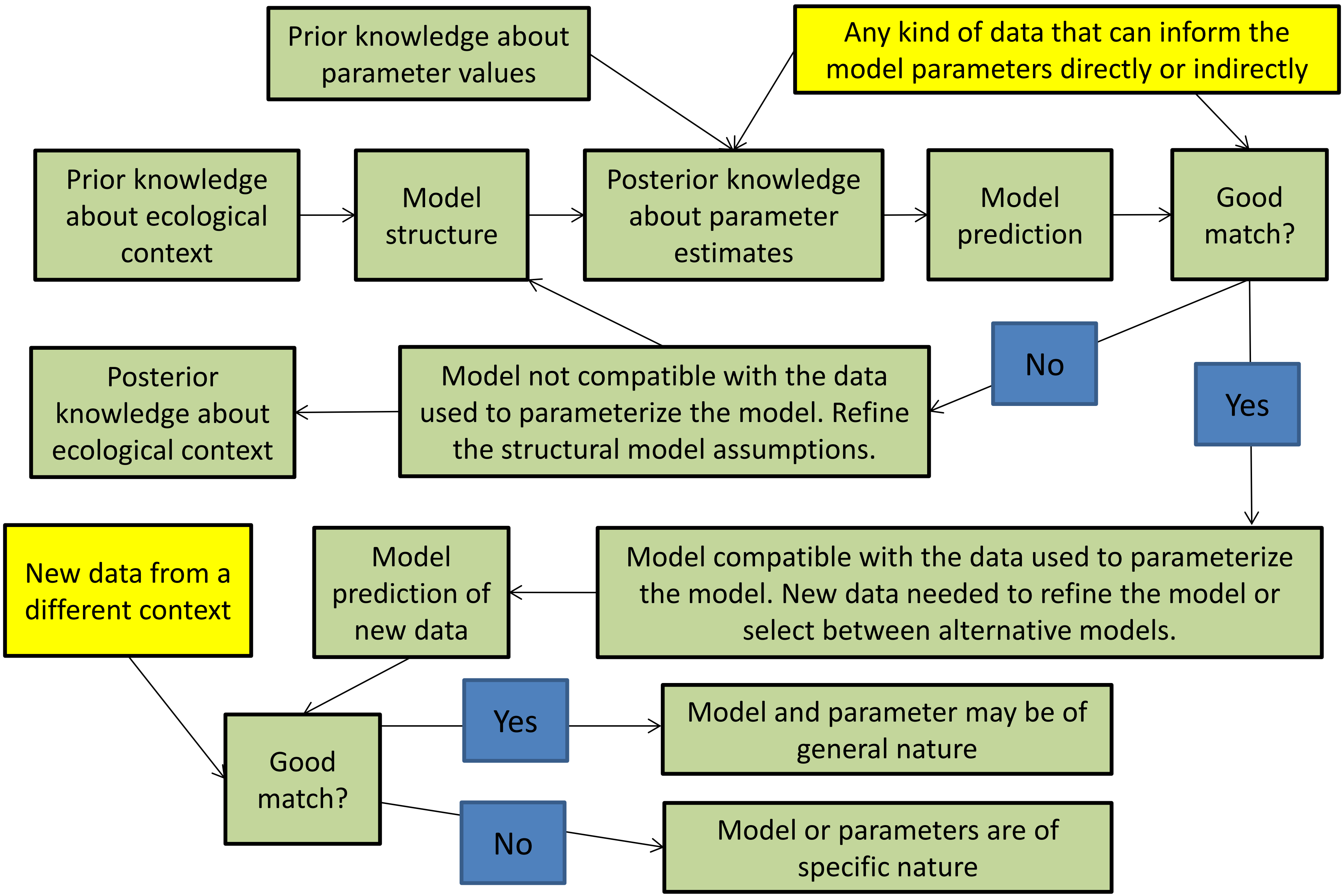
. L5. Quantitative population genetics (math + stat)

. L6. Community ecology (stat)

The forward approach to model parameterization



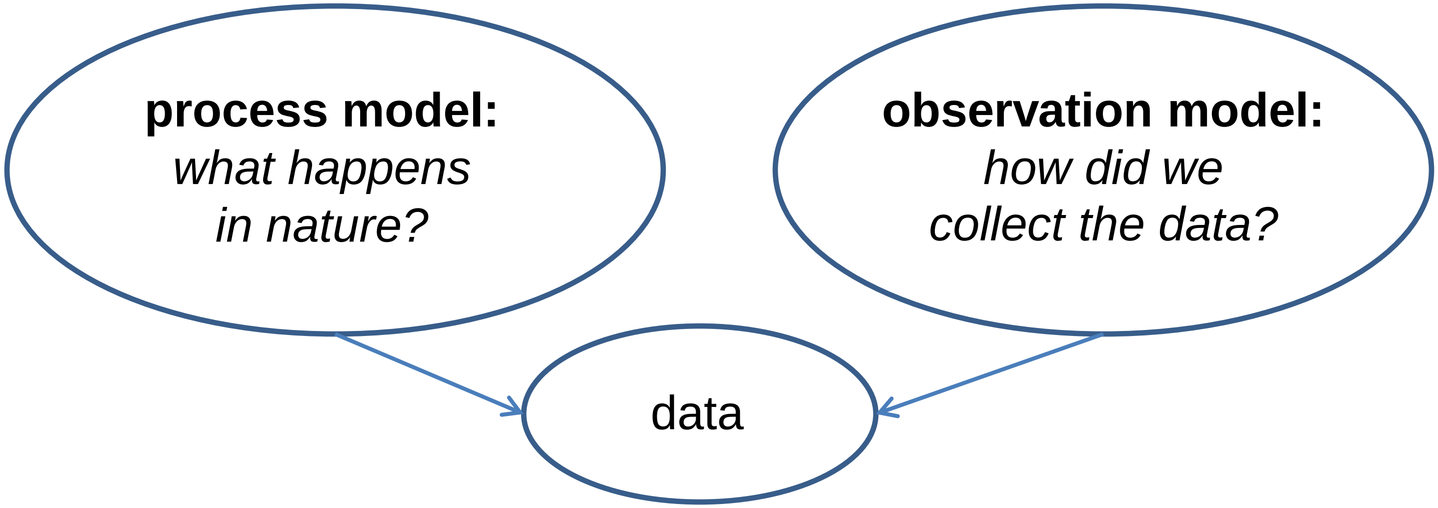
The Bayesian state-space approach to model parameterization



How to combine the forward and inverse-approaches in practice?

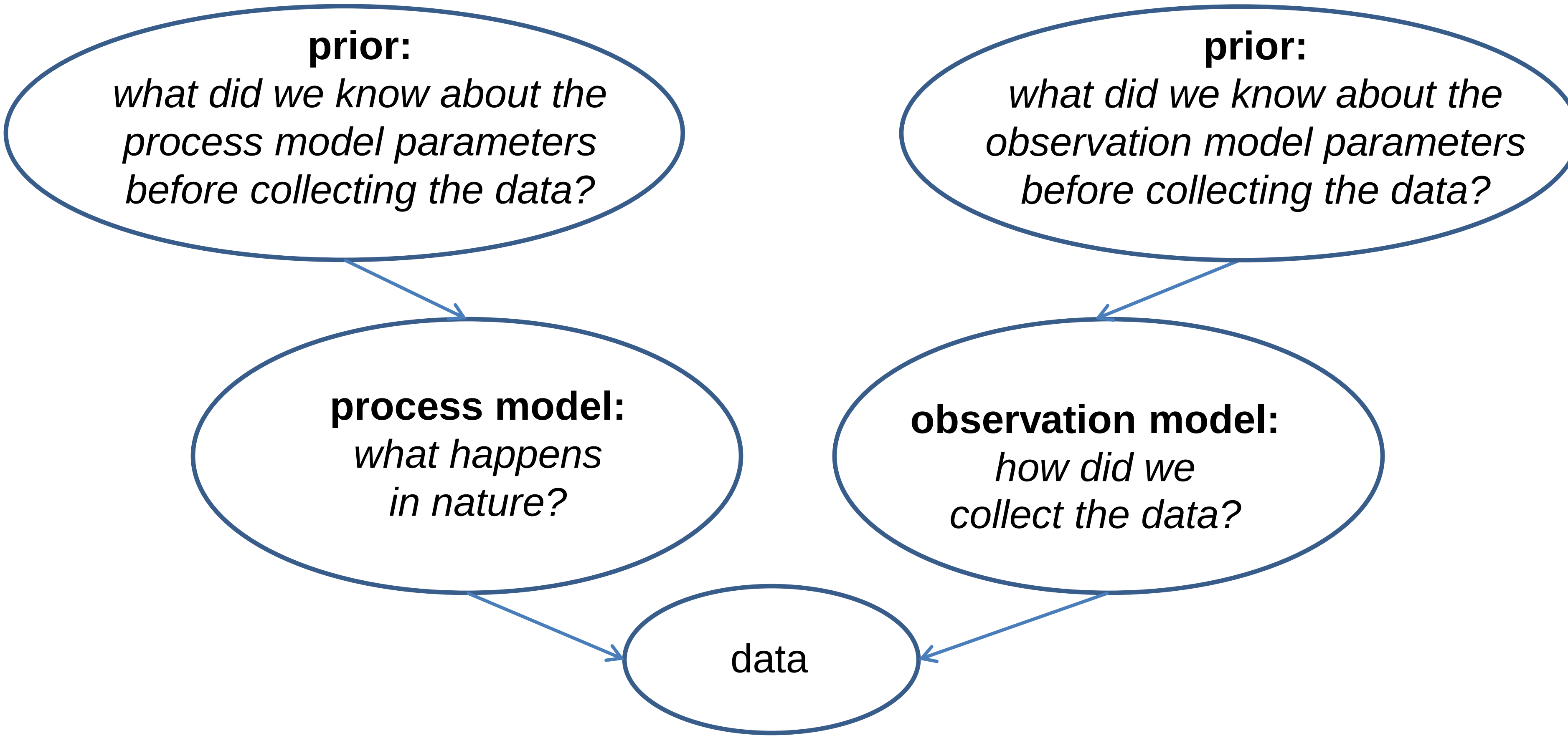
State-space models

Also called Hidden Markov models, process based models...



How to combine the forward and inverse-approaches?

