Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

Wrangling outliers to avoid telling lies

— A primer for exploring, cleaning & filtering data prior to analysis

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

INTRODUCTION

You may have heard of the expression “Damn Lies and Statistics” and perhaps read the book ‘Damn Lies and Statistics: Untangling numbers from the Media, Politicians, and Activists’ written by Joel Best, released in 2001 and updated in 2012. The book has been aptly described as “a classic guide to understanding how numbers can confuse us.”

- Are you familiar with the expression?
- What do you think about when you hear it?
- How might it apply in the context of fisheries biology?

As scientists working with numbers, we have the capacity to confuse and mislead if we are not careful with how we analyse and interpret data. This can be avoided by ensuring that we undertake some preliminary steps to explore our data, prepare it for analysis and then select an analytical method that is appropriate for the data to answer the question posed.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

ABOUT THIS MODULE

It is important for us to know how and when to prepare and adjust data to avoid providing scientific advice that could potentially lead to inappropriate decisions about managing fisheries resources. Taking these preliminary steps can avoid costly repetition of analysis when something about the statistical outputs indicates that there may be a flaw in the information. It is also important to avoid unnecessarily discarding information that may be providing an important signal, even of it would be more convenient to omit.

By the end of this training module, you will be able to:

- describe some of the common issues and types of data observed in datasets,
- undertake an exploratory data analysis (EDA),
- identify outliers and decide whether to retain, eliminate or minimise their effects,
- filter and ‘cleanse’ data prior to analysis,
- select an appropriate model statement for a statistical analysis to answer the central question, and
- describe several of the implications of omitting EDA & filtering,
- describe the limitations and pitfalls of this approach.

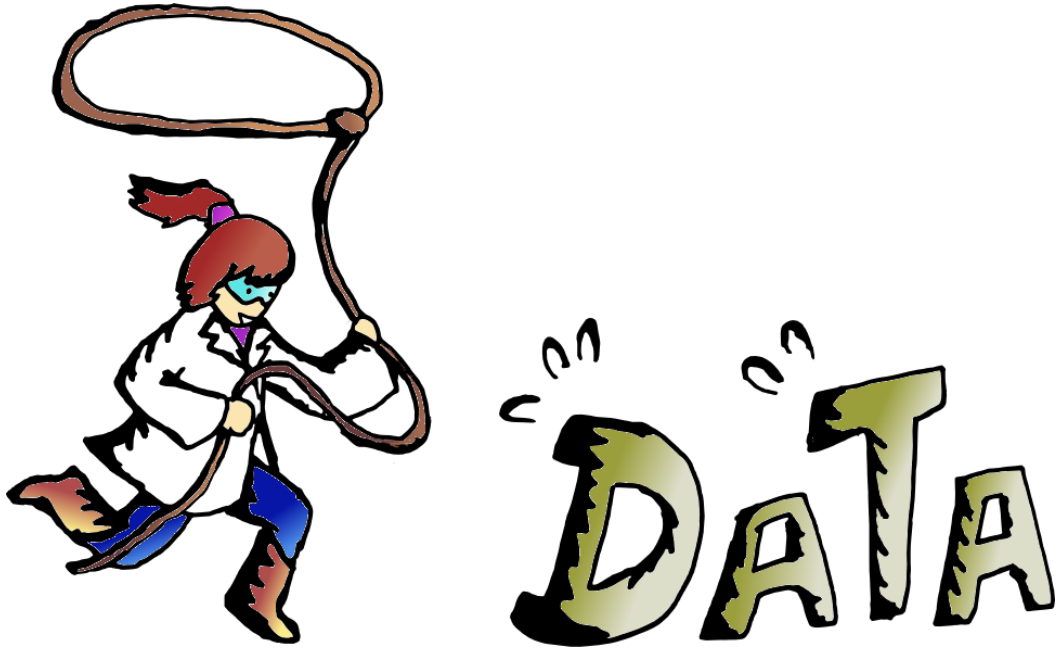
This will enable you to have confidence in your results from knowing that you applied a thorough and generally accepted process, prior to proceeding with your chosen analysis.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

WHAT IS DATA WRANGLING?

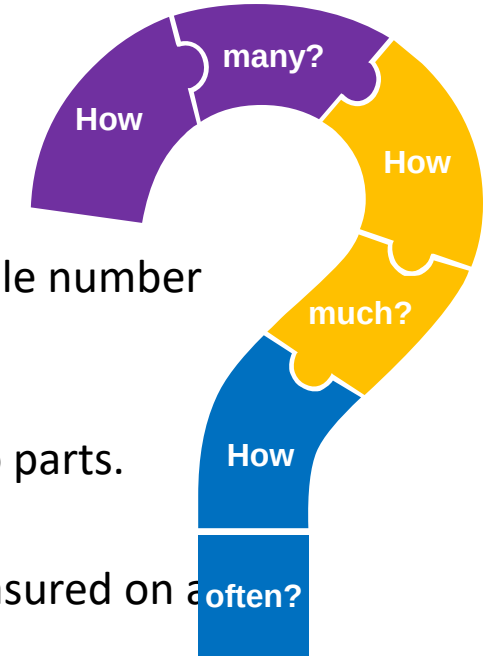
- ‘Data wrangling’ foregrounds the problems that prevent raw data from being effectively used in analytics;
- rigorous data cleaning and pre-processing is fundamental to preparing data for analytics;
- essential when converting raw data into actionable insights that align with the objectives of a stock status assessment.



Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

QUALITATIVE VS QUANTITATIVE DATA



Quantitative data

Answers key questions such as “how many, “how much” and “how often”. Expressed as a measurable number or can be quantified.

- *Discrete data*
  Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts.
- *Continuous data*
  Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.

Qualitative data

Qualitative data can’t be expressed as a number and can’t be measured. Qualitative data consist of words, pictures, and symbols, not numbers. Qualitative data is also called categorical data because the information can be sorted by category, not by number.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

NOMINAL VS ORDINAL DATA

Nominal data

Nominal data is used just for labeling variables, without any type of quantitative value. The name ‘nominal’ comes from the Latin word “nomen” which means ‘name’. The nominal data just name a thing without applying it to order. Actually, the nominal data could just be called “labels.”

Ordinal data

Ordinal variables are considered as “in between” qualitative and quantitative variables.

Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data. Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.

Cannot do arithmetic with ordinal numbers because they only show sequence, but can assign numbers to ordinal data to show their relative position.

In other words, the ordinal data is qualitative data for which the values are ordered.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

FISHERIES DATA

Fisheries data often take the form of counts, scores, ratios, and frequencies.

Fisheries researchers often use nominal data to refer to untreated or raw data that have not been standardised. It is generally preferable to use characters as labels to avoid confusion, although sometimes there is logic to using numerals for ease of sorting or sequencing in some software packages.

Data can be collected incidentally during fishing operations or angling activities including competitions, or independently at times and locations chosen by researchers to conform to a specific statistical design.

