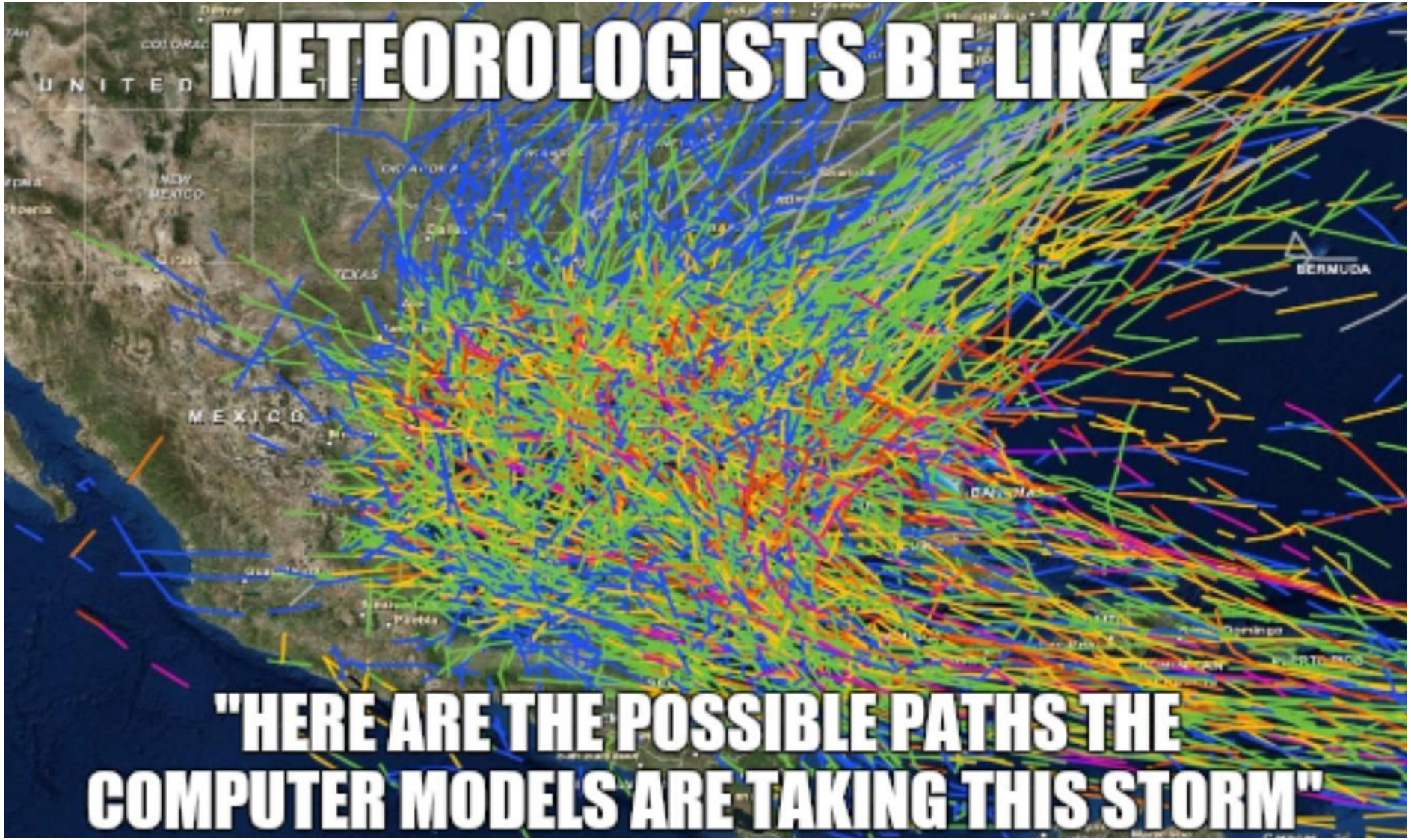
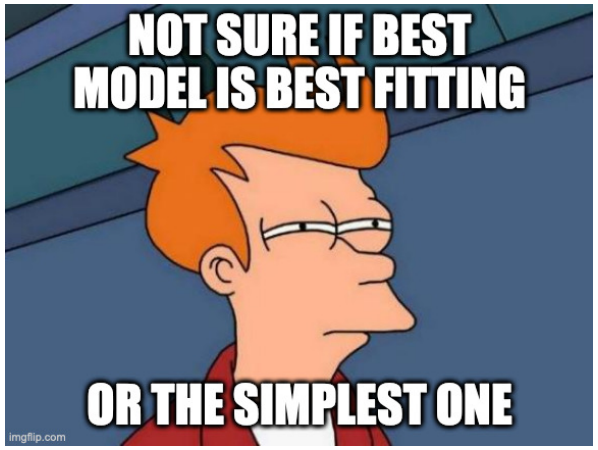
Model Selection

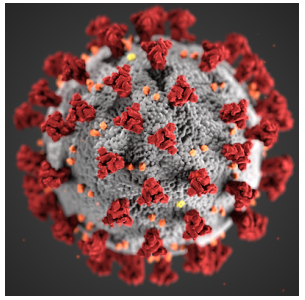


What is model selection?