Bayesian state-space models



Example: stochastic logistic model

Number of individuals is $n = n(t) = 0,1,2,3, \ldots$

Individuals produce new individuals at per-capita fecundity rate f

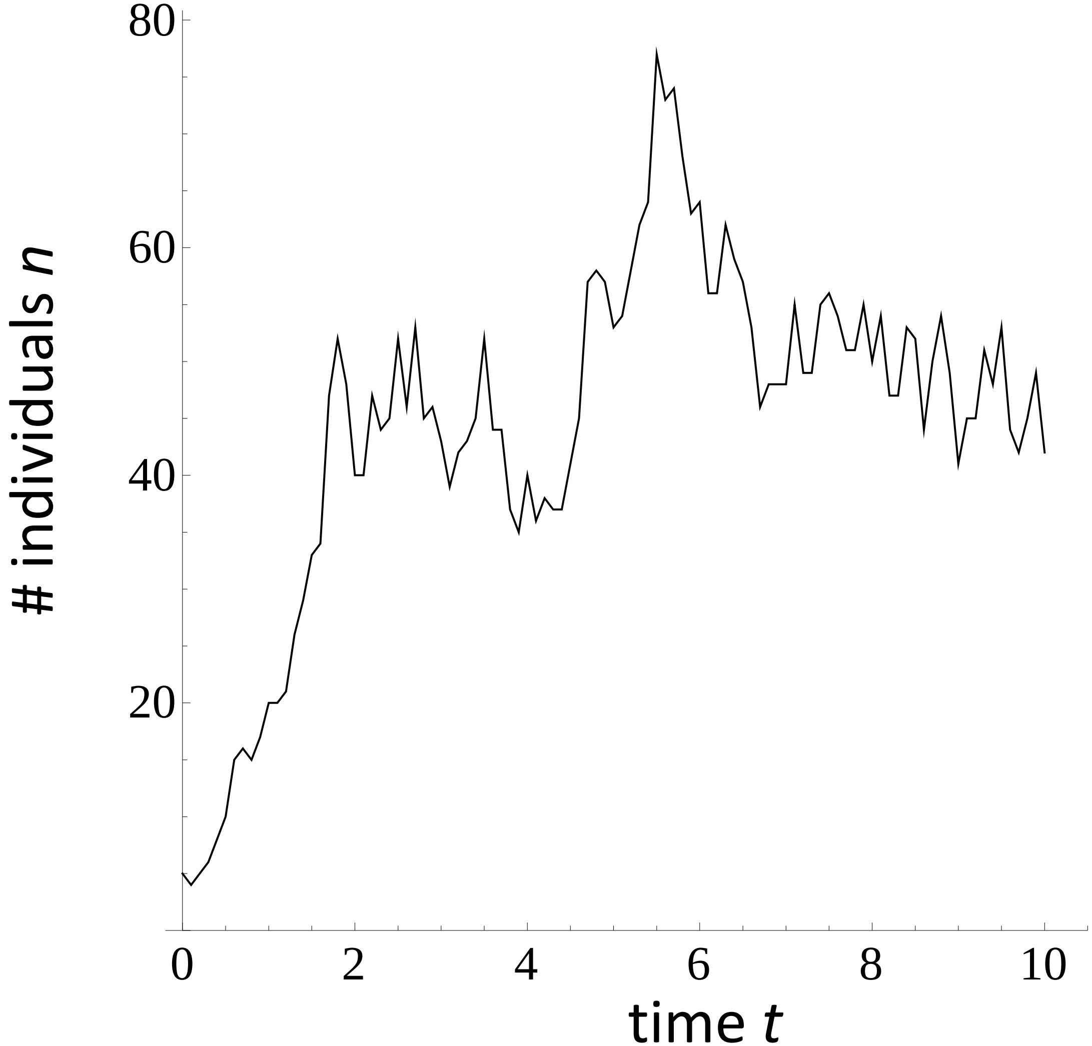
The per-capita death rate is $d + c(n-1)$, where d is the density-independent background mortality rate and the parameter c describes the additional death rate imposed by competitive effects of the $n-1$ individuals to the focal individual

The model is a stochastic Markov process. The deterministic mean-field model is

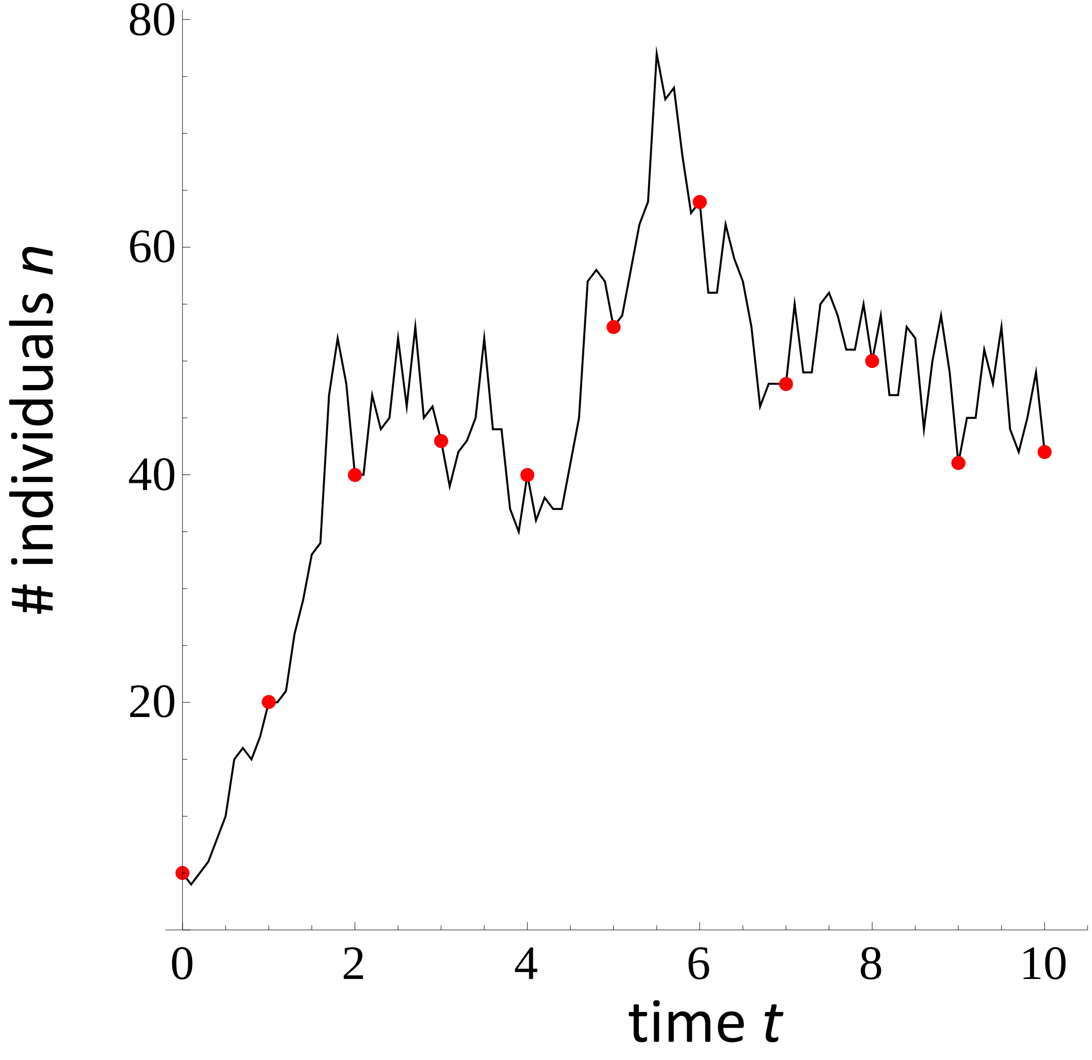
$$\frac{dn}{dt} = (f-d)n - cn^2 = rn(1 - n/K),$$

$r = f - d$ is the growth rate of the population at low density
$K = r/c$ is the carrying capacity

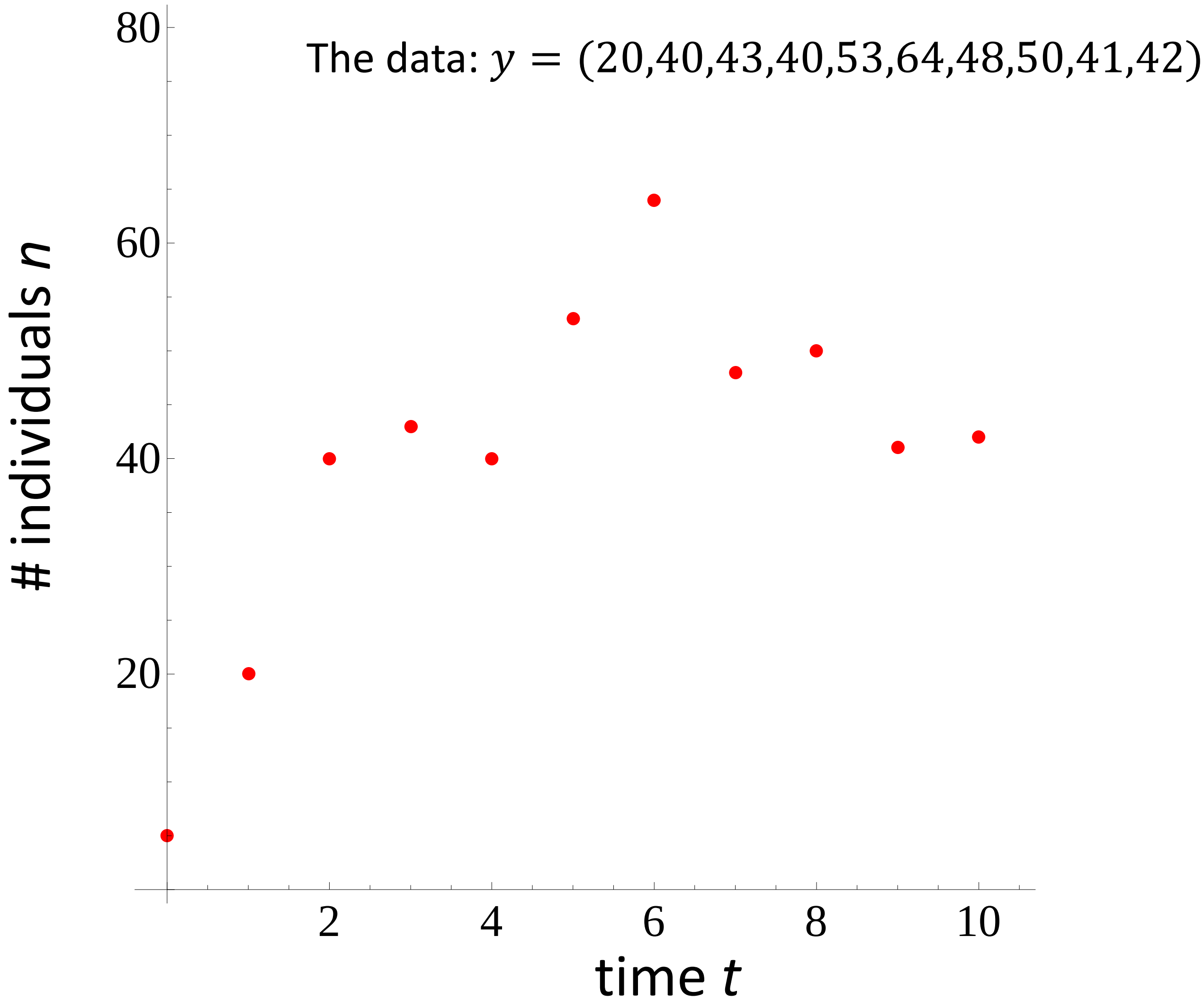
Model simulation with "true" parameter values $f = 3$, $d = 1$ and $K = 50$. Initial state $n(0) = 5$.