Fishery dependent

- CPUE - The most commonly applied data in fisheries assessments is catch-per-unit-effort (CPUE) usually via logbooks mandated by regulation in commercial fisheries. CPUE can be described as the quantitative ratio between catch and effort and is a continuous data form. Although catch expressed in terms of weight is continuous, when its abundance is enumerated in numbers it is discrete. Effort can also be discrete or continuous depending on the units e.g., number of shots or hooks for discrete and time for continuous.
- CREEL – structed questionnaires of recreational anglers delivered by trained interviewers at the shoreline, on jetties or at boat ramps. Participation is voluntary.
- Onboard observation & commercial catch sampling by trained scientific observers.
- Angler diarists – selected anglers acting as ‘citizen scientists’ measure and record their catches in diaries and in some instances may be requested to fish in particular locations using specific gear e.g. hooks.
- Mobile phone Apps e.g. GoFish

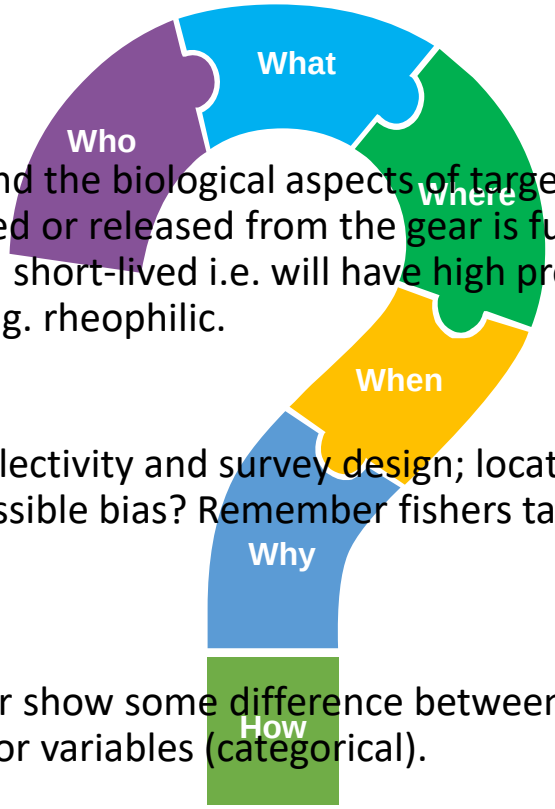
Fishery independent

Survey or monitoring data acquired independently of fishing have the advantage of conforming to a controlled design that can be readily analysed with conventional statistical methods. Their disadvantage can be cost of achieving adequate replication for statistical power to detect changes.
Some use fishing equipment, alternatives include cameras, mounted underwater on baited stands (baited remote underwater video or BRUV), on poles at boat ramps (CCTV), or on unmanned aerial vehicles (UAV).

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

EXPLORATORY DATA ANALYSIS



Fisheries biological context

Knowledge about how a fishery operates and the biological aspects of target and bycatch species such as life history and behaviour of the species caught and retained or released from the gear is fundamental. As yourself is this species fast growing, early maturing, highly fecund, and short-lived i.e. will have high productivity or is it long-lived. How does it feed? Habitat and environmental requirements e.g. rheophilic.

How were the data acquired?

Considerations such as method e.g. gear selectivity and survey design; location, time of day, duration, season. Repeated measures, stratified random, sources of possible bias? Remember fishers target so their decision are not random but there are still stochastic elements at play.

What are the questions being posed?

Are you trying to detect a temporal trend or show some difference between species. Hypothesis. Statistical model: dependent variable, explanatory or predictor variables (categorical).

Properties of the data

Which categories do the data conform with as described earlier? What was being measured and is it a direct measure of, say, abundance or biomass or a proxy e.g. CPUE?

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

EXPLORATORY DATA ANALYSIS

Fisheries biological context

Knowledge about how a fishery operates and the biological aspects of target and bycatch species such as life history and behaviour of the species caught and retained or released from the gear is fundamental. As yourself is this species fast growing, early maturing, highly fecund, and short-lived i.e. will have high productivity or is it long-lived. How does it feed? Habitat and environmental requirements e.g. rheophilic.

How were the data acquired?

Considerations such as method e.g. gear selectivity and survey design; location, time of day, duration, season. Repeated measures, stratified random, sources of possible bias? Remember fishers target so their decision are not random but there are still stochastic elements at play.

What are the questions being posed?

Are you trying to detect a temporal trend or show some difference between species. Hypothesis. Statistical model: dependent variable, explanatory or predictor variables (categorical).

Properties of the data

Which categories do the data conform with as described earlier? What was being measured and is it a direct measure of, say, abundance or biomass or a proxy e.g. CPUE?

THE TRUE SITUATION IS NOT ALWAYS HOW IT FIRST APPEARS

This photo is real and was not edited. The stone is real, the trees are real, the soil is real and the sky is real.

Now, the only thing you have to do is change your point of view. Look at the photo, upside down!

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

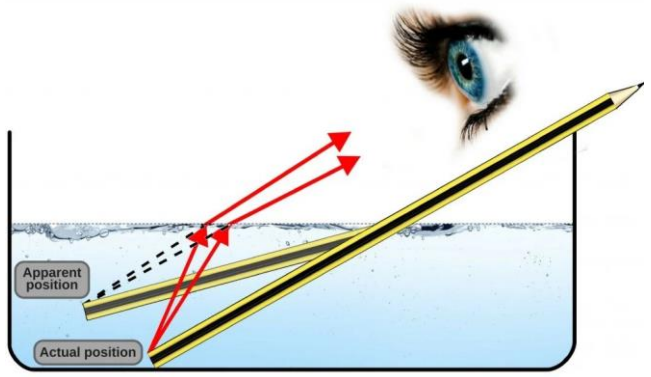
Situations are not always how they my initially appear to be, and it is worthwhile examining from different angles

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

PROXIES

In many areas of biological research, it is not possible to directly measure the variable of interest, so we select a **proxy** measure. One that we think mimics the variable we want to measure.



This is true in other fields such as biochemistry or pathology where spectrophotometric methods or changes in the colour of a chemical indicator are used. The are many instances where something can be detected at much higher resolution or sensitivity than if it was to be observed directly but the critical requirement is that what is being measure changes proportionally with the variable of interest (following some treatment procedure) and not some other extraneous variable. Microscopy involves the use of various treatments of a specimen including the application of stain or fluorescent dye and then subjecting the treated specimen to light of specific wavelengths through a series of lenses. What is observed or measured e.g. luminosity is a refractive image of the subject, not reality. Knowledge of the procedures used and how they work is essential for appropriate scientific interpretation of data from the image.

Proxy measures in fisheries science are also not reality and there is far less control over the “image”. There is a much greater prospect that what is measured is not the effect of variable we are interested in but some other factor. This is where data filtering and cleansing are important and often researchers are not bound by a particular protocol as would occur in a medical pathology laboratory for instance. It is essential that all manipulations of data are explicitly documented to facilitate reproducibility of results.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

SCIENTIFIC INTEGRITY

It is easy to envisage how we can make mistakes with biological research data, the key here is HONESTY and TRANSPARENCY. As professional scientists we must endeavour to work ethically.

Avoiding confirmation bias

All of us have knowledge constructs about how we think the world works and as scientists it behooves us to be aware of our own inherent bias. Whilst fabricating data is the most heinous of scientific frauds, confirmation bias can occur when we have a belief which is stronger than the supportive evidence for our hypothesis. It is too easy to b selective of data that conforms our theories and to ignore that which refutes them.

Firstly, we must acknowledge that we are prone to bias, then try diligently mitigate this by avoiding any *post-hoc* data selection and analysis that is motivated by a desire to seen as successful by “proving” our hypothesis. Removing outliers without justification, filtering inconvenient levels of a factor, or applying the wrong denominator in an F-test are some examples of how we can err.

Science is as much about refuting hypotheses as it is about generating new knowledge. It is not a personal failure if the evidence is inconsistent, instead it means a revised or new hypothesis is needed. Our current constructs will inevitably be updated or replaced as more evidence is acquired.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

“I already wrote the paper. That’s why it’s so hard to get the right data.”

"You are completely free to carry out whatever research you want, so long as you come to these conclusions."

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

“There is a fundamental disconnect between the biological interactions that we observe and common reductionist (linear) assumptions of the framework we use to study them”

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

GENERALISED LINEAR MIXED EFFECTS MODEL (GLMM)

model <- glmmTMB(cpue+0.01 ~ Fyear + (1|pfn) + (1|ShortMonthName) + (1|AreaCode),
data = xxx, family = Gamma(link="log"))

- **Fixed effects** – Fiscal-Year,
- **Random effects** – Fisher, Month, Area-Code
  [inclusion of Fisher × Area-Code produced a poorer fit]

- Run variants of the model with diff combinations of explanatory variables & their interactions e.g. diff fishers fish diff areas & mo.

