Fitting Models to Data in Ecology and Evolution

CMEE Masters

Samraat Pawar

Imperial College London

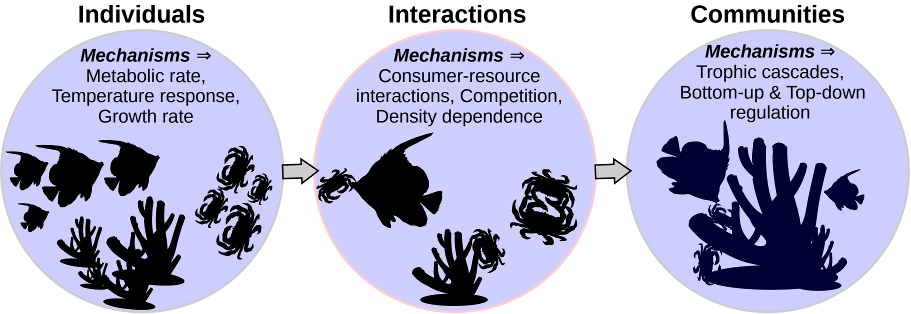
November 18, 2016

Mechanistic vs. phenomenological models

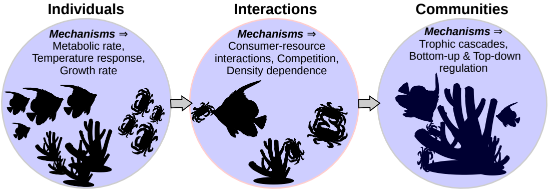
- Mechanistic models aim to explain the PROCESSES underlying observed patterns
- Empirical or phenomenological models show relationships between observed data (e.g. population size as a function of temperature or rainfall), but provide no insights into why they are related

Mechanistic vs. phenomenological models

- Mechanistic models aim to explain the PROCESSES underlying observed patterns
- Empirical or phenomenological models show relationships between observed data (e.g. population size as a function of temperature or rainfall), but provide no insights into why they are related

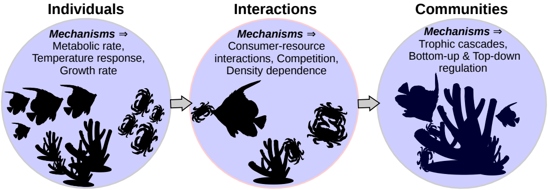


What are Mechanisms?



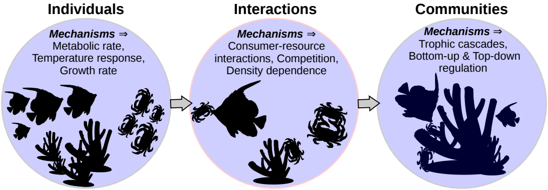
- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.

WHAT ARE MECHANISMS?



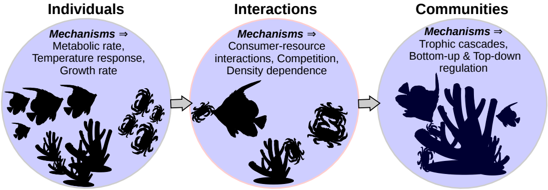
- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.
- For example, insect invasions, outbreaks and spread (http://www.sandyliebhold.com/pubs/science_DC1/) — papers in your `Readings` directory; (http://www.imperial.ac.uk/qmee-cdt/supervisors/)

What are Mechanisms?



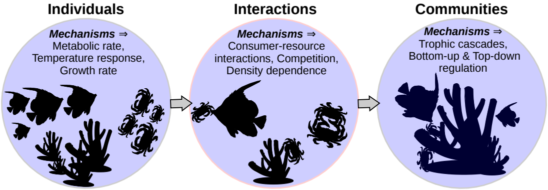
- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.
- For example, insect invasions, outbreaks and spread (http://www.sandyliebhold.com/pubs/science_DC1/) — papers in your `Readings` directory; (http://www.imperial.ac.uk/qmee-cdt/supervisors/)
  - Why the cycles?, Why the travelling waves?

WHAT ARE MECHANISMS?



- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.
- For example, insect invasions, outbreaks and spread (http://www.sandyliebhold.com/pubs/science_DC1/) — papers in your `Readings` directory; (http://www.imperial.ac.uk/qmee-cdt/supervisors/)
  - Why the cycles?, Why the travelling waves?

WHAT ARE MECHANISMS?



- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.
- For example, insect invasions, outbreaks and spread (http://www.sandyliebhold.com/pubs/science_DC1/) — papers in your `Readings` directory; (http://www.imperial.ac.uk/qmee-cdt/supervisors/)
  - Why the cycles?, Why the travelling waves? What mechanisms operate? (budmoth/parasitoid interaction? (budmoth/food quality interaction?) Are these truly mechanisms?
- Another example, disease outbreaks (Papers in your `Readings` directory)

What are Mechanisms?

- Somewhat subjective!

WHAT ARE MECHANISMS?

- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1-\frac{N_t}{k}\right)} \tag{1}$$

What are Mechanisms?

- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1-\frac{N_t}{k}\right)} \quad (1)$$

- What is the mechanism? — Density dependence through scramble competition (Brannstrom & Sumpter 2005)

What are Mechanisms?

- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1-\frac{N_t}{k}\right)} \tag{1}$$

- What is the mechanism? — Density dependence through scramble competition (Brannstrom & Sumpter 2005)
- If the Ricker model and another model with contest competition were compared with data — some would call it mechanistic modelling because one is trying to get at the underlying mechanism, scramble or contest competition

WHAT ARE MECHANISMS?

- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1-\frac{N_t}{k}\right)} \tag{1}$$

- What is the mechanism? — Density dependence through scramble competition (Brannstrom & Sumpter 2005)
- If the Ricker model and another model with contest competition were compared with data — some would call it mechanistic modelling because one is trying to get at the underlying mechanism, scramble or contest competition
- But is this REALLY mechanistic? What are *r* and *k* really?

What are Mechanisms?

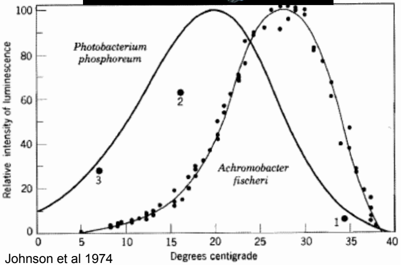
- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1 - \frac{N_t}{k}\right)} \tag{1}$$

- What is the mechanism? — Density dependence through scramble competition (Brannstrom & Sumpter 2005)
- If the Ricker model and another model with contest competition were compared with data — some would call it mechanistic modelling because one is trying to get at the underlying mechanism, scramble or contest competition
- But is this REALLY mechanistic? What are *r* and *k* really?
- Many (including yours truly!) now argue that we have not progressed far enough because the first level has been ignored!

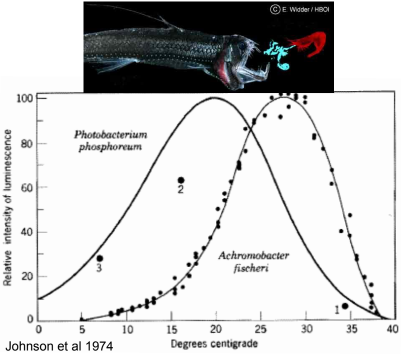
AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM

AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM



Johnson et al 1974

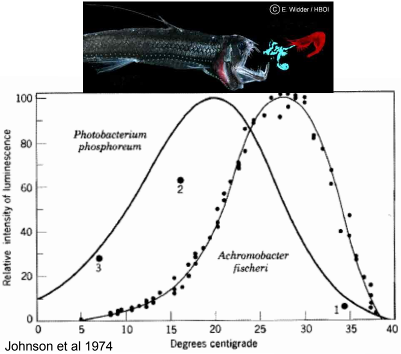
AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM



$$B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$$

T = temperature (K)
k = Boltzmann constant (eV K^{-1})
E = Activation energy (eV)
T_{pk} = Temperature of peak performance
E_D = Deactivation energy (eV)
(J H van't Hoff 1884, S Arrhenius 1889)

AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM

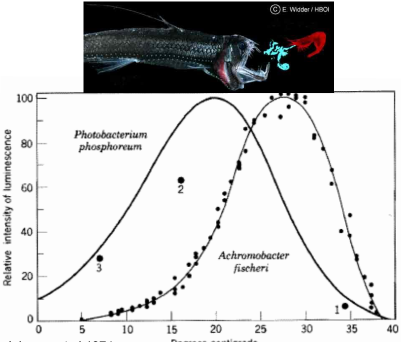


$$B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$$

T = temperature (K)
k = Boltzmann constant (eV K^{-1})
E = Activation energy (eV)
T_{pk} = Temperature of peak performance
E_D = Deactivation energy (eV)
(J H van't Hoff 1884, S Arrhenius 1889)

- Surely there is more to thermal responses?
  - Oxygen limitation
  - Complexity of metabolic network
  - Hormonal regulation

AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM



Johnson et al 1974

$$B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$$

T = temperature (K)
k = Boltzmann constant (eV K^{-1})
E = Activation energy (eV)
T_{pk} = Temperature of peak performance
E_D = Deactivation energy (eV)
(J H van't Hoff 1884, S Arrhenius 1889)

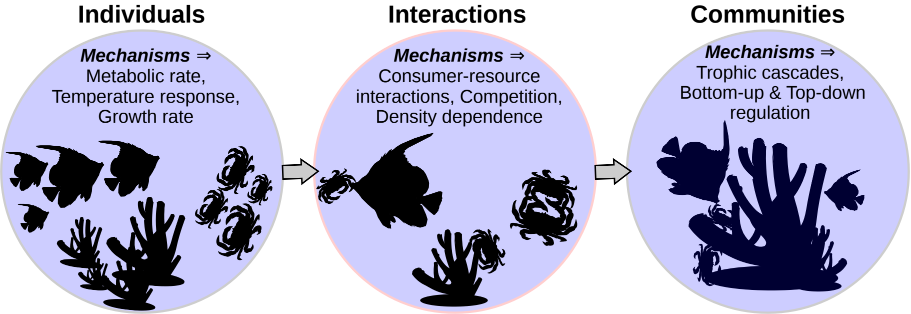
- Surely there is more to thermal responses?
  - Oxygen limitation
  - Complexity of metabolic network
  - Hormonal regulation
- *What about alternative models?*

MODELLING, AND FITTING MODELS TO DATA: WHAT'S THE BIG IDEA?

- If possible, use biological knowledge to construct models
- See if the models "agree well" with data
- Whichever model "agrees best" is most likely to have the right mechanisms
- That's the one that's best for predictions (e.g. population cycles), estimating rates (e.g. growth rates), etc.
- Don't use models you already know have the wrong mechanisms!
- Phenomenological models often perform better than mechanistic ones

MODELS: HOW TO BUILD THEM?

- It's an art, take practice (look at Levins' paper on the strategy of model building in biology)
- Build models one mechanism at a time — in biology, it means start at the right level of organization!
- Always consider a alternative that is more parsimonious, even if it is phenomenological (the TPC example: Sharpe-Schoolfield, or Polynomial?)!



MODELS: HOW TO BUILD THEM?

- For example, the Boltzmann-Arrhenius model is a good first try describe and uncover mechanisms underlying individual level rates
- The next step would be to include high-temperature effects (e.g., the Schoolfield model)
- The next step would be to include species interactions with temperature dependence of individuals (or go in an evolutionary direction!)

Next: A primer on mechanistic model fitting

Next: A primer on mechanistic model fitting

But first: A preview of the Miniproject (back to Notes)

FITTING MODELS TO DATA

Two common ways to do it:

- One-step forecasting (appropriate for discrete models)
- Ensemble fitting (appropriate for full time series or responses) – this is what you will be doing in NLLS

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)
4. Adjust the parameters to make the curve come closer to the data points. This the tricky part — you will use the Levenberg-Marquardt algorithm in the `lmfit` package in `python`

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)
4. Adjust the parameters to make the curve come closer to the data points. This the tricky part — you will use the Levenberg-Marquardt algorithm in the `lmfit` package in `python`
5. Adjust the parameters again so that the curve comes even closer to the points (rss decreases)