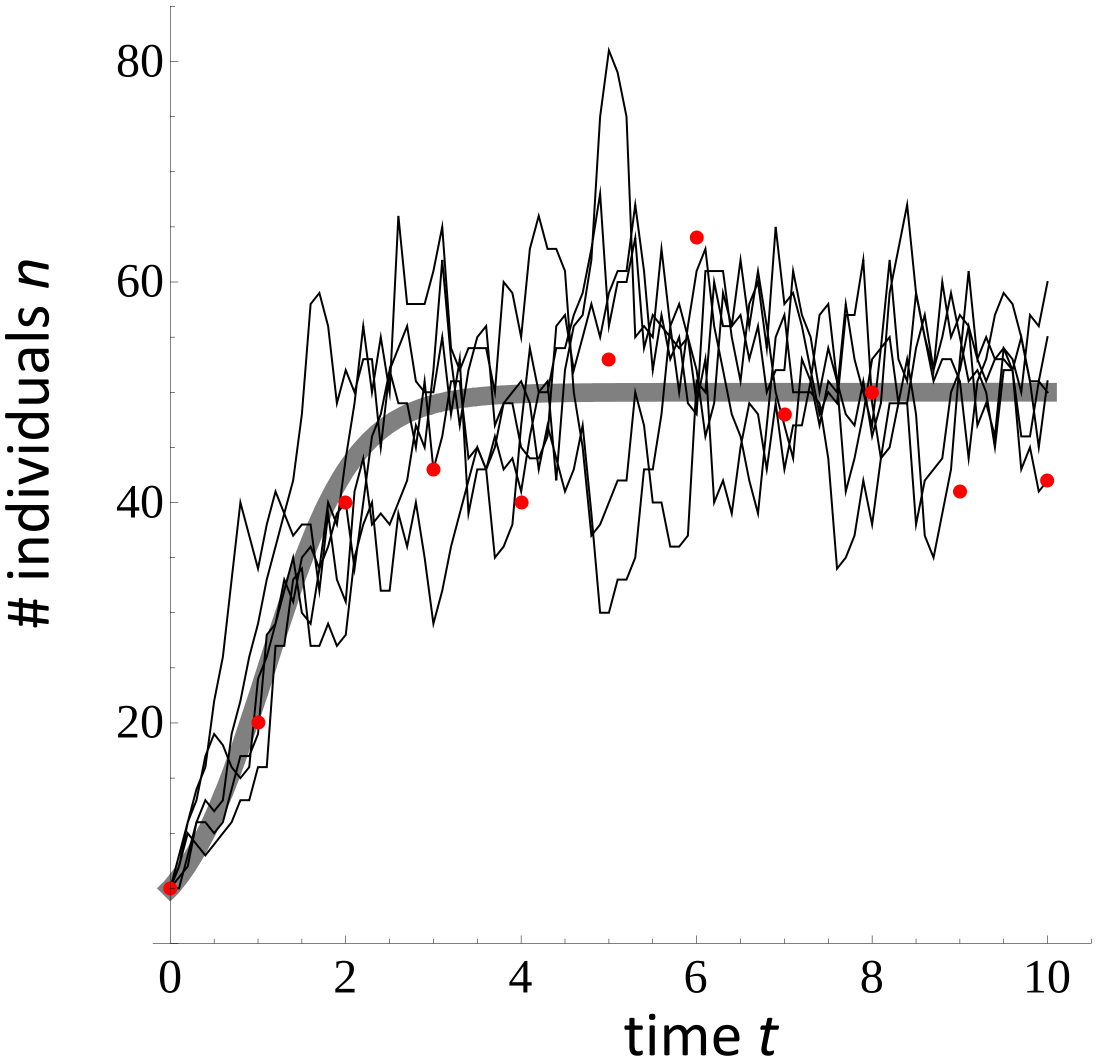
Model simulation with “true” parameter values $f = 3$, $d = 1$ and $K = 50$.
Initial state $n(0) = 5$.



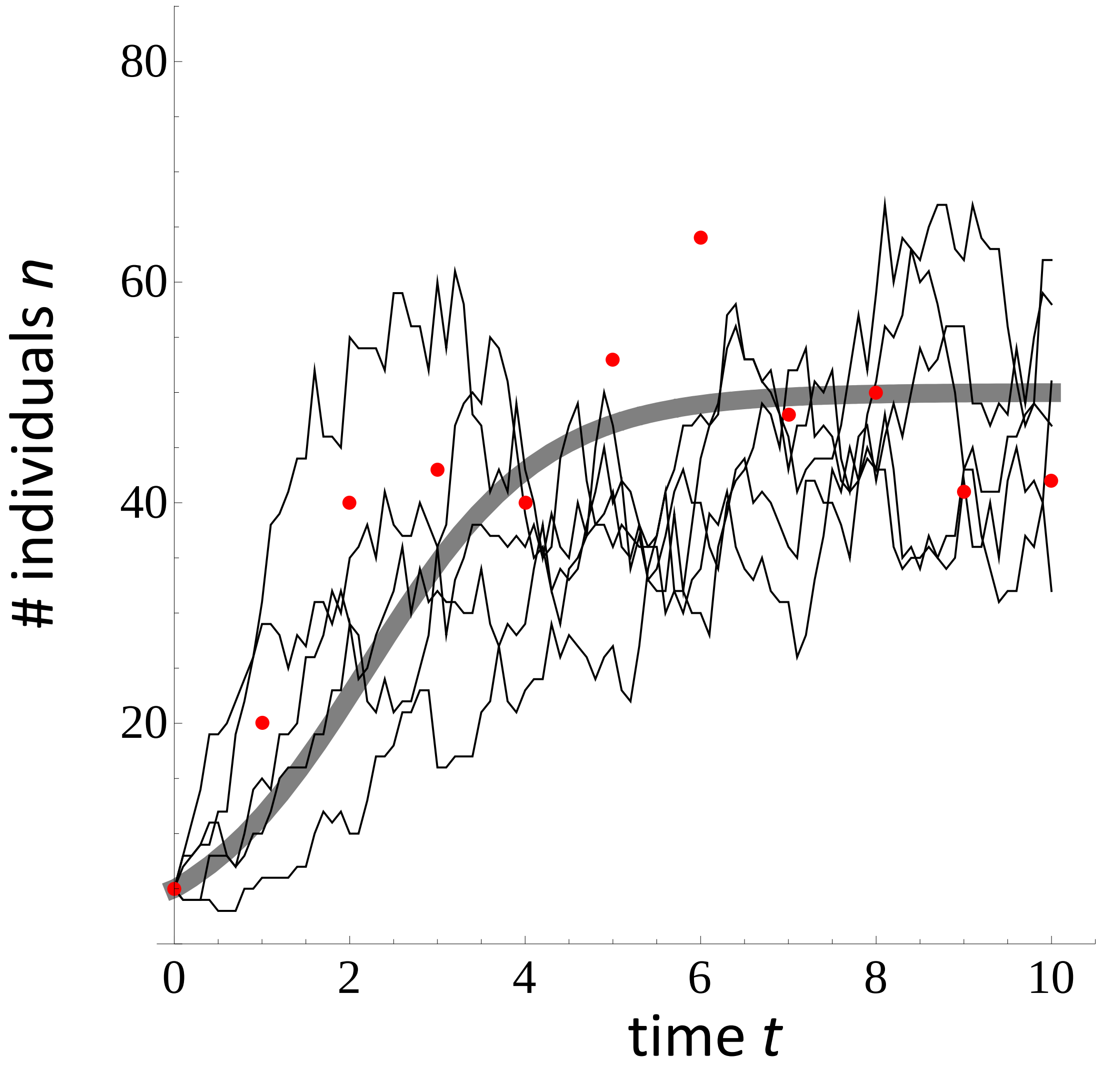
Model simulation with "true" parameter values $f = 3$, $d = 1$ and $K = 50$. Initial state $n(0) = 5$.



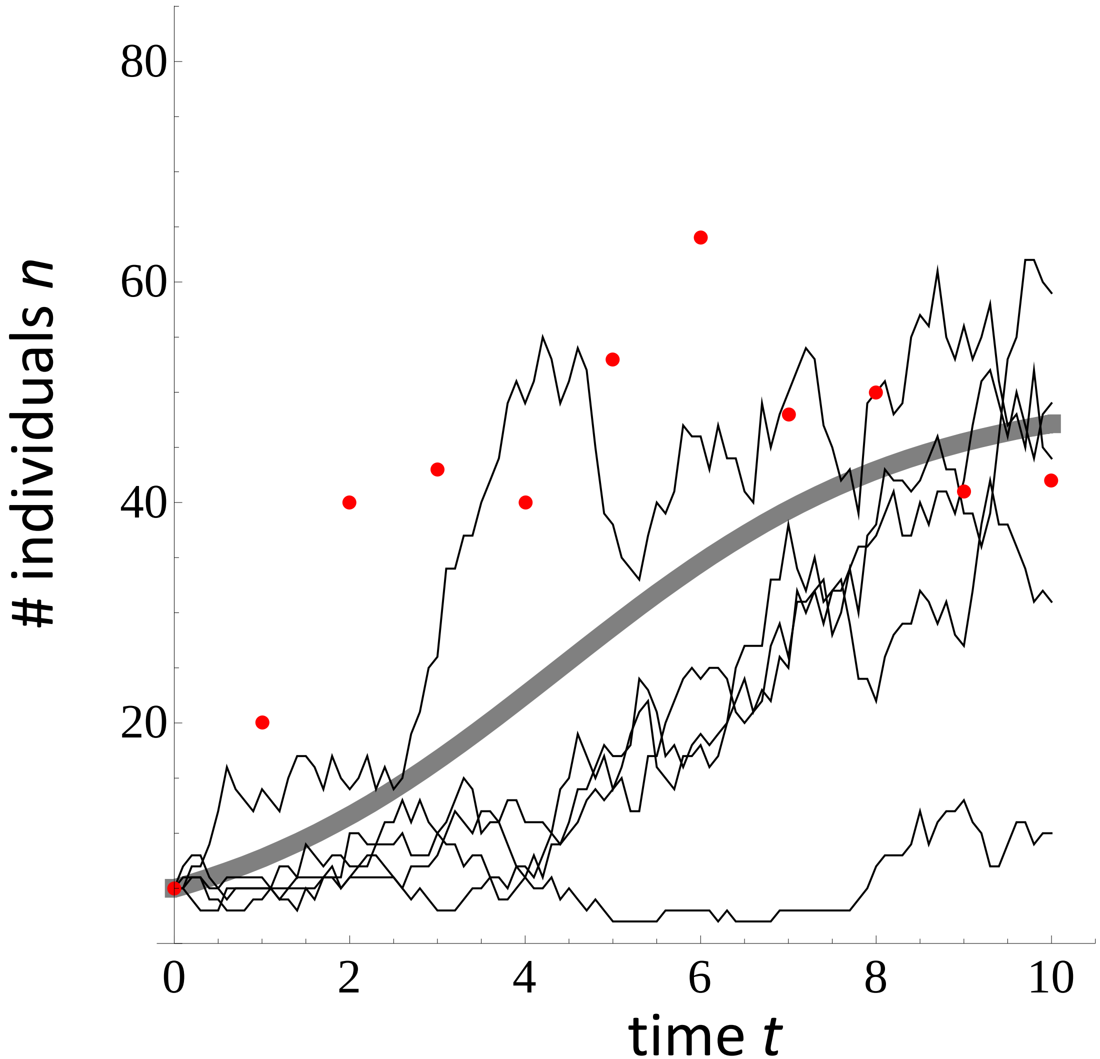
“True” parameter values $f = 3$, $d = 1$ and $K = 50$.
Simulations with $f = 3$, $d = 1$ and $K = 50$.