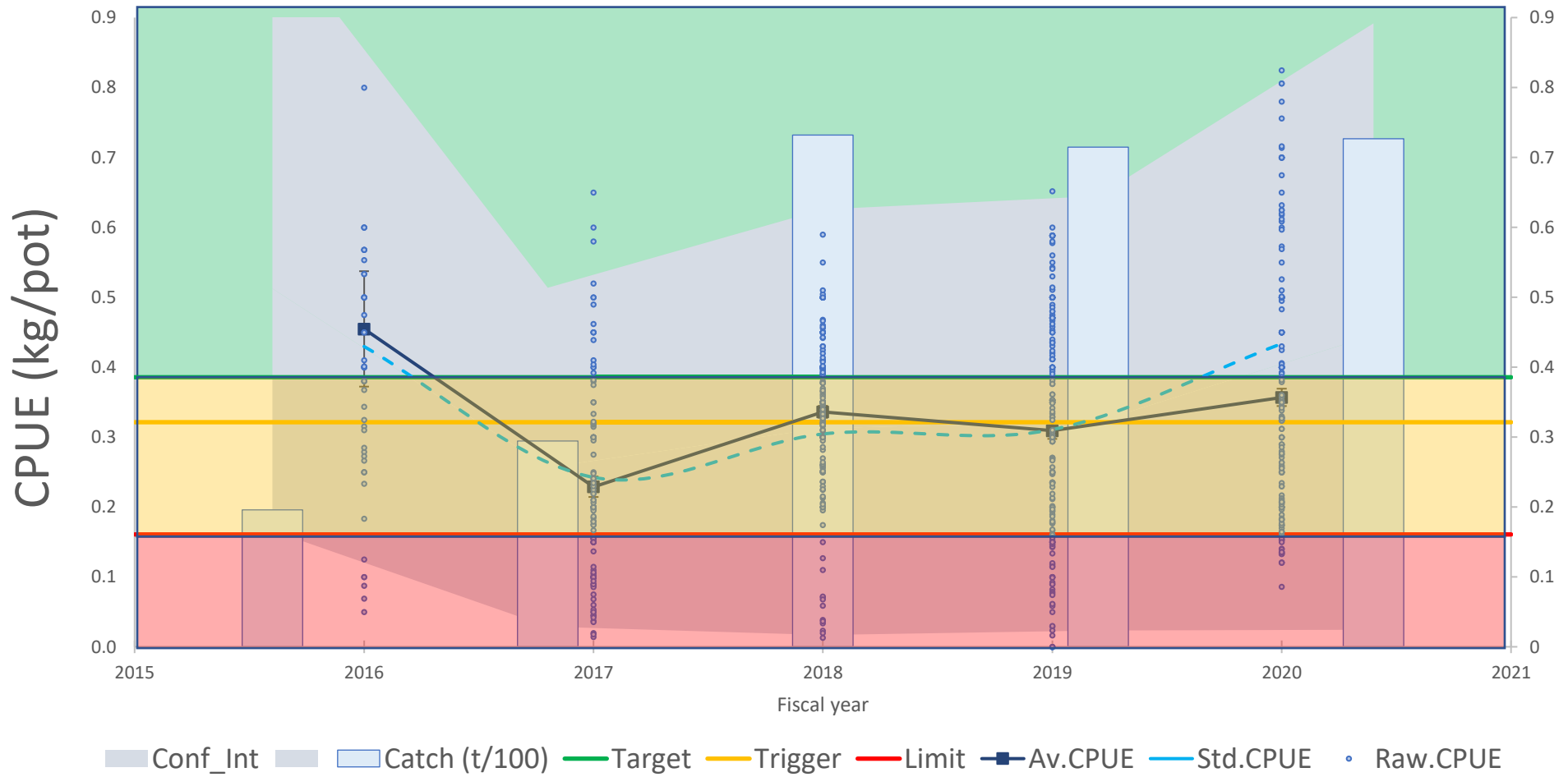
- AIC = Akaike Information Criterion – lowest value is best fit i.e. explains the most variation with fewest variables.

Frequentist approach commonly used for detecting trends in CPUE; alternatives are Bayesian and EDM

Sustainable Inland Fisheries
Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

DATA SERIES RELATIVE TO REFERENCE POINTS

Note diff bt'n average and standardized & broad spread among raw values for each year



Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

CONNECTING ASSESSMENT DATA WITH MANAGEMENT DECISIONS VIA CATCH CONTROL RULES

This can be as simple as:

- If a performance measure (PM) is for an indicator above target reference point value → increase harvest
- PM is between threshold & target reference point values → maintain harvest
- PM is between threshold & limit reference point values → reduce harvest
- PM is below limit reference point value → fishery closure

Definitions:

Performance indicator = biological or stock attribute chosen for assessing management performance e.g. stock biomass
Performance measure = the variable being measured or estimated e.g. CPUE
Reference point = a value of a performance measure chosen to represent a particular stock status i.e. target (TaRP), threshold (TRP), and limit (LRP)
Catch or harvest control rule = action to be taken in response to the current (or future) performance measure relative to a reference point value
Harvest strategy = the decision-making process linking a suite of performance indicators with catch management decisions

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

EXAMPLE OF CATCH CONTROL RULES (3 levels)

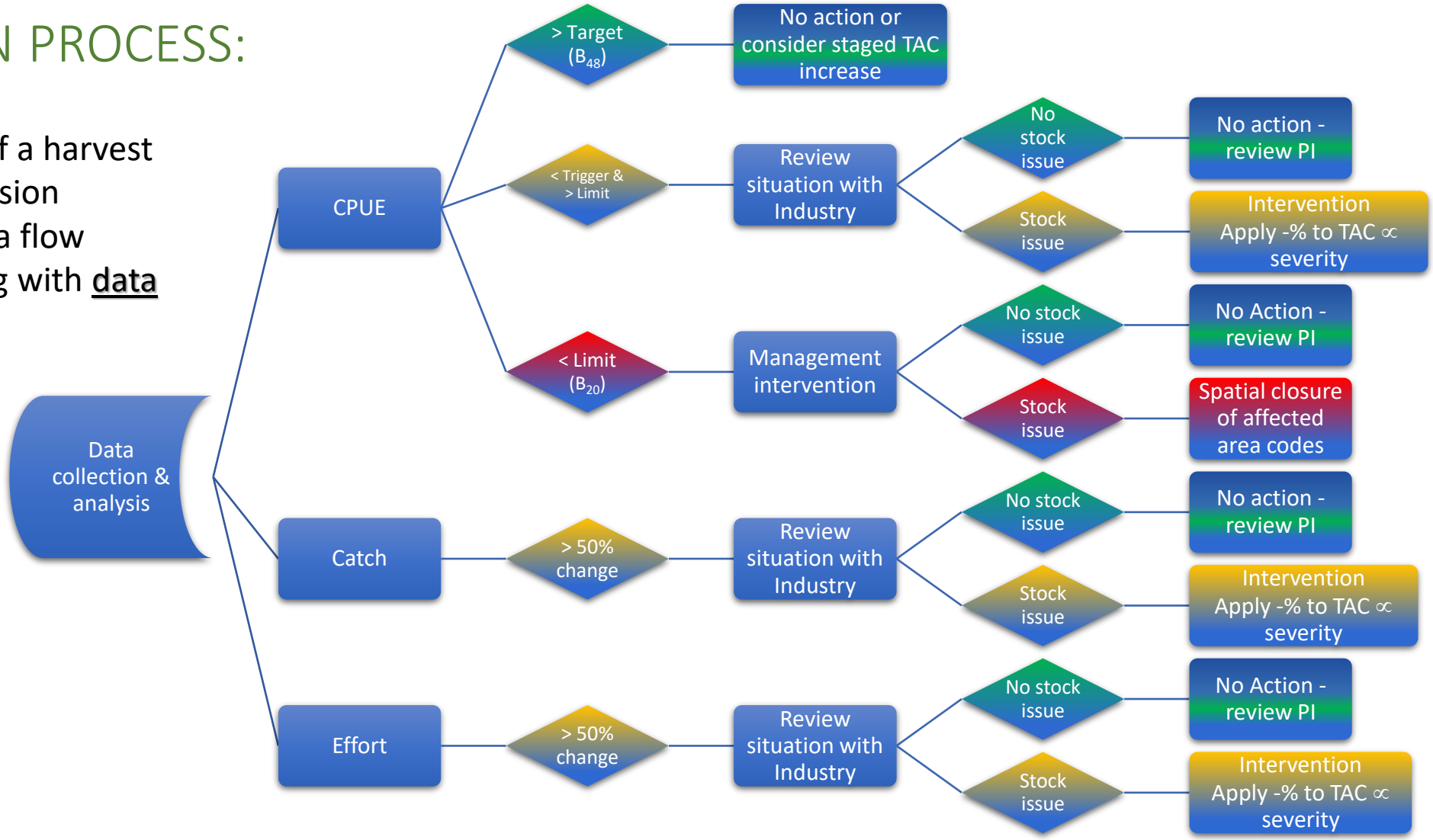
1. **Green Zone** – CPUE ≥ TRP – by at least 10% for 2 consecutive years increase ≤ 15% available; increase > 5% maintained for min 2 y without further increase. Staged increase shall be calculated on the initial TAC and not as a compound increase on the previous increase i.e. 5% + 5% + 5% not 5% × 5% × 5%.
2. **Amber Zone** – LRP ≤ Standardised CPUE < TRP; in this instance a reduction in TAC should be considered either Zone-wide or for one or a group of Area Codes.
3. **Red Zone** – drastic reduction in TAC or mandatory spatial closure should be mandatory; could be on an Area Code basis. Spatial Area Code catch or effort (boat days) caps.
4. **Min Weight**– if more than 5% of the catch sampled and recorded in the research logbook has a mean individual female weight of < 300 g then the rules associated with the Amber Zone shall apply.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