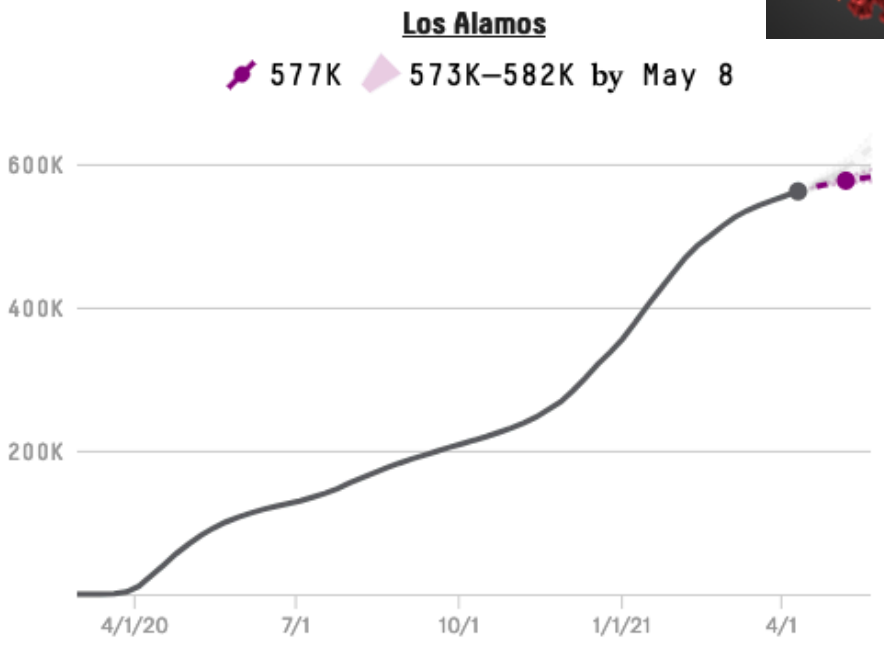
- Process in choosing a statistical model from a set of models based on your data
- Rule of thumb: the simplest model balanced with a decent goodness of fit is generally the best model



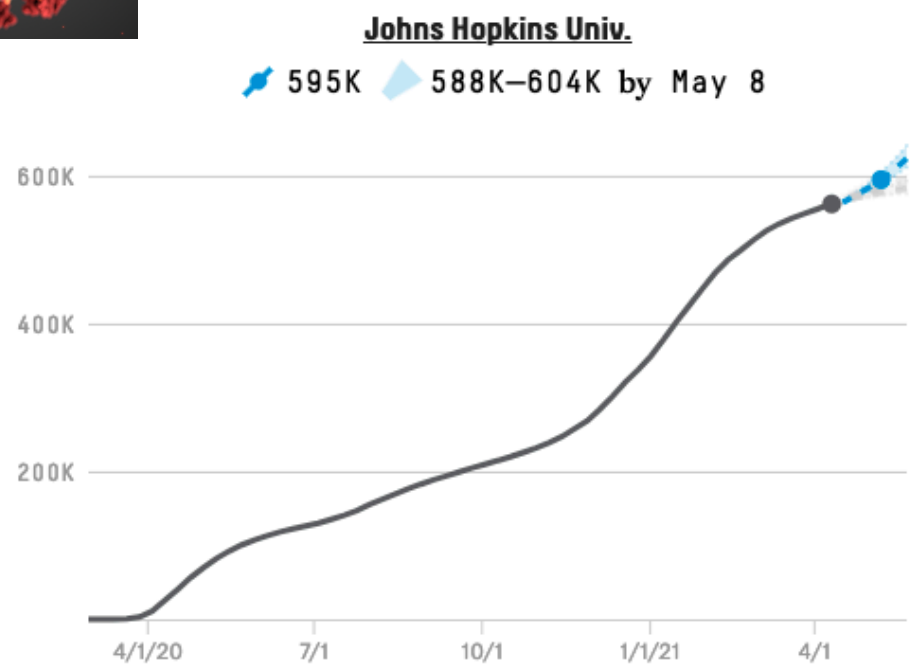
COVID model selection



- Competing models
- Different assumptions
  - True number of infections
  - Stay-at-home orders
- Predictability is key
- 562,066 Deaths as of April 11th
- By May 8th models predict:
  - Los Alamos: 15,000 more
  - Johns Hopkins:
  - 33,000 more



This model assumes that there will continue to be interventions, such as stay-at-home orders, but it does not specifically assume what those interventions will be. Instead, it considers various possible interventions to arrive at its forecast, which typically results in wider prediction intervals than a model with stricter assumptions.