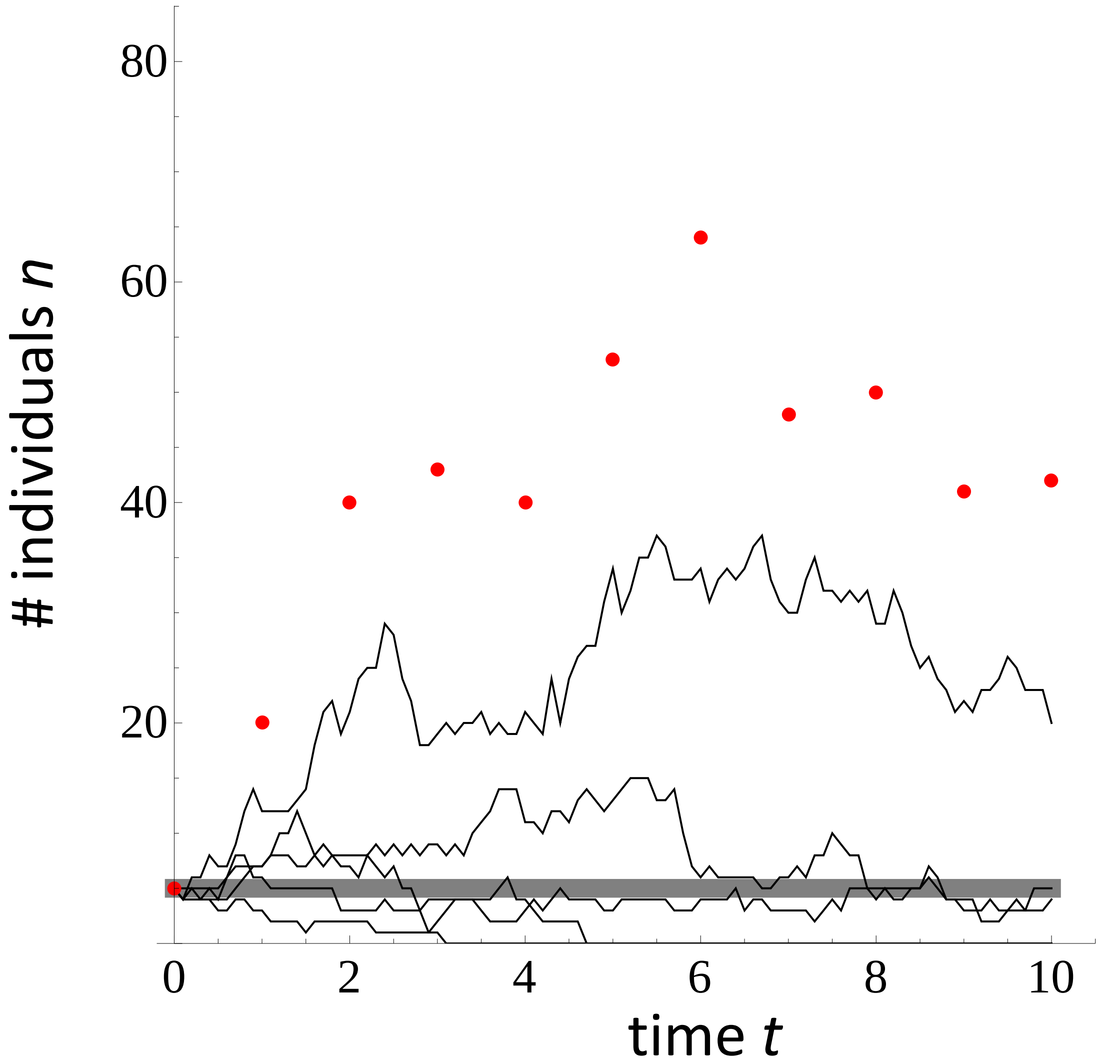
“True” parameter values $f = 3$, $d = 1$ and $K = 50$.
Simulations with $f = 2$, $d = 1$ and $K = 50$.



“True” parameter values $f = 3$, $d = 1$ and $K = 50$.
Simulations with $f = 1.5$, $d = 1$ and $K = 50$.



“True” parameter values $f = 3$, $d = 1$ and $K = 50$.
Simulations with $f = 1$, $d = 1$ and $K = 50$.



Likelihood of observing the data

The data: $y = (20,40,43,40,53,64,48,50,41,42)$

The parameters: $\theta = (f, d, K)$

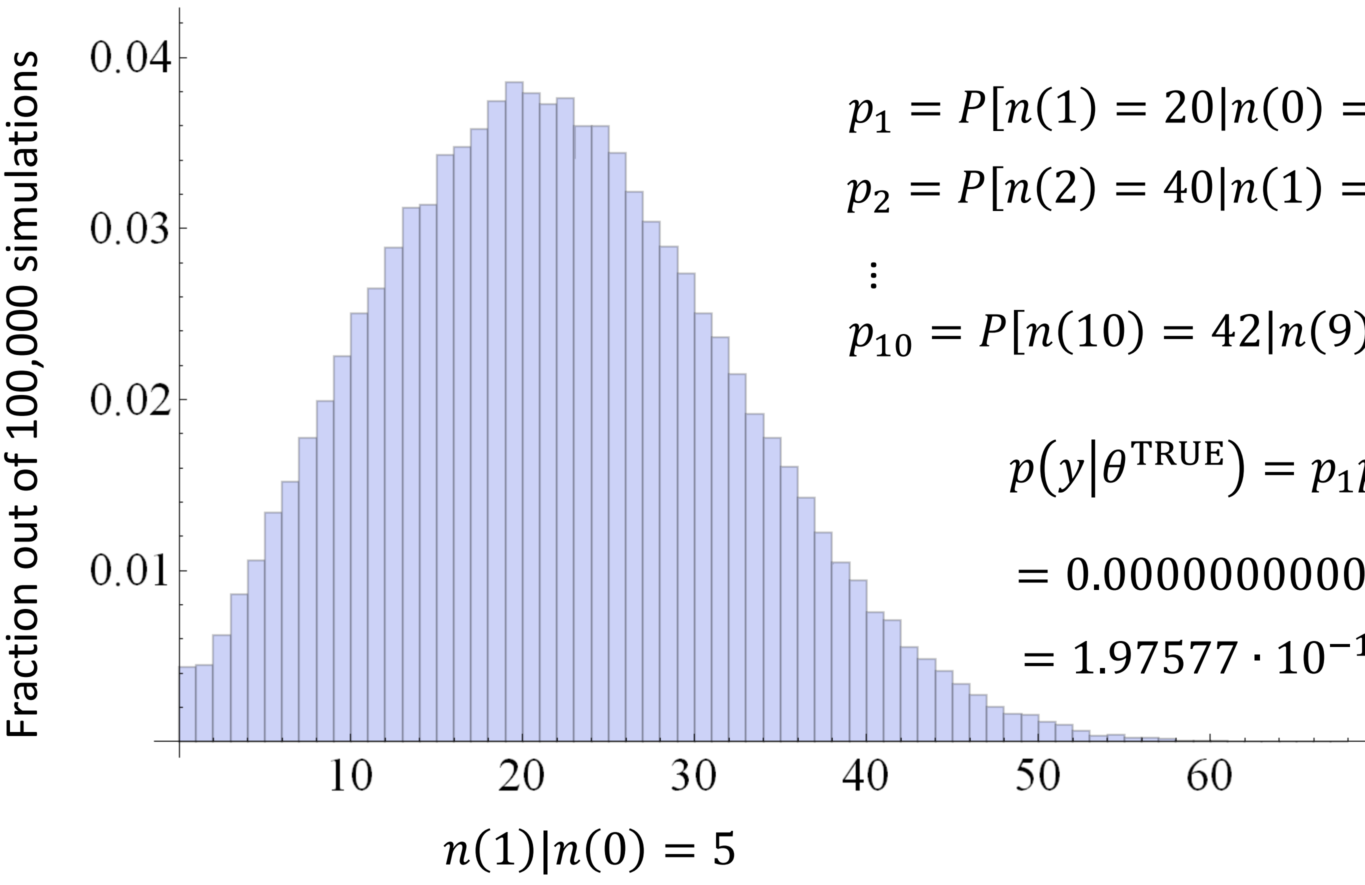
$$\theta^{\mathrm{TRUE}} = (3,1,50)$$

The probability (likelihood) of observing the data, given the model and the model parameters: $p(y|\theta)$

How to compute the likelihood of observing the data

The data: $y = (20,40,43,40,53,64,48,50,41,42)$

The experiment was initiated at day 0 with 5 individuals. What is the probability that there would be 20 individuals at day 1, assuming θ^{TRUE}?