DECISION PROCESS:

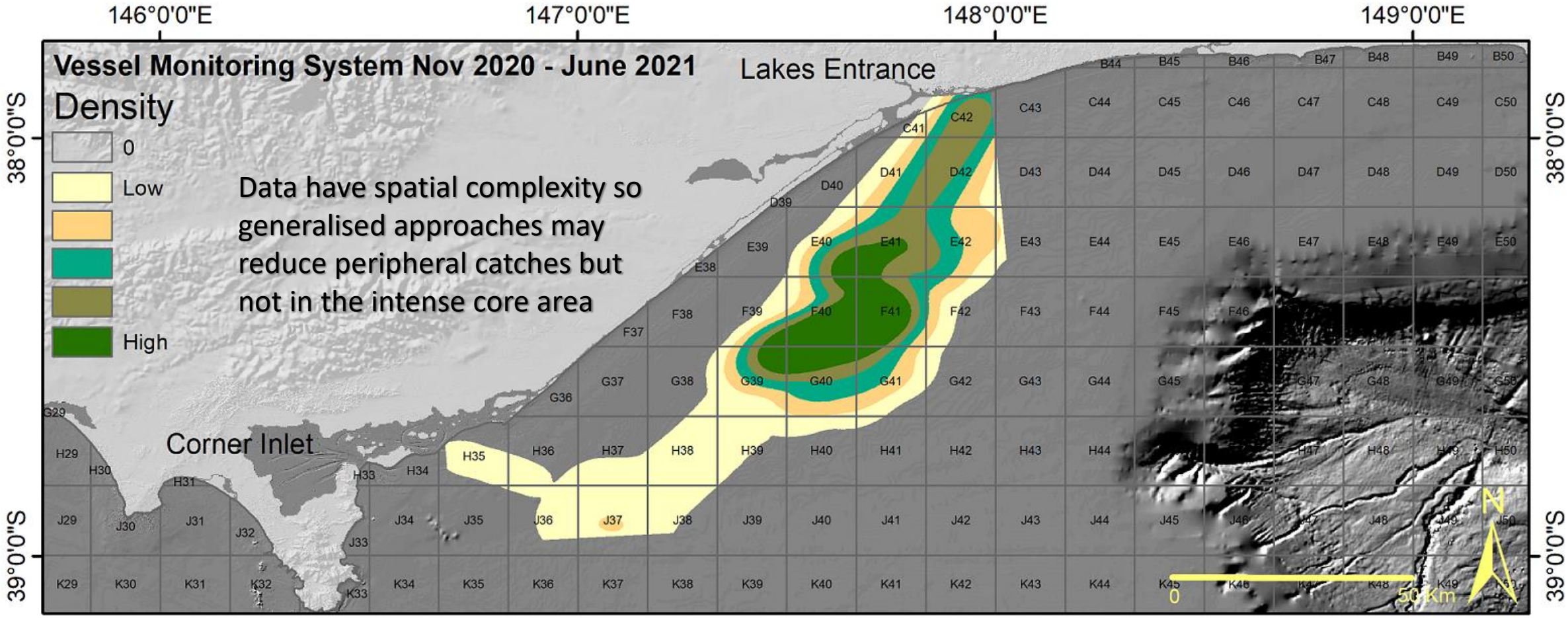
Illustration of a harvest strategy decision structure as a flow chart starting with **data**



Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

SPATIAL INTENSITY OF LOGGED VESSEL



Kernel density estimates from VMS data show south westerly expansion of the fishing grounds

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

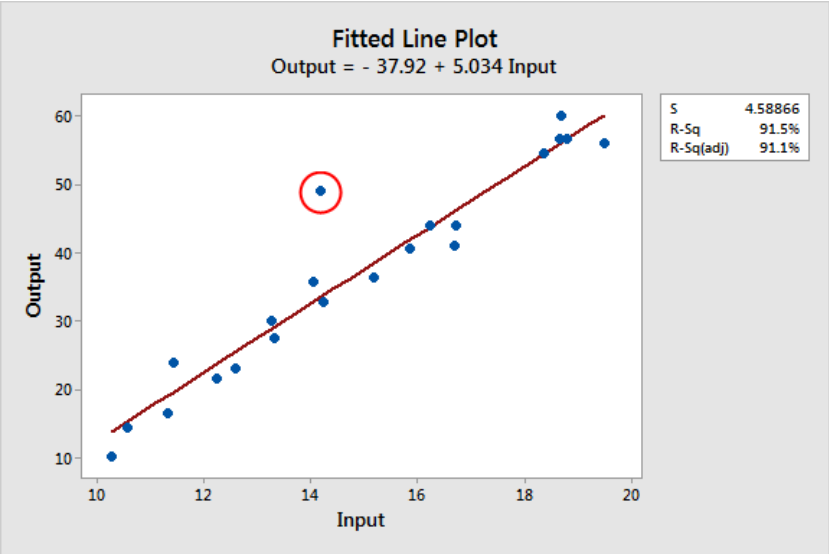
DATA WRANGLING

Outliers or signals?

Outliers are aberrant values that may arise from observation, measurement or sampling error. An important but overlooked and vital step in exploratory data analysis to plot raw data to observe their pattern. Outliers will appear as extreme values which lie away from most data points in the plot [See … https://statisticsbyjim.com/basics/outliers/]

Identifying outliers:

- sorting
- plotting
- Z-scores
- interquartile range bounds
- hypothesis tests

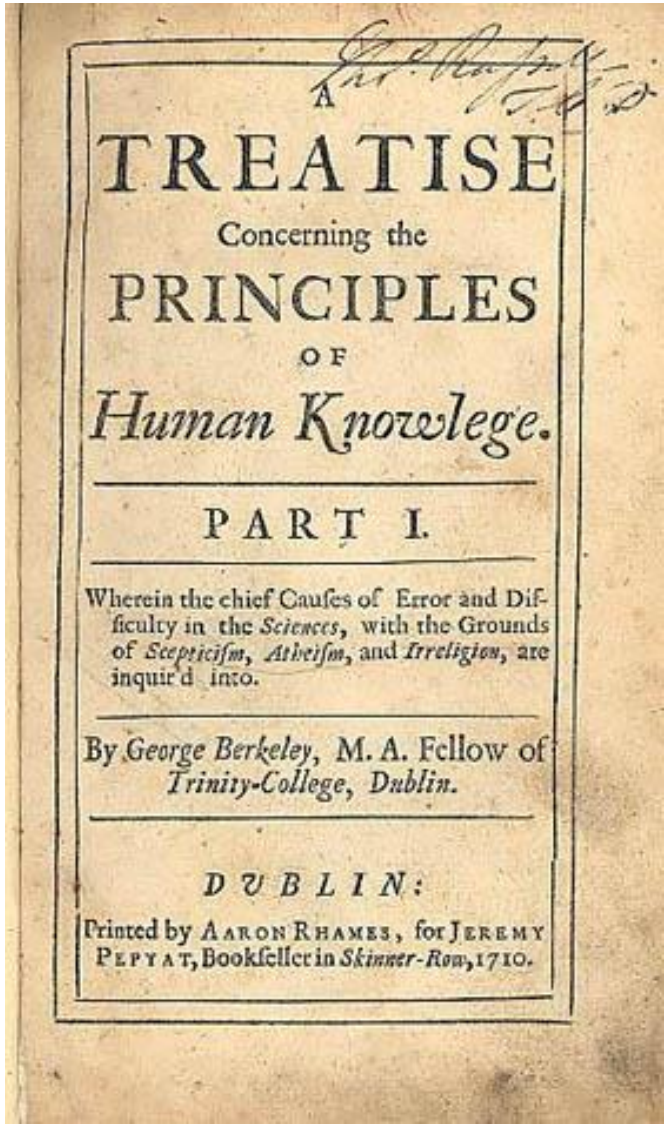


Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

"CORRELATION DOES NOT NECESSARILY IMPLY CAUSATION"

Bishop George Berkley (1710) ***Treatise on the nature of human knowledge***



A
TREATISE
Concerning the
PRINCIPLES
OF
Human Knowlege.

PART I.

Wherein the chief Cauſes of Error and Difficulty in the *Sciences*, with the Grounds of *Scepticiſm*, *Atheiſm*, and *Irreligion*, are inquir'd into.

By *George Berkeley*, M. A. Fellow of *Trinity-College*, *Dublin*.

DUBLIN:

Printed by AARON RHAMES, for JEREMY PEPYAT, Bookſeller in *Skinner-Row*, 1710.

Yet, as scientists we are pursuing causation, hence controlled experimentation, and prediction using mathematical modelling.

Many dynamical systems involve stochastic processes and non-linear feedback loops so the concept of equilibrium, assumed extensively in fisheries population modelling, is invalid.

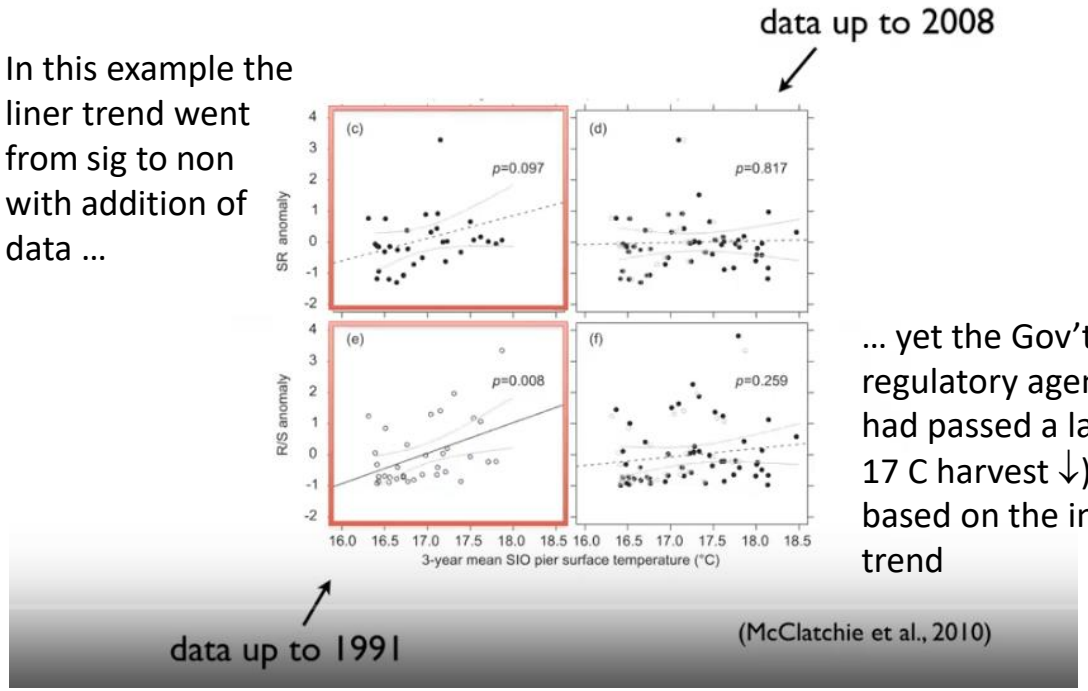
Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

ISSUES WITH CORRELATION

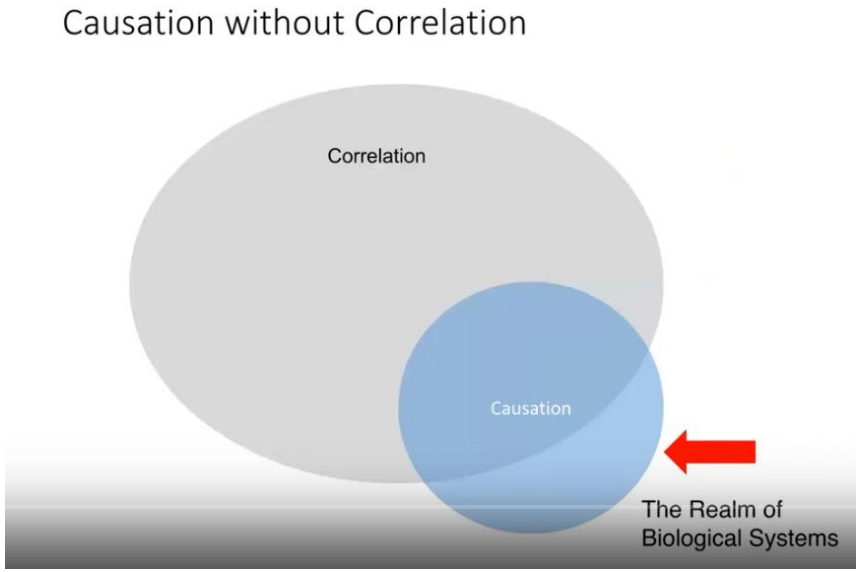
Pacific sardine productivity and SST

In this example the liner trend went from sig to non with addition of data …



… yet the Gov't regulatory agency had passed a law (< 17 C harvest ↓) based on the initial trend

Can also have a causative relationship despite absence of a correlation:



As with long-run odds e.g., where multiple tosses of a fair coin the probability will invariably lead to a 50:50 outcome, adding more years to a time series will inevitably reveal that apparent trends are often transient and without any predictive value.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

Static Theoretical Ideal vs. Dynamic Reality

- **Static Theoretical Ideal** (classical linear framework)
  - equilibrium
  - stable
  - separable (decomposable, study piecewise)
  - Granger
  - classic parametric models
- **Dynamic Reality** (nonlinear empirical dynamics)
  - non-equilibrium
  - non-stable
  - non-separable (interdependent, study as a whole)
  - CCM
  - empirical dynamic models

CCM = cross correlation matrix

"**Granger** representation theorem states that if a set of non-stationary variables are cointegrated then they can be characterized as generated by an error correction mechanism…. Cointegration places too much importance on the long run and excludes interesting short run dynamics."