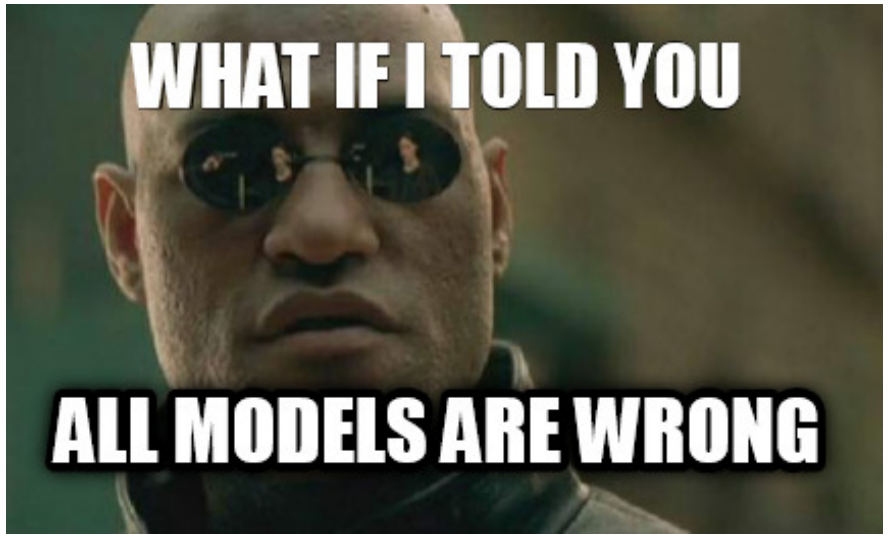


This model incorporates information about stay-at-home orders and assumes that the effectiveness of social distancing measures in a given state decreases by roughly 25 percent after those orders are lifted.

https://projects.fivethirtyeight.com/covid-forecasts/

Semantics and philosophy

- Model selection can get messy and opinionated
- *“All models are wrong, but some are useful”*
- Word choice: “best” vs “most plausible” vs “most parsimonious”



Recommended read:

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

CONCEPTS & SYNTHESIS

A practical guide to selecting models for exploration, inference, and prediction in ecology

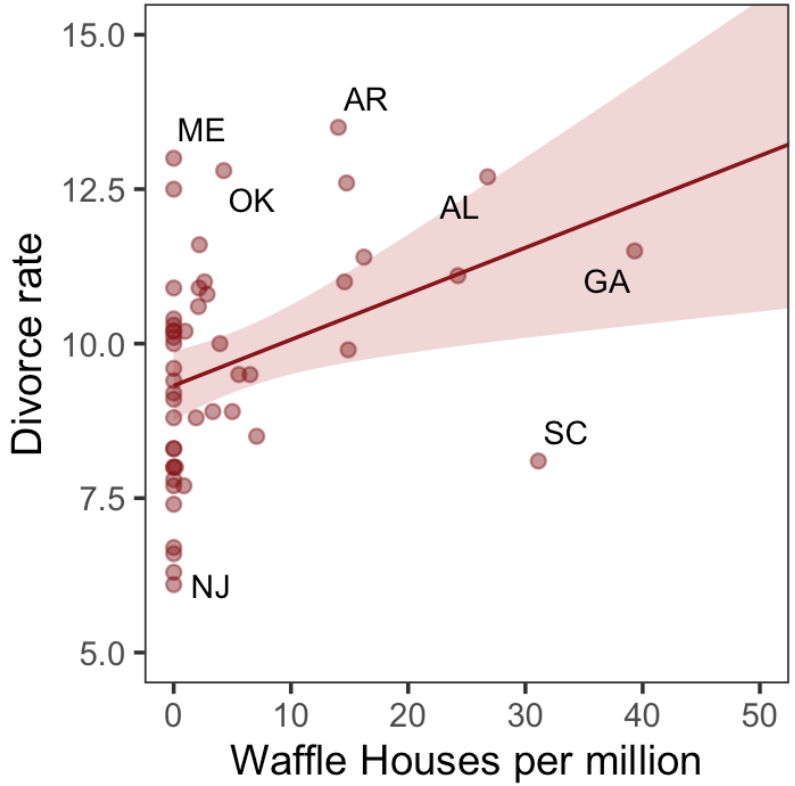
Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, Peter B. Adler