$$p_1 = P[n(1) = 20|n(0) = 5] = 0.03815$$

$$p_2 = P[n(2) = 40|n(1) = 20] = 0.043095$$

$$\vdots$$

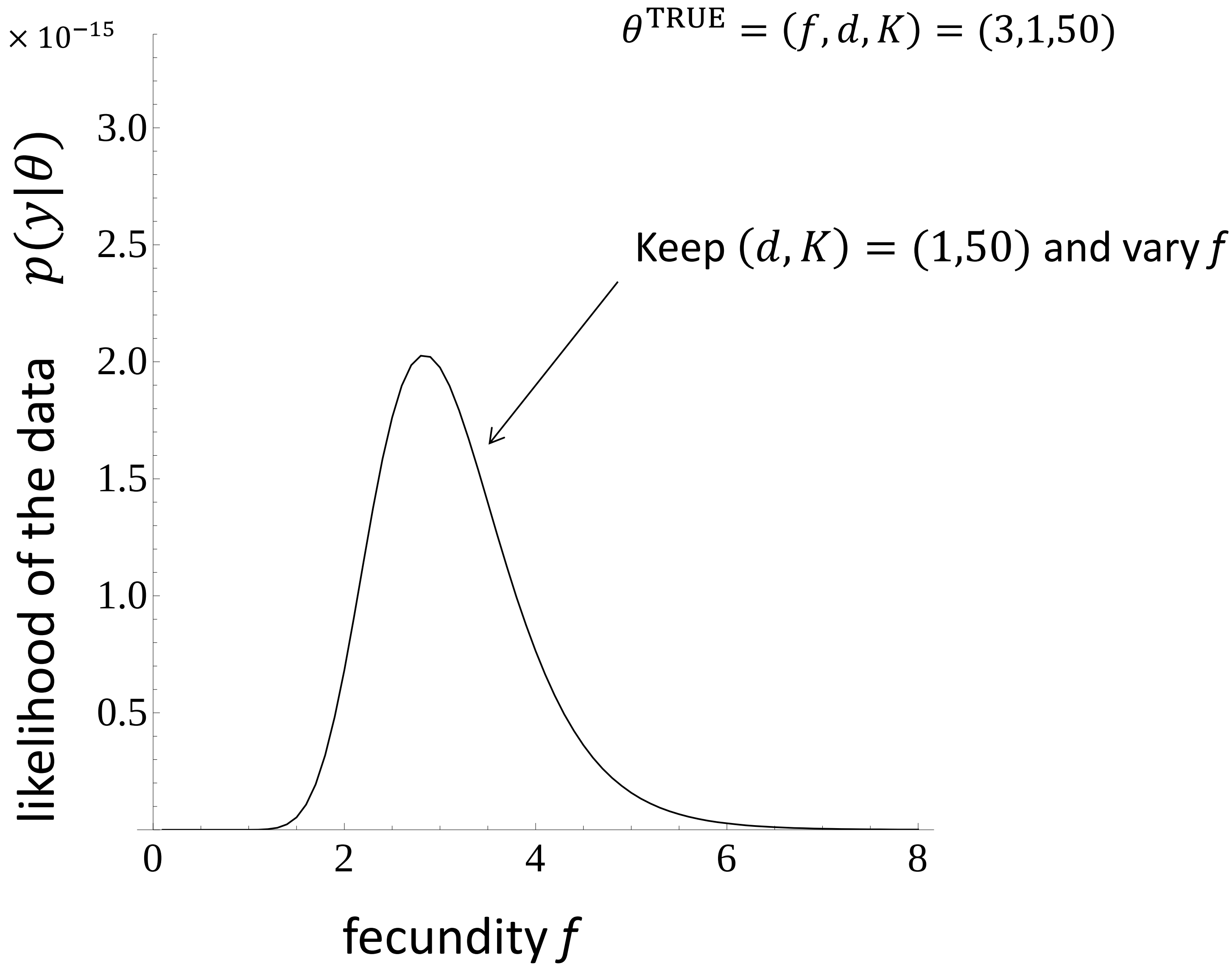
$$p_{10} = P[n(10) = 42|n(9) = 41] = 0.0368283$$

$$p\left(y\middle|\theta^{\mathrm{TRUE}}\right) = p_1 p_2 \dots p_{10}$$

$$= 0.00000000000000197577$$

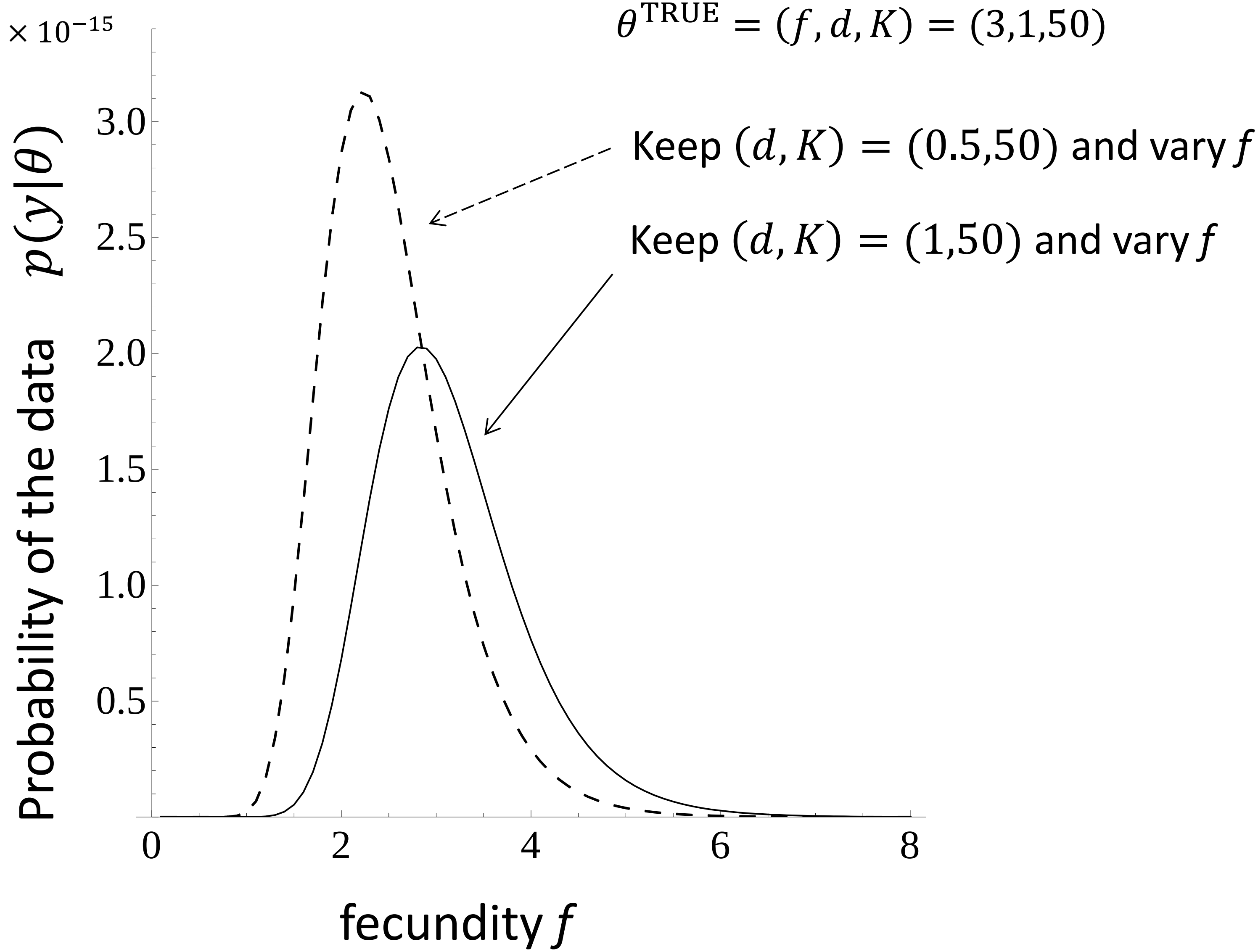
$$= 1.97577 \cdot 10^{-15}$$

How does the likelihood depend on model parameters?



How does the likelihood depend on model parameters?

$\theta^{\mathrm{TRUE}} = (f, d, K) = (3,1,50)$



Bayesian inference

$p(y|\theta)$: the likelihood of observing the data y conditional on the parameters θ

$p(\theta|y)$: the posterior: the probability distribution of parameters, given y.

$p(\theta)$: the prior: what we assumed about the parameters before seeing the data

Bayes theorem:

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

How to choose the prior

Sometimes there is prior information, e.g. from other studies.

> *"We followed singly grown individuals through their life-times, from which data we estimate the density-independent death rate d to be between 0.7 and 1.1."*

Often there is no prior information. Then one may assume an "uninformative prior".

Let us assume, for the sake of illustration, the following prior:

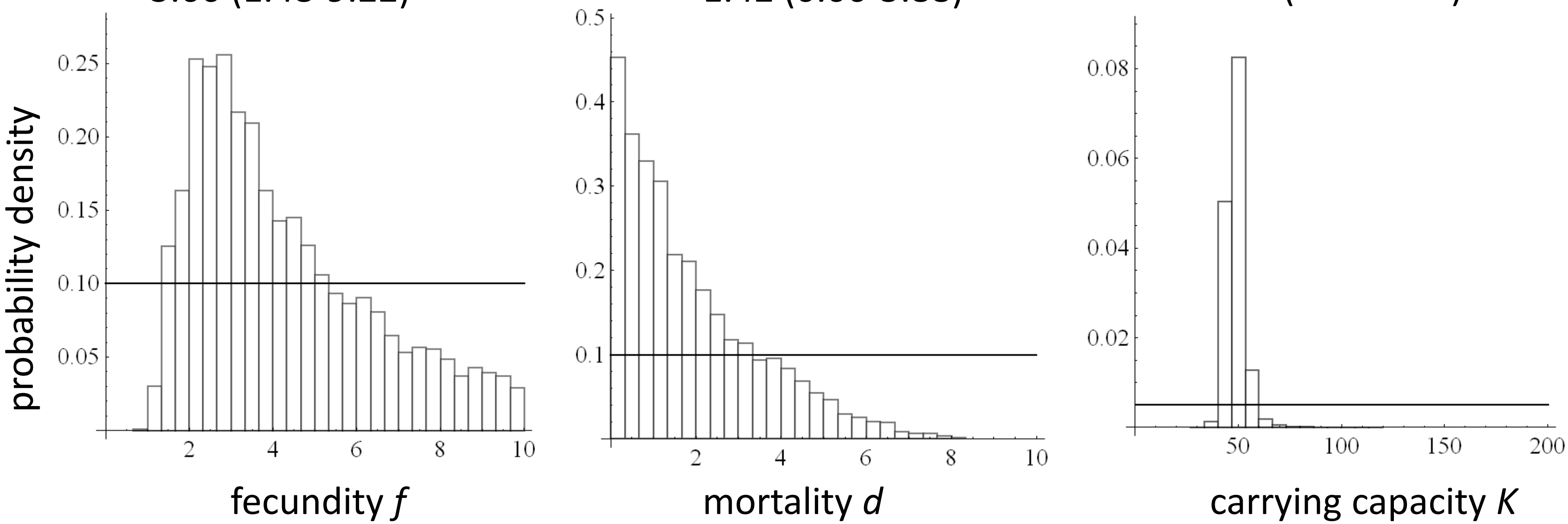
$$p(\theta) = 1/20000 \text{ if } 0 \leq f, d \leq 10 \text{ and } 0 \leq K \leq 200$$