Sustainable Inland Fisheries

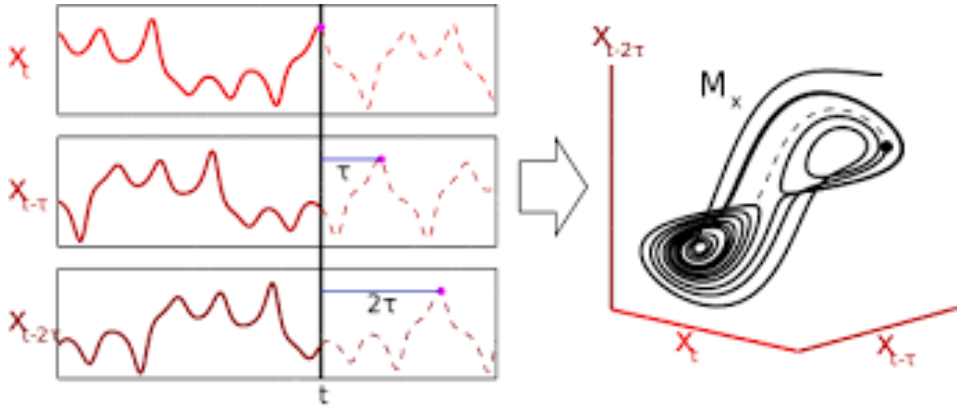
Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

DATA WRANGLING

Could it be a signal?

Most time series are short, typically less than 2 decades and in our attempts to detect cycles and trends, these are too brief to detect important effects which occur at intervals greater than a decade. Professor George Sugihara and his collaborators and students at Scripps Institution of Oceanography have developed what he calls empirical dynamics models (EDM) using mathematical techniques such as S-mapping with notable success.

Lorenzian “butterfly” attractor (Takens Theorem)



If you are interested then I recommend viewing his lectures posted on YouTube where there are some of the graphical illustrations, so-called “butterfly attractors”, but for this module the key point is that observations that may appear to be outliers could possibly be the most important ones. Ignoring or omitting them risks deceiving ourselves and discarding valuable data. This can become apparent when more years’ data are added to a series.
https://cran.r-project.org/web/packages/rEDM/vignettes/rEDM-tutorial.pdf
https://www.youtube.com/watch?v=uhONGgfx8Do

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

THE PROBLEM WITH CORRELATION

Myers, 1998

A meta-analysis of 74 environment-recruitment (fish productivity) correlations reported in the literature.

- Only 28 out of 74 held to retest when data subsequent to the original study was added.

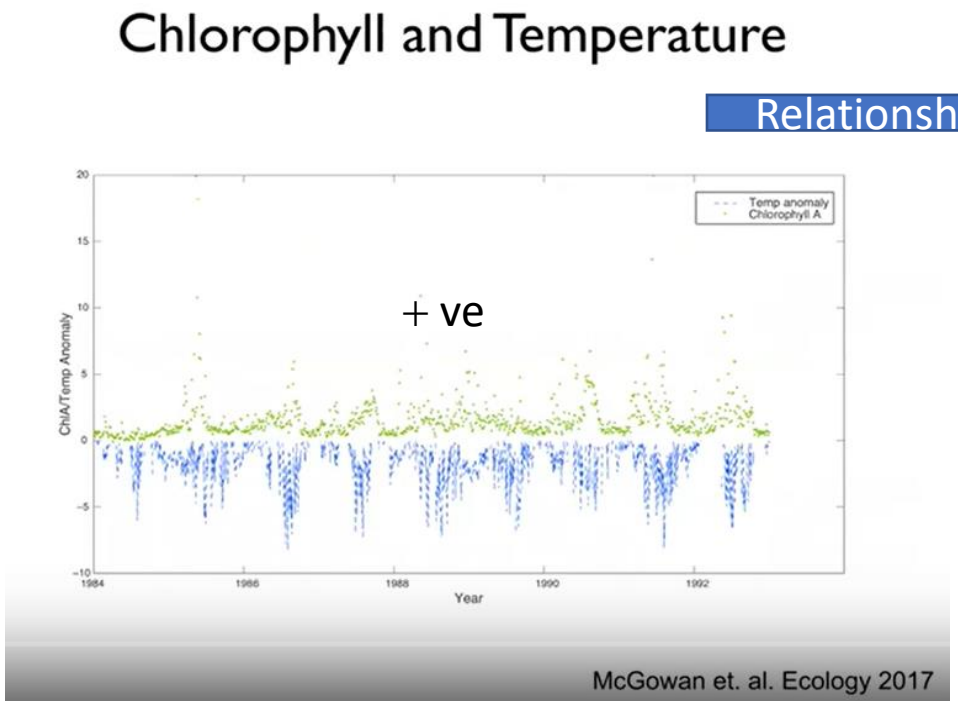
(Fewer now: sardine-temperature was still successful at that time)

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

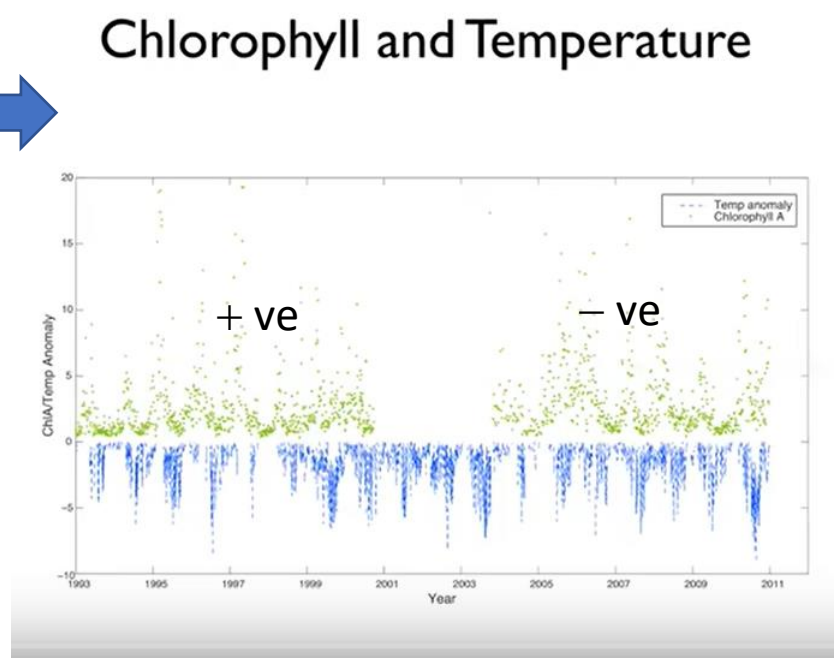
DATA EXAMPLES: MIRAGE CORRELATION ARISING FROM NON-LINEAR DYNAMICS

Red tide algal blooms v. SST anomaly that causes stratification in water column:



Positive correlation until 2001 when funding ceased …

Relationship flips



… after monitoring resumed in 2004 the correlation had flipped to become negative!