Ensemble fitting

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)
4. Adjust the parameters to make the curve come closer to the data points. This the tricky part — you will use the Levenberg-Marquardt algorithm in the `lmfit` package in `python`
5. Adjust the parameters again so that the curve comes even closer to the points (rss decreases)
6. Repeat 4–5

ENSEMBLE FITTING

There include maximum likelihood, bayesian methods, and Non-linear least squares (NLLS) optimization or fitting. Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)
4. Adjust the parameters to make the curve come closer to the data points. This the tricky part — you will use the Levenberg-Marquardt algorithm in the `lmfit` package in `python`
5. Adjust the parameters again so that the curve comes even closer to the points (rss decreases)
6. Repeat 4–5
7. Stop simulations when the adjustments make virtually no difference to the rss

ENSEMBLE FITTING

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit

Ensemble fitting

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit
- Then compare multiple models (e.g., Schoolfield vs. cubic)

ENSEMBLE FITTING

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit
- Then compare multiple models (e.g., Schoolfield vs. cubic)

ENSEMBLE FITTING

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit
- Then compare multiple models (e.g., Schoolfield vs. cubic)

The precise parameter values you obtain will depend in part on the initial values chosen and the stopping criteria –

ENSEMBLE FITTING

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit
- Then compare multiple models (e.g., Schoolfield vs. cubic)

The precise parameter values you obtain will depend in part on the initial values chosen and the stopping criteria – so different programs will not always give exactly the same results

Comparing models

- It's all about the "Likelihood" of a model:
- That is, the likelihood of a set of parameter values (of a model), θ, given outcomes x, equals the probability of those observed outcomes given those parameter values, that is,

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

COMPARING MODELS

The easiest thing to do for you is to use information theory (including AIC and BIC) to compare models.
Both use estimated likelihoods of a model

COMPARING MODELS

The easiest thing to do for you is to use information theory (including AIC and BIC) to compare models.
Both use estimated likelihoods of a model
The lower the AIC or BIC, the better. This is how you calculate them (using python syntax):

COMPARING MODELS

The easiest thing to do for you is to use information theory (including AIC and BIC) to compare models.
Both use estimated likelihoods of a model
The lower the AIC or BIC, the better. This is how you calculate them (using python syntax):

- residuals = Observations - Predictions
- rss = sum(residuals ** 2)
- Then, AIC is n * log((2 * pi) / n) + n + 2 + n * log(rss) + 2 * k (*what is n and k?*)
- And BIC is n + n * log(2 * pi) + n * log(rss / n) + (log(n)) * (k + 1)
- For both AIC and BIC, If model **A** has AIC lower by 2-3 or more than model **B**, it's better — Differences of less than 2-3 don't really matter

Comparing models

The easiest thing to do for you is to use information theory (including AIC and BIC) to compare models.
Both use estimated likelihoods of a model
The lower the AIC or BIC, the better. This is how you calculate them (using python syntax):

- residuals = Observations - Predictions
- rss = sum(residuals ** 2)
- Then, AIC is n * log((2 * pi) / n) + n + 2 + n * log(rss) + 2 * k (*what is n and k?*)
- And BIC is n + n * log(2 * pi) + n * log(rss / n) + (log(n)) * (k + 1)
- For both AIC and BIC, If model **A** has AIC lower by 2-3 or more than model **B**, it's better — Differences of less than 2-3 don't really matter

Also note that:

- R^2 = 1 - (rss/tss), where tss is total sum of squares:
  tss = sum((Observations - mean(Predictions)) ** 2)
  (a useful measure of goodness of fit – you should report it)

READINGS

- Levins, R. 1966 The strategy of model building in population biology. Am. Sci. 54, 421–431.
- Johnson, J. B. & Omland, K. S. 2004 Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.
- For the suggested fitting TPCs project: Papers in the `Temperature_response_papers` directory, but especially Schoolfield, R. M., Sharpe, P. J. & Magnuson, C. E. 1981 Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory. J. Theor. Biol. 88, 719–31.