First published: 12 March 2021 | https://doi.org/10.1002/ecy.3336

- Three main frameworks of model selection:
  - Exploration, Inference, Prediction

Exploration

- Trade-off: Thoroughness vs spurious relationships.
- We don’t want to lose out on any important relationships so we “cast a wide net” – prone to false discoveries (type-II errors)
- How to deal with this?
  - Only consider plausible relationships based on your knowledge
  - Adjust p–values
  - BE HONEST, it’s ok to have an exploratory study



- Brian McGill’s blogpost:https://dynamicecology.wordpress.com/2013/10/16/in-praise-of-exploratory-statistics/

If exploratory statistics weren’t treated like the crazy uncle nobody wants to talk about and everybody is embarrassed to admit being related to, science would be much better off.

Inference

- Trying to confirm hypotheses about the system
- Hypotheses are structured as competing models (Read Chamberlin 1890)
  - BUT if you're testing 20 different models/hypotheses → question whether this is turning into a exploratory expedition.

Chamberlin 1890, reprinted in 1965

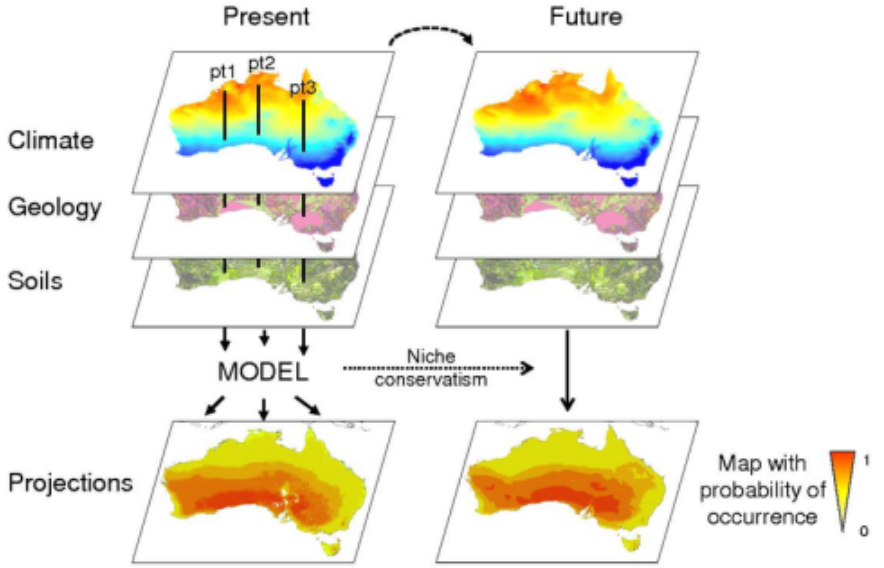
The Method of Multiple Working Hypotheses

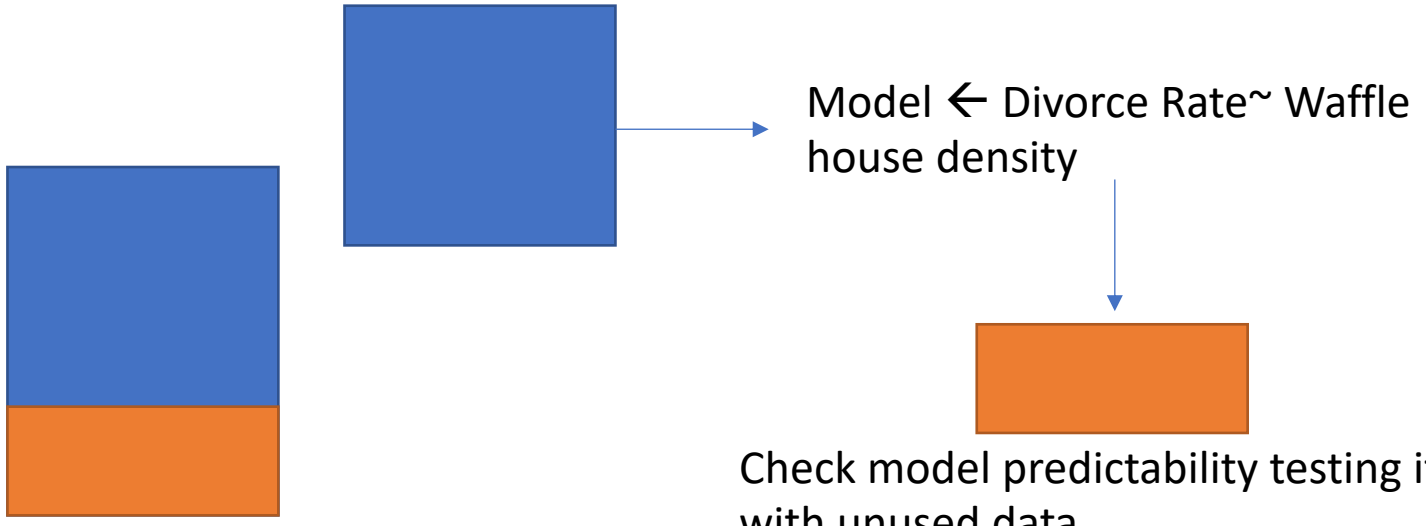
With this method the dangers of parental affection for a favorite theory can be circumvented.