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

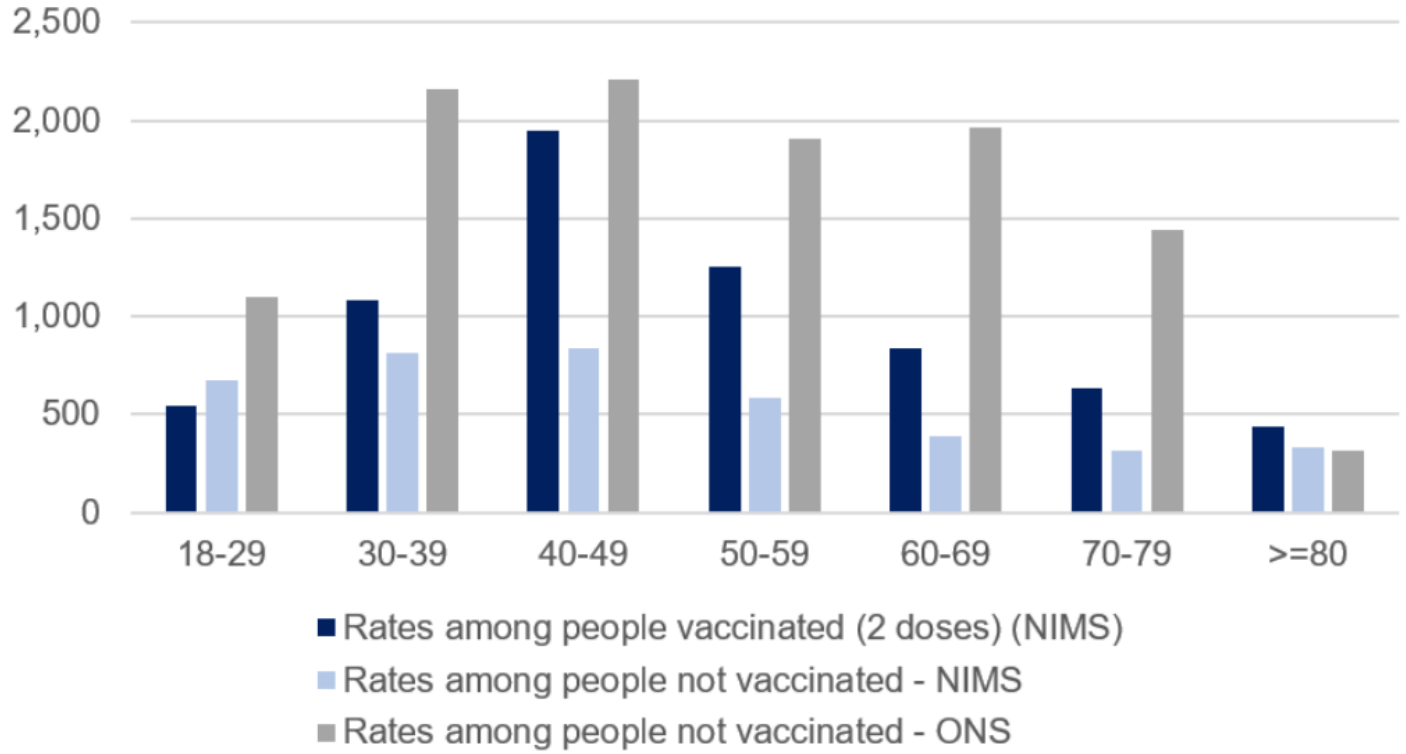
EXAMPLES

As per effort in CPUE consistency and accuracy in the denominator is important.

In empirically comparing rates of COVID-19 between vaccinated and unvaccinated people in the UK, contradictory inferences can be drawn depending upon the source of the population estimates.

NIMS data shows that those who were vaccinated have higher rates of infection than those vaccinated. ONS data show the unvaccinated are more at risk. Although the difference in the 40 – 49 yo age group is much less than for other age categories and the unvaccinated 80+ yo appear to be slightly better off.

INTERPRETATION CHANGES WITH DENOMINATOR DATA SOURCE



Different population estimates can create different results. *(UK OSR, via ONS, NIMS and UKHSA)*

National Immunisation Management Service (NIMS)
Office for National Statistics (ONS)

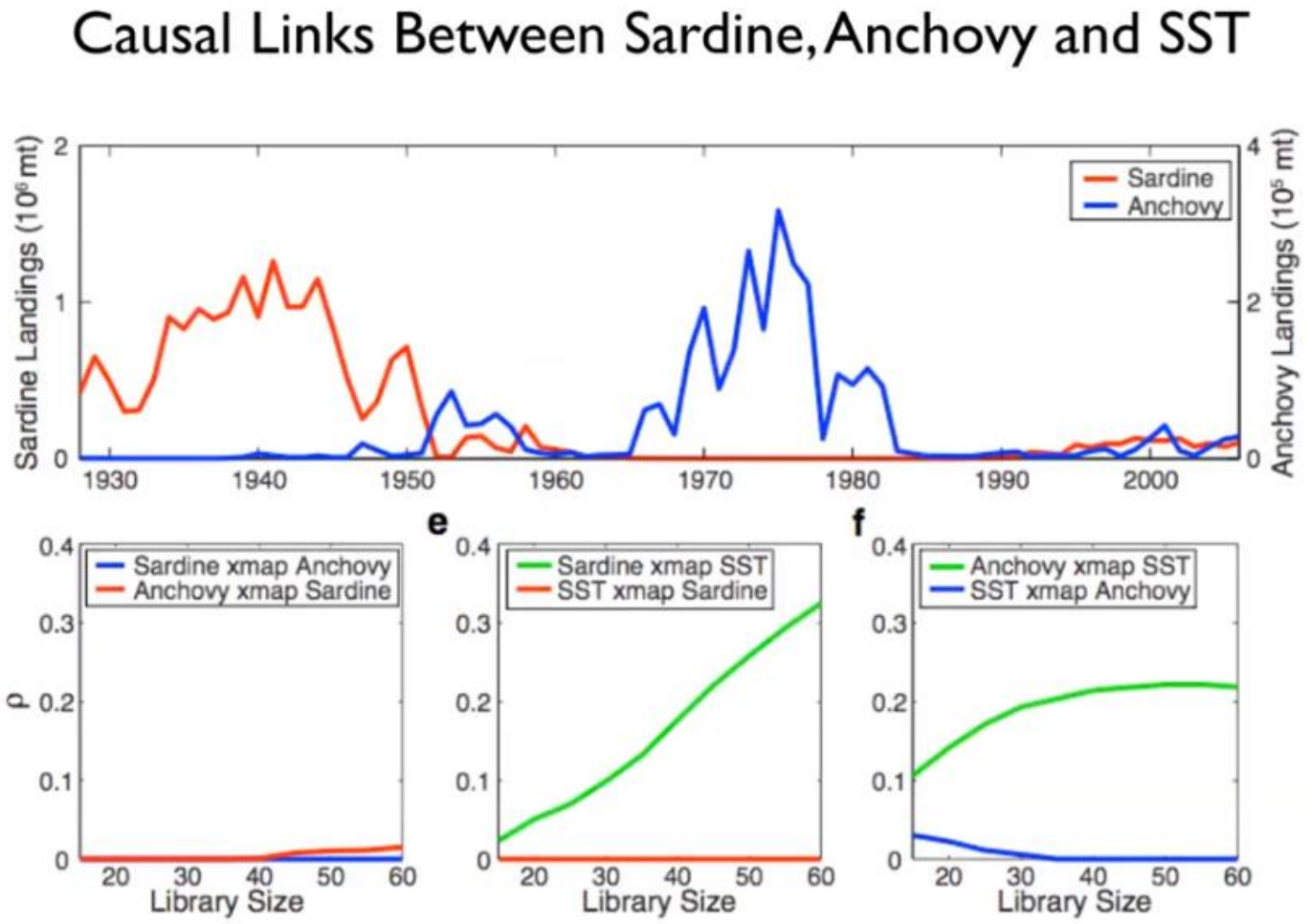
Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

EXAMPLES

For some time it was thought that sardine and anchovies exhibited alternating patterns of abundance (based on catch weight) caused by competition for food.

Application of cross-mapping using empirical dynamics modelling (EDM) showed an absence of a relationship between these fish species, but the temporal patterns in landings related to differing responses to SST.



Sustainable Inland Fisheries

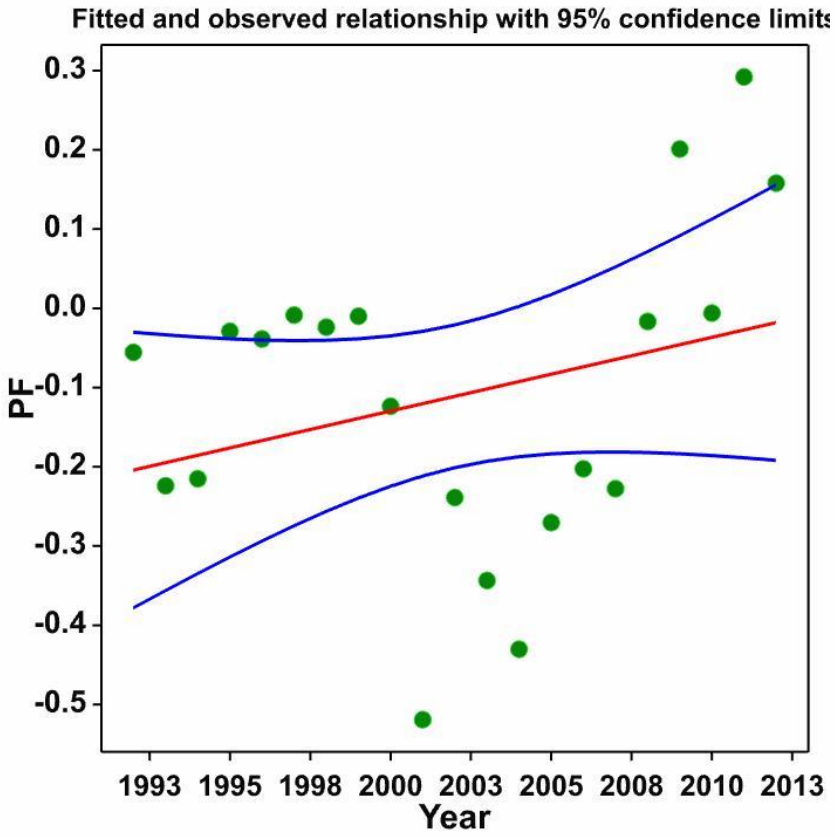
Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