otherwise $p(\theta) = 0$

The marginal posterior distributions

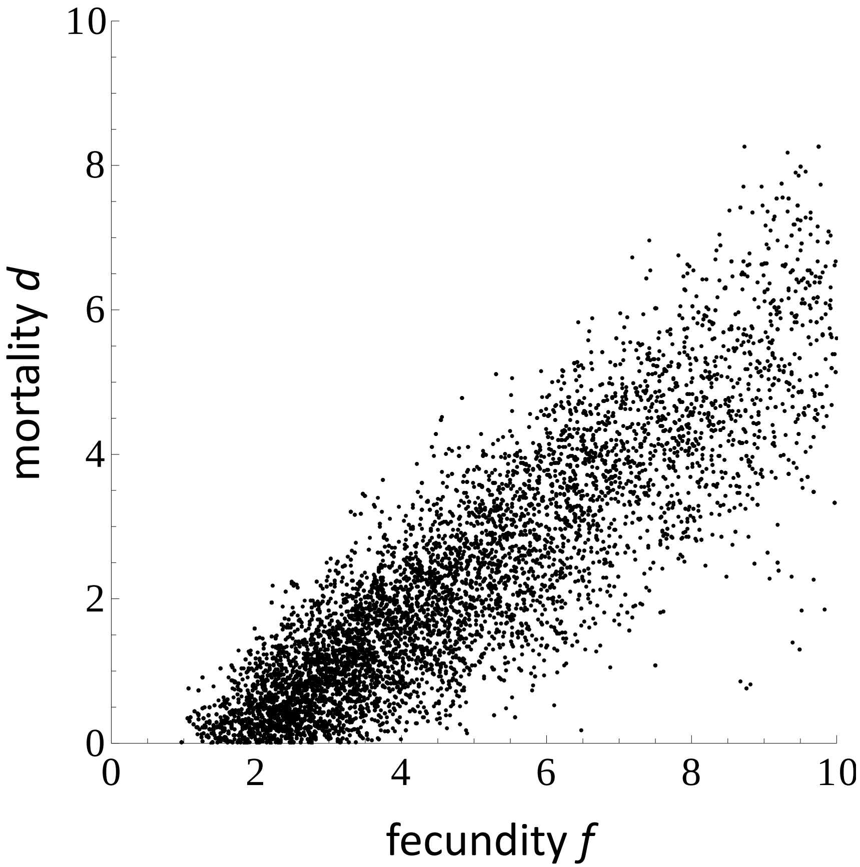
posterior mean (95% credibility interval)

3.66 (1.48-9.22) 1.42 (0.06-5.88) 48.1 (41.6-58.6)



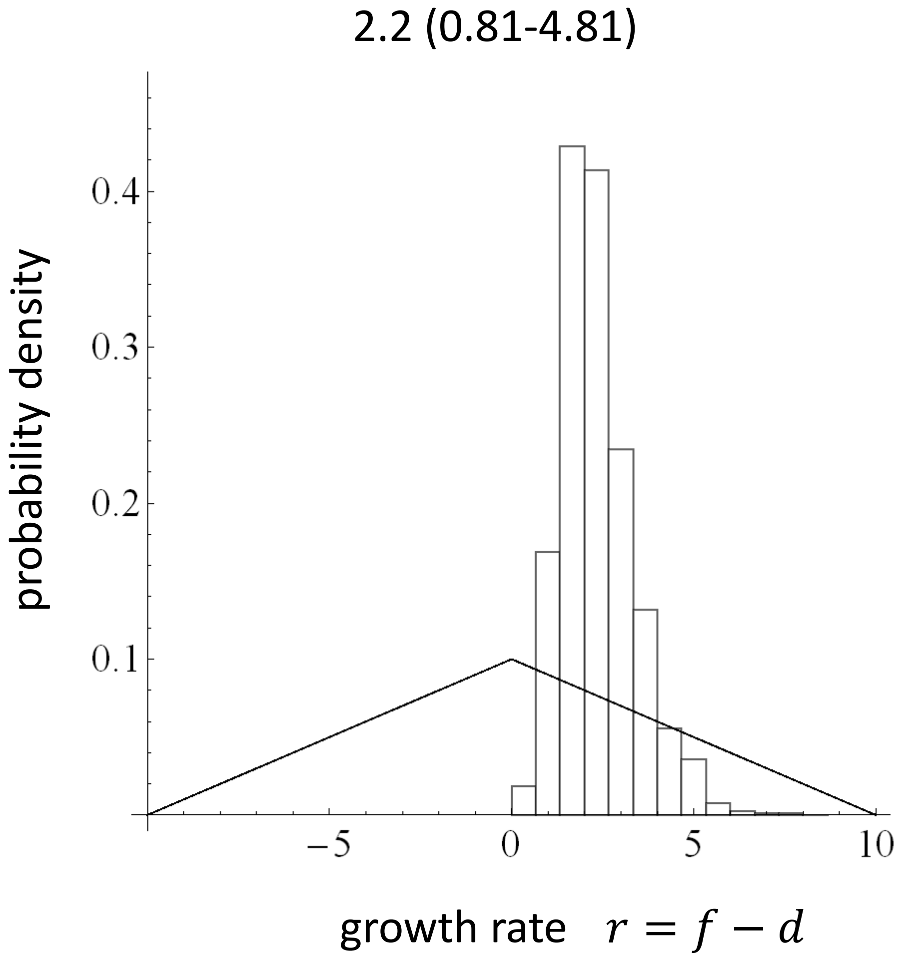
$$\theta^{\text{TRUE}} = (f, d, K) = (3,1,50)$$

The joint posterior distribution



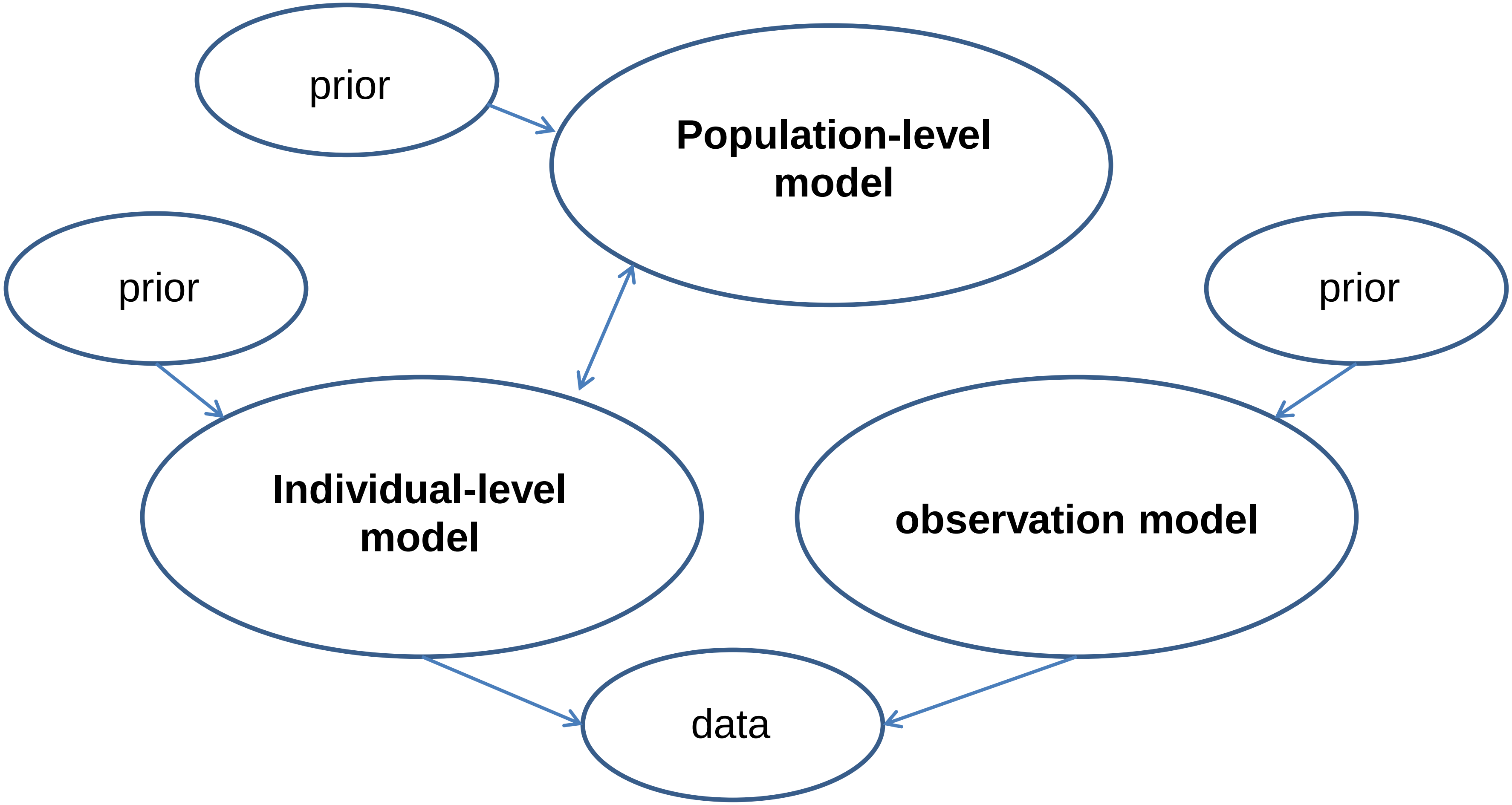
$$\theta^{\mathrm{TRUE}} = (f, d, K) = (3,1,50)$$