T. C. Chamberlin

Prediction

- Accuracy and predictability are key
- Often a blend of both inference and exploration
- Best model for prediction may not be the best model for inference
- DISTINGUISHING FEATURE is the testing of the model using independent data validation!





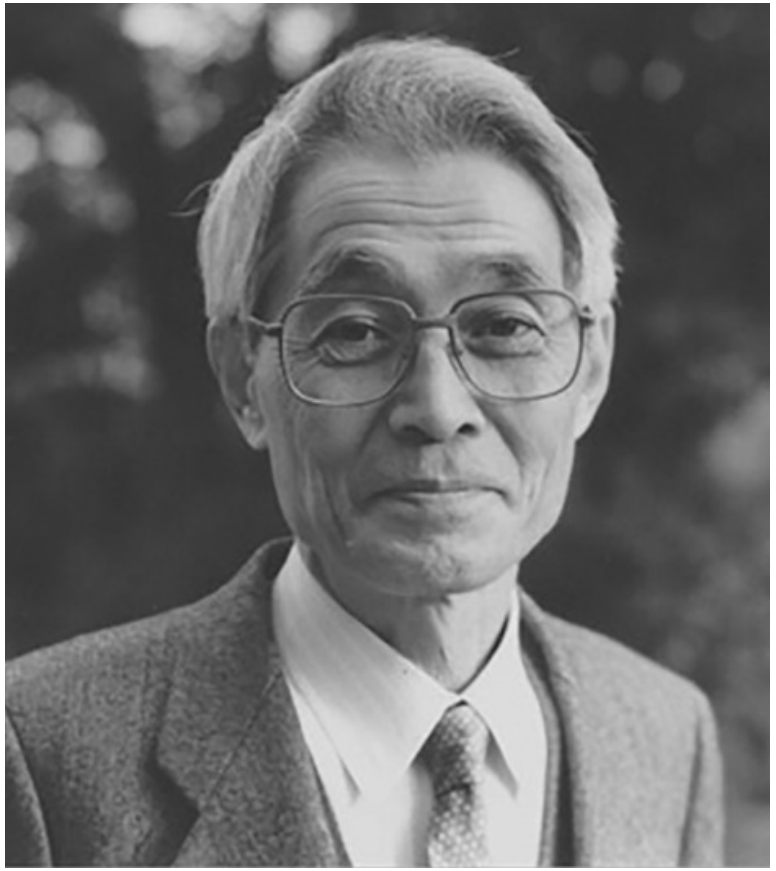
Inference Part 1: Null hypothesis testing

- Example: Likelihood ratio test based on t-values that tests one model vs another model
  - Limited to two models
  - Models must be nested
    - Y~A+b
    - Y~A

- Null hypothesis testing works well for **experiments**
- Less common when we move to **observational or empirical studies** (AIC comes to play)

Inference Part 2: Akaike Information Criterion (AIC)

- Developed by Hirotogu Akaike
- A metric of **relative** model quality for a given set of data
- $AIC = -2 * \ln(model\ likelihood)$ +2K
  - K = number of parameters in the model
  - Lower AIC is better
  - Model likelihood is a measure of model fit
- Don't need to worry about nestedness, get rid of the p-value problem, but there still are arbitrary cutoffs (explained later)
- **Note that your data and your response variable have to be the same for AIC selection!**



How does AIC work?

$$AIC = -2 * \ln(model\ likelihood) + 2K$$

- AIC gets **lower** when:
  - Model likelihood increases and likelihood is a measure of model fit
  - K (number of parameters is low)

```
data("mtcars")
m1<-lm(data=mtcars, mpg~hp)
m2<-lm(data=mtcars, mpg~wt)
```

```
> AIC(m1)
[1] 181.2386
> logLik(m1)
'log Lik.' -87.61931 (df=3)
```

```
> -2*logLik(m1) + 2* 3
'log Lik.' 181.2386 (df=3)
```

```
> AIC(m2)
[1] 166.0294
> logLik(m2)
'log Lik.' -80.01471 (df=3)
```

```
> -2*logLik(m2) + 2* 3
'log Lik.' 166.0294 (df=3)
```