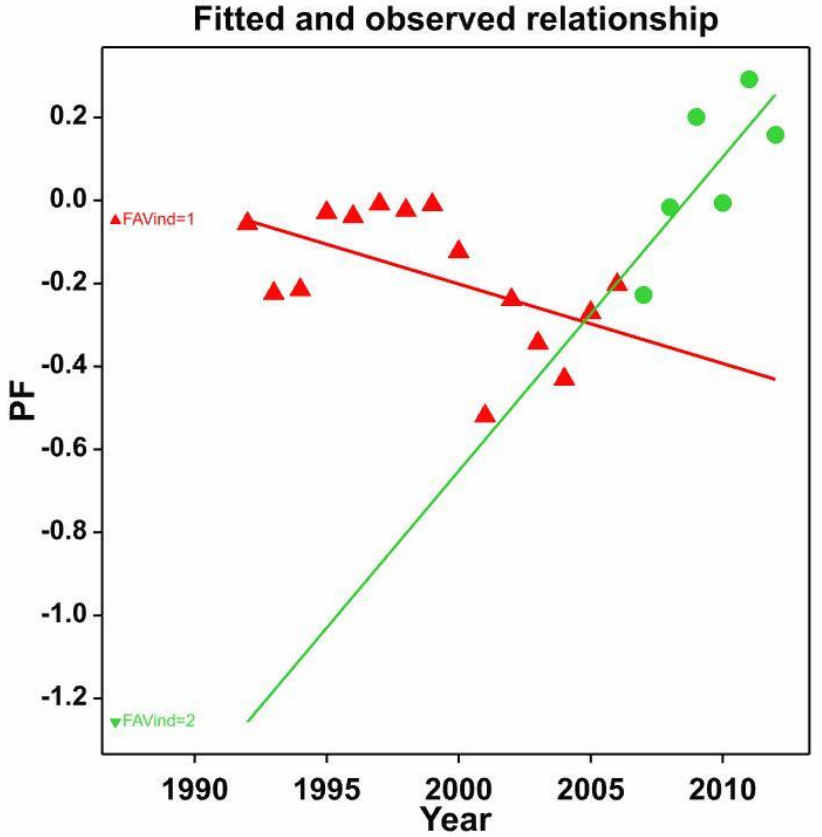
EXAMPLES

Fishery independent survey counts of gastropod abundance during a disease outbreak in 2006.

Some of the points appear to be outliers, but are they really?

How the data are grouped is critical and in this instance a binary variable was added to the model:
0 = healthy,
1 = diseased





The impact of disease is clear when the binary variable is added to the GLMM; this required specific knowledge about the progress of the disease outbreak.

Sustainable Inland Fisheries

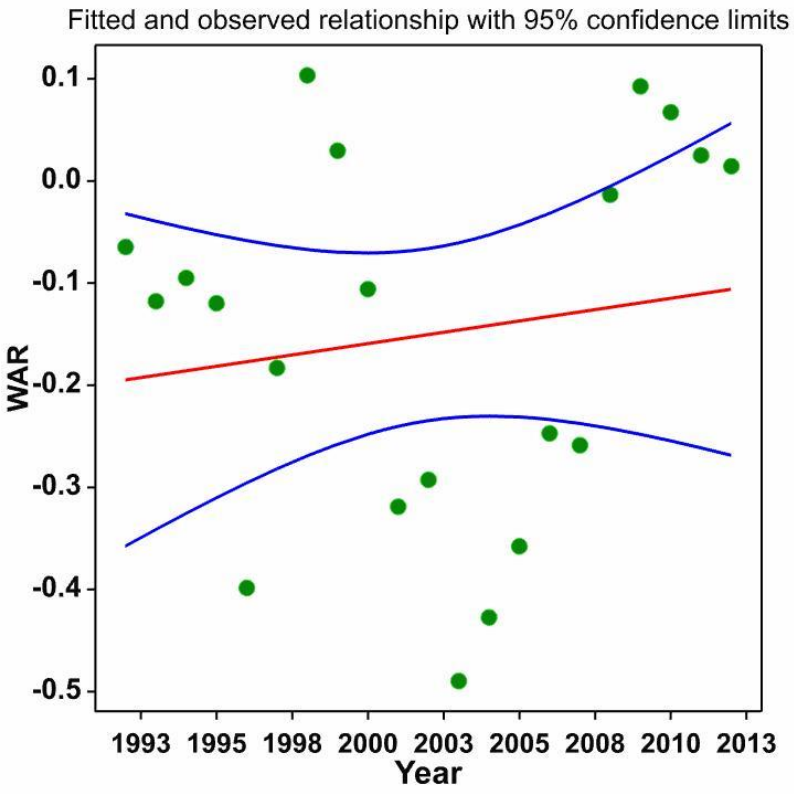
Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

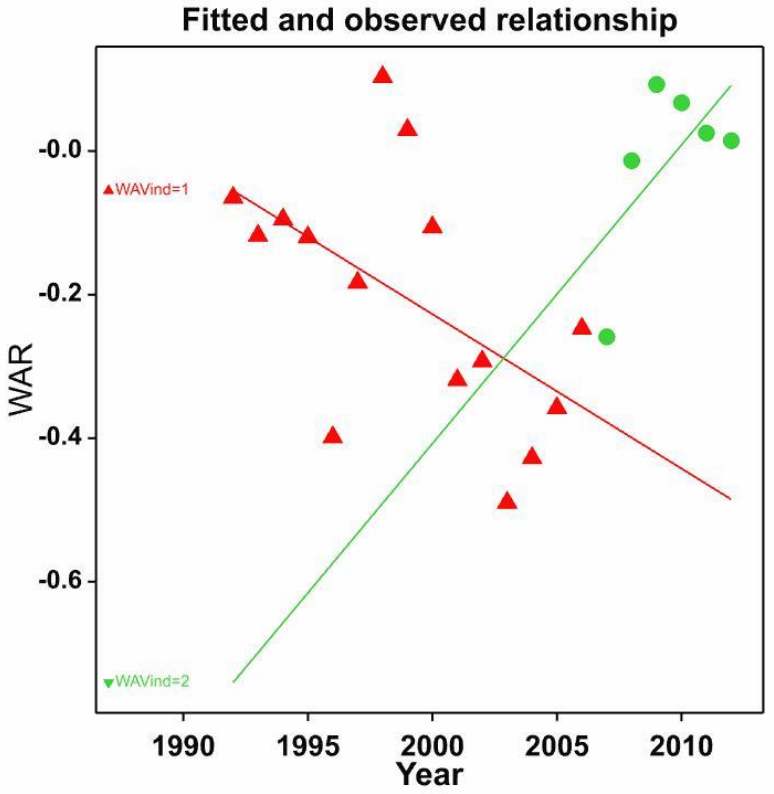
EXAMPLES

Fishery independent survey counts of abundance during a disease outbreak in 2006.

Some of the points appear to be outliers, but are they really?

How the data are grouped is critical and in this instance a binary variable was added:
0 = healthy,
1 = diseased





As at the previous location the impact of disease is clear when the binary variable is added; this required specific knowledge about the progress of the disease outbreak.

Sustainable Inland Fisheries