The posterior distribution of a derived parameter

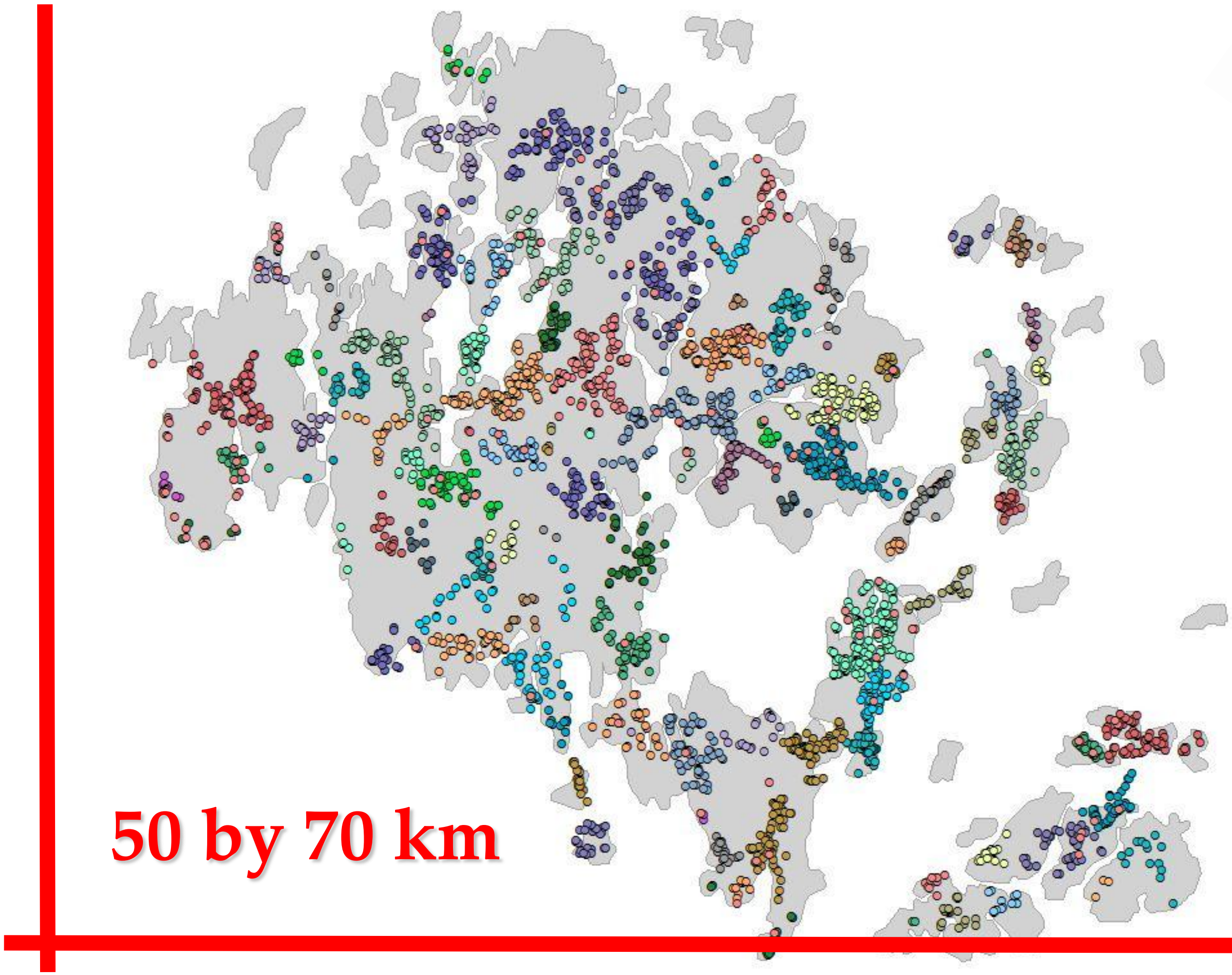


$$\theta^{\mathrm{TRUE}} = (f, d, K) = (3,1,50)$$

State-space models often have a hierarchical structure

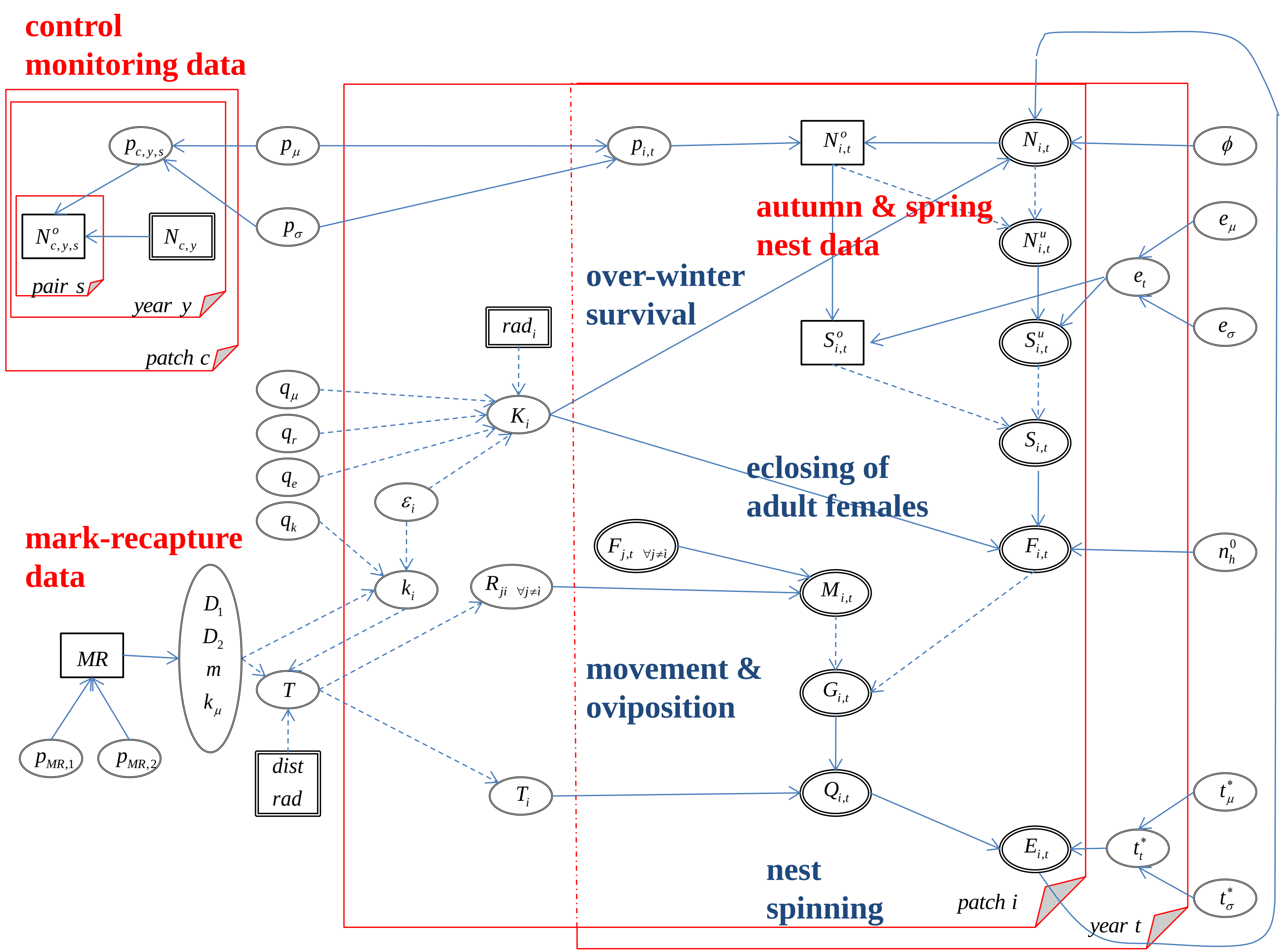


Example: Glanville fritillary metapopulation dynamics



prof. Ilkka Hanski

Building metapopulation dynamics from individual behavior



Harrison, P. J., Hanski, I. and Ovaskainen, O. 2012. Bayesian state-space modeling of metapopulation dynamics in the Glanville fritillary butterfly. *Ecological Monographs* **81**, 581-598.

Part of the individual-based model

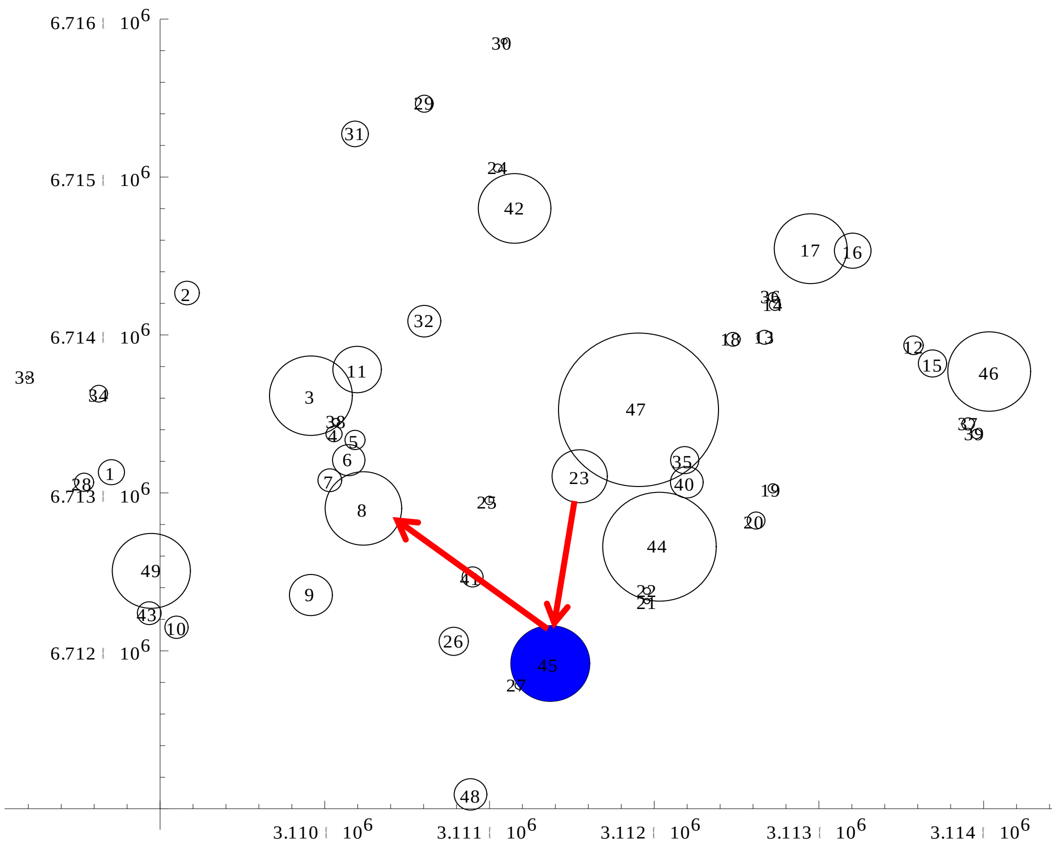
The amount of time all females spend in a patch

The number of larval groups that survive through the winter: $N \sim Bin(E, \phi)$