AICc

- Small sample sized can bias AIC to select for models with too many parameters
- To account for this AICc (Akaike Information Criterion Corrected for Small Sample Sizes) was developed:

- AICc = AIC + $\frac{2k^2+2k}{n-k-1}$
  - n = sample size
  - k = no. of parameters

- Generally use AICc with sample size <30 (I personally use it with sample size <50)

AIC opinions

- Many people have differing opinions on AIC usage
- My opinion: A lot of it depends on your philosophy of the scientific process
- Great read: Dynamic Ecology blogpost about this

Why AIC appeals to ecologist's lowest instincts

It is my sense of the field that AIC (Akaike information criteria) has moved past bandwagon status into a fundamental and still increasingly used paradigm in how ecologists do statistics. For some quick and dirty evidence I looked at how often different core words

https://dynamicecology.wordpress.com/2015/05/21/why-aic-appeals-to-ecologists-lowest-instincts/

2) You never assess how good your best AIC model is – in my experience the "best" model often has an r2 of 0.10 yet with AIC we get to all slap each other on the back and congratulate ourselves on finding the best model.

SKEPTIC!

HOPEFUL! But in some ways sad…

AIC will only be good for the advance of science (and ecology) if we use it to advance the development and predictability of theory.

Very nice text, thank you for posting it. AIC has become a plague in Ecology. As you said, the problem is not the tool itself, but the misuse that has become commonplace in the past years. As a reviewer of manuscripts and theses, to my experience, something like 95% of the studies use AIC and model/variable selection as a replacement for deduction, induction, and abduction. Some model

CRITIC!

CONSPIRACY!

I think the problem is two-fold.

1) AIC was sort of "marketed" as a panacea for all of your p-value woes. I remember attending a seminar given by David Anderson which felt a lot like a sales pitch. He even got into a debate with one of the other attendees about the AIC>2/

So why AIC?

- Because we as humans like to rank things and AIC justifies ranking things and makes things easier! – one of the arguments from the blogpost

But also:

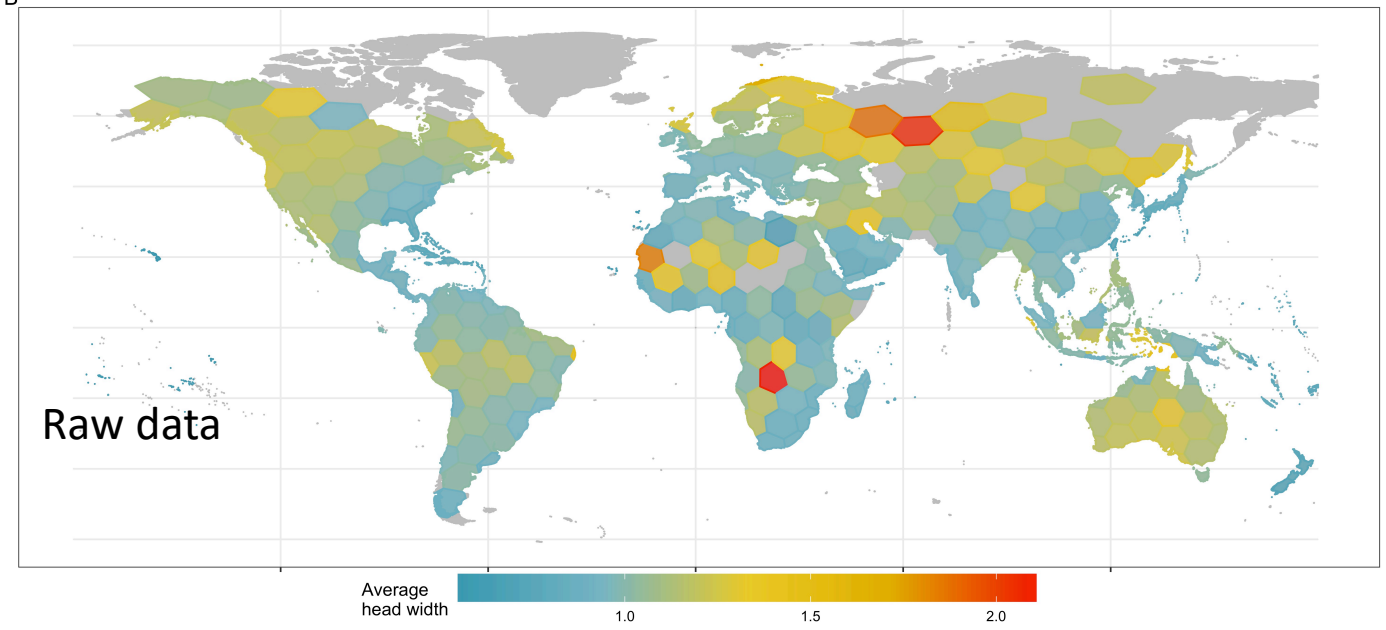
- AIC is commonly used and accepted
- Works well within a multiple competing hypothesis framework
- A step forward in getting away from the *p*-value problem

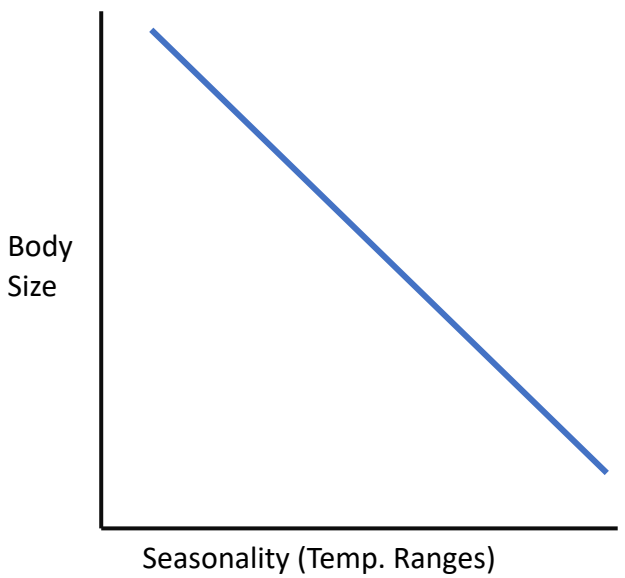
Example with AIC

- I have the average body size of an ant community
- Here is what the data looks like:

	cell	geo_avg	SR	ATR	NPP_mean	MAT	lat	lon
1	1	1.0912129	56	26.58891	550.4055176	4.7616038	5.894358e+01	11.83942092
2	9	1.6880459	7	23.71292	332.4631958	1.1910487	6.948915e+01	19.20741261
3	10	1.2090465	50	29.55614	420.8560791	0.9827181	6.443302e+01	14.98824540
4	12	1.2123414	15	34.34307	418.0303955	-0.6845881	6.092071e+01	-156.22049831
5	13	1.0968534	22	46.19601	287.1639099	-7.2203846	6.706155e+01	-149.96528280
6	22	1.1284582	19	38.76278	330.6351624	-3.7842741	6.156315e+01	-141.02015659
7	23	1.0858013	18	48.80699	369.2464294	-7.6180682	6.613260e+01	-131.53257329
8	32	1.2057155	90	38.47393	421.6273193	-1.6192271	5.864949e+01	-128.02748195