$$\lambda \longrightarrow E \longrightarrow N$$

The number of egg groups in the autumn: $E \sim Poisson(\lambda)$

Example of model prediction at the individual level



5 km

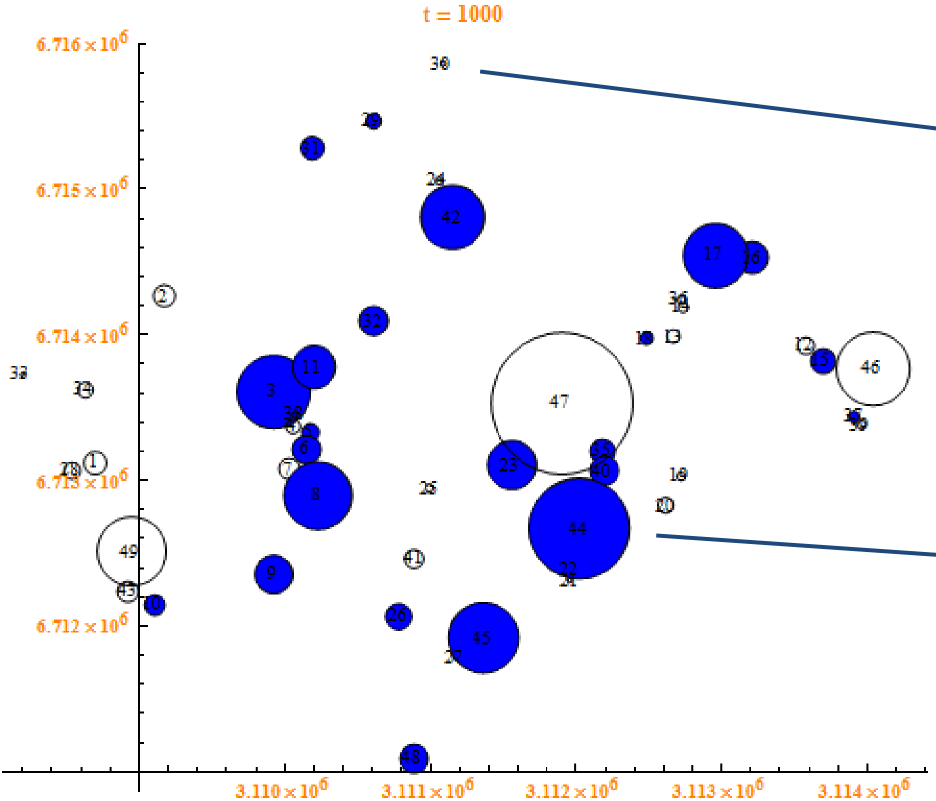
Circle = habitat patch

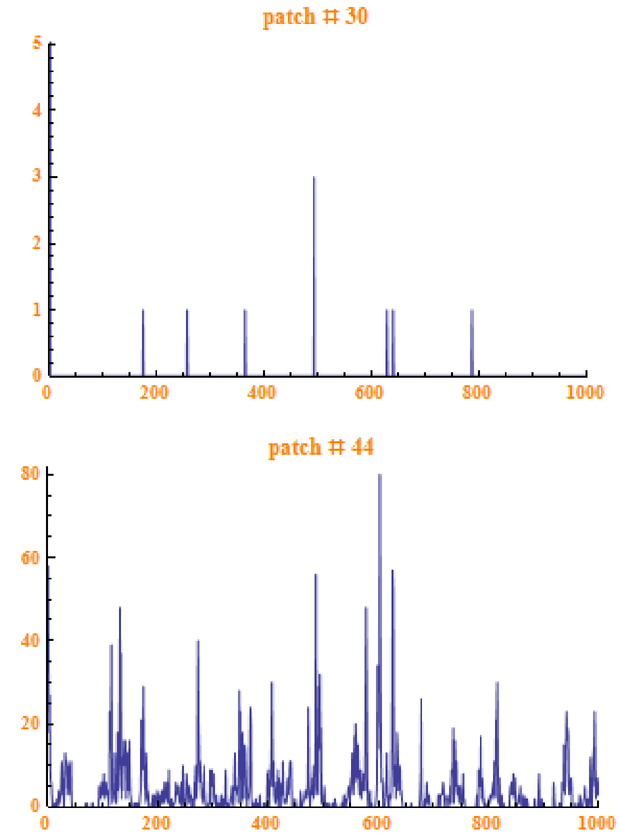
Life of a single female

LOCATION	TIME	EGG GROUPS
Patch 23	1.92444	0
Matrix	0.267237	0
Patch 45	1.19267	1
Matrix	0.607294	0
Patch 8	0.124928	0

Examples of model prediction at the population level

Number of larval groups



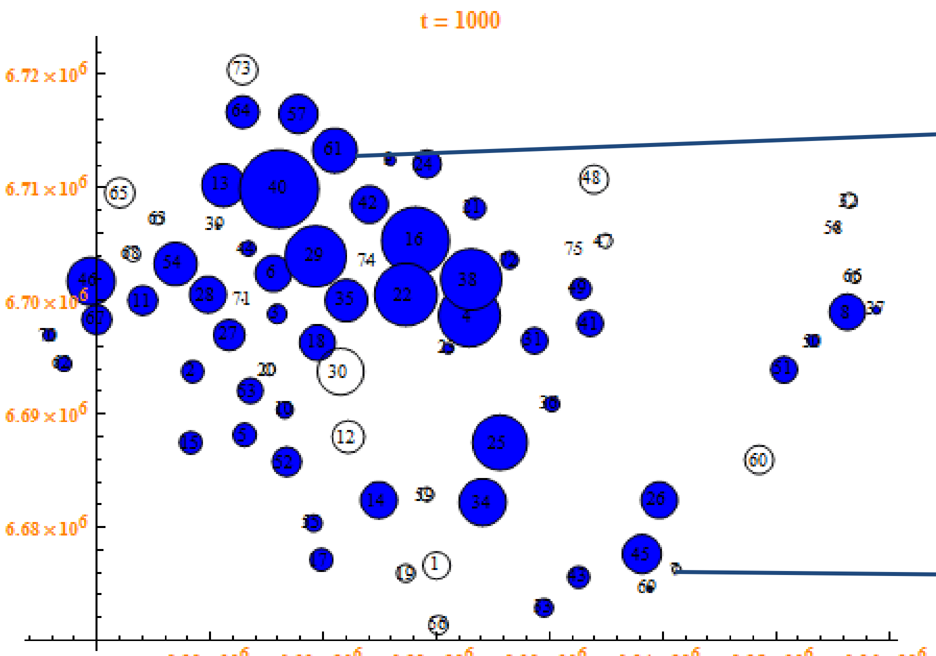


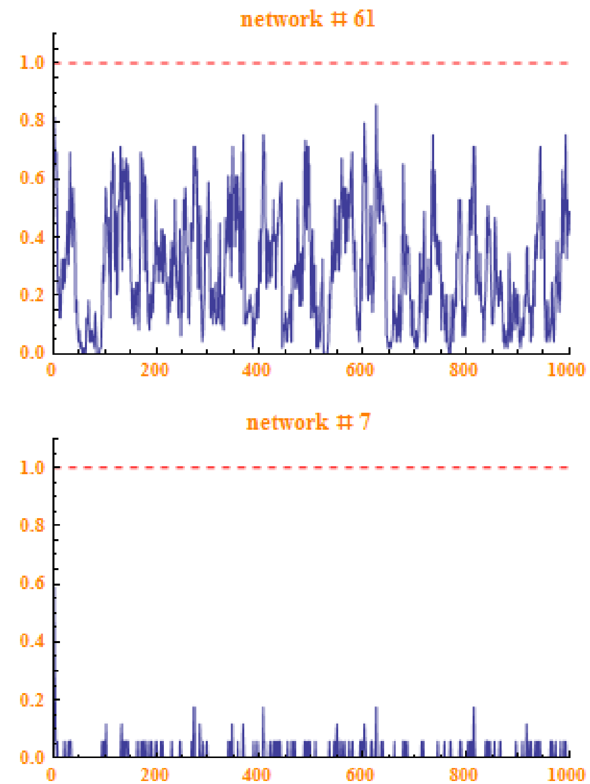
5 km

Circle = habitat patch

Examples of model prediction at the metapopulation level

Fraction of occupied patches





70 km

Circle = habitat patch network

Strategies for model validation

1. Do nothing, just trust the model (still most common option!)

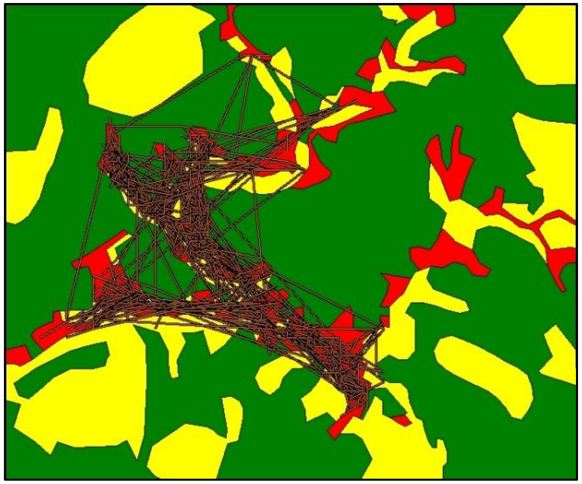
2. Fit model to data, then check if the model can reproduce the same data

3. Cross-validation: split the data into two parts. Use data 1 for fitting the model, and check if the model is able to reproduce data 2

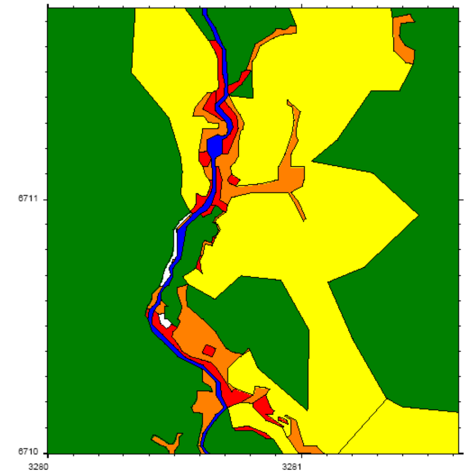
4. If you have data from different situations, see if the model fitted to situation 1 can reproduce the data collected from situation 2

Example of model validation (strategy 4)

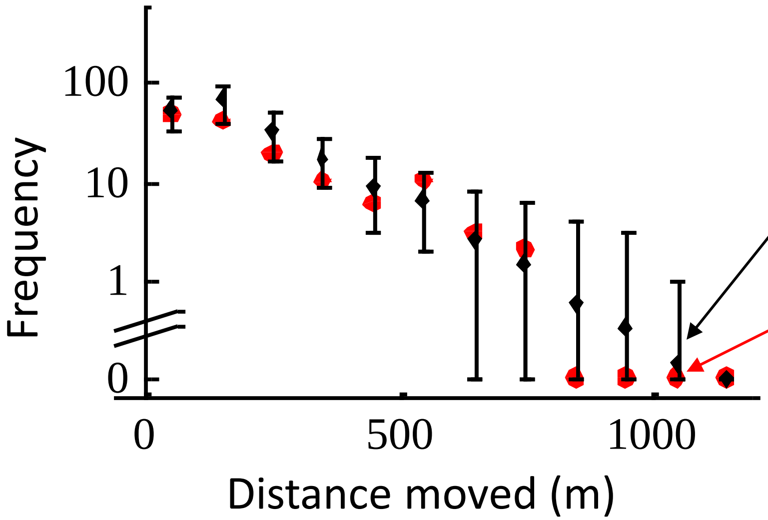
Landscape A



prediction



Landscape B



Model parameterized with data from Landscape A, prediction for Landscape B

Empirical data on Landscape B

Ovaskainen, O., Luoto, M., Ikonen, I., Rekola, H., Meyke, E. and Kuussaari, M. 2008. An empirical test of a diffusion model: predicting clouded apollo movements in a novel environment. *American Naturalist* **171,** 610-619.

Strategies for model selection

1. Try only one model and hope it fits nicely enough (still most common option!)

2. See which model reproduces the data (or preferably some independent data) using a summary statistic you best like / think is biologically relevant

3. Use formal model selection methods: AIC, BIC, DIC, Bayes factor, ...

L2: take home messages

- State-space models combine a process model with an observation model. They provide a very general framework of formulating and fitting movement models.

- State-space models can be visualized using a DAG (directed acyclic graph). DAG is a very useful way to illustrate how the components of the model link to each other.

- State-space models allow one to bring biological knowledge into statistical inference, parameterize dynamic models of movement, and to use data with missing observations.

- Fitting state-space models to data can be technically challenging. A great number of methods exist (essentially variants of MCMC approaches).

- All models are wrong, but some are still useful. Take model selection and validation seriously!