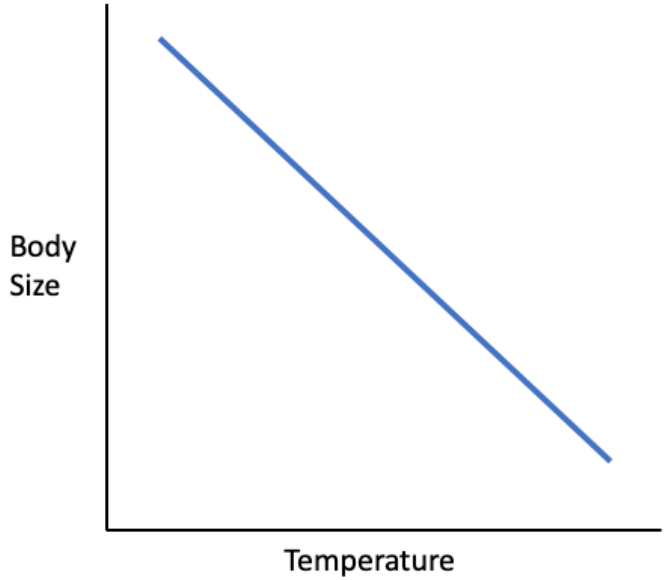
B





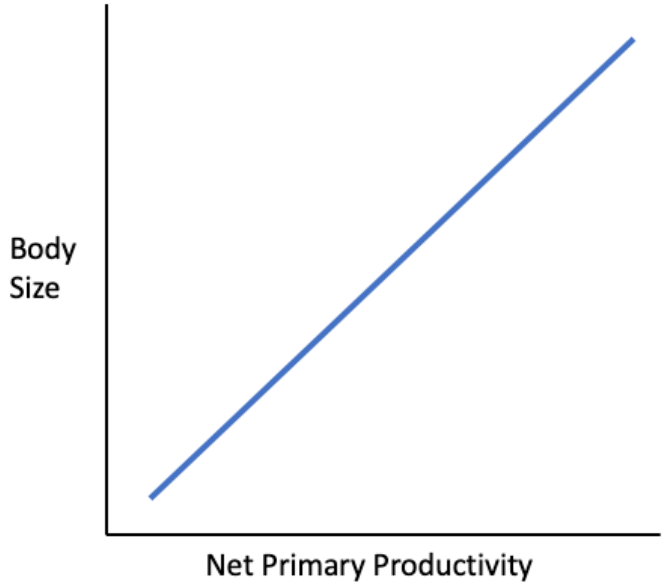
Seasonality hypotheses:

- Growth can be limited by the length of growing season (Mousseau 1997, Chown & Gaston 2010). As seasonality increases, the growing season should decrease leading to smaller body-sizes



Thermoregulatory hypothesis:

- Colder temperatures are likely to hold larger body sizes due to thermal-tolerance as heat is lost more slowly with larger body-sized animals (Stevenson 1985).



Resource-Availability hypothesis:

- Animals grow larger in high-resource environments (Atkinson & Sibly 1997). Areas with higher net primary productivity are likely to hold larger body sizes.

- **Sampling effect hypothesis:** The no. of species impacts the body size
- **Null hypothesis:** There is no effect on body size

- **<u>QUESTION ASKED:</u>**

Which hypothesis best explains the distribution of body size in ant assemblages?

Each hypothesis as a model

```
Seasonality <- lm(log(geo_avg) ~ ATR, data = model_df_haver)

Sampling <- lm(log(geo_avg) ~ SR, data = model_df_haver)

resource_energy <- lm(log(geo_avg) ~ NPP_mean, data = model_df_haver)

Temperature <- lm(log(geo_avg) ~ MAT, data = model_df_haver)

NULLMod <- lm(log(geo_avg) ~ 1, data = model_df_haver)
```

Calculate AIC using ‘bbmle’

```
> AICtab(seasonality, sampling, resource_energy,
+        Temperature, NULLMod, weights = T, delta = T, base =T)
                 AIC    dAIC   df weight
seasonality      -447.5    0.0 4  1
Temperature      -421.1   26.4 4  <0.001
resource_energy -419.5   28.0 4  <0.001
sampling         -299.8  147.7 4  <0.001
NULLMod          -290.9  156.6 3  <0.001
```

Calculate AIC using ‘bbmle’

```
> AICtab(seasonality, sampling, resource_energy,
+        Temperature, NULLMod, weights = T, delta = T, base =T)
                 AIC
seasonality     -447.5
Temperature     -421.1
resource_energy -419.5
sampling        -299.8
NULLMod         -290.9
```

- **Raw AIC values (lower is better). Note they can be negative!**

Calculate AIC using 'bbmle'

```
> AICtab(seasonality, sampling, resource_energy,
+        Temperature, NULLMod, weights = T, delta = T, base =T)
                AIC    dAIC
seasonality     -447.5    0.0
Temperature     -421.1   26.4
resource_energy -419.5   28.0
sampling        -299.8  147.7
NULLMod         -290.9  156.6
```

- **Difference in AIC between models**
- **AIC difference of <2 would mean models are comparable with one another. NOTE: this is an arbitrary cutoff!**

Calculate AIC using ‘bbmle’

```
> AICtab(seasonality, sampling, resource_energy,
+        Temperature, NULLMod, weights = T, delta = T, base =T)
                AIC    dAIC   df
seasonality     -447.5    0.0 4
Temperature     -421.1   26.4 4
resource_energy -419.5   28.0 4
sampling        -299.8  147.7 4
NULLMod         -290.9  156.6 3
```

- **Df represents the number of parameters/ degrees of freedom in each model.**

Calculate AIC using ‘bbmle’

```
> AICtab(seasonality, sampling, resource_energy,
+        Temperature, NULLMod, weights = T, delta = T, base =T)
                AIC    dAIC   df weight
seasonality     -447.5    0.0 4  1
Temperature     -421.1   26.4 4  <0.001
resource_energy -419.5   28.0 4  <0.001
sampling        -299.8  147.7 4  <0.001
NULLMod         -290.9  156.6 3  <0.001
```

- **Weight: The probability that the model is best from the set of competing models**

Interpret most plausible model

2 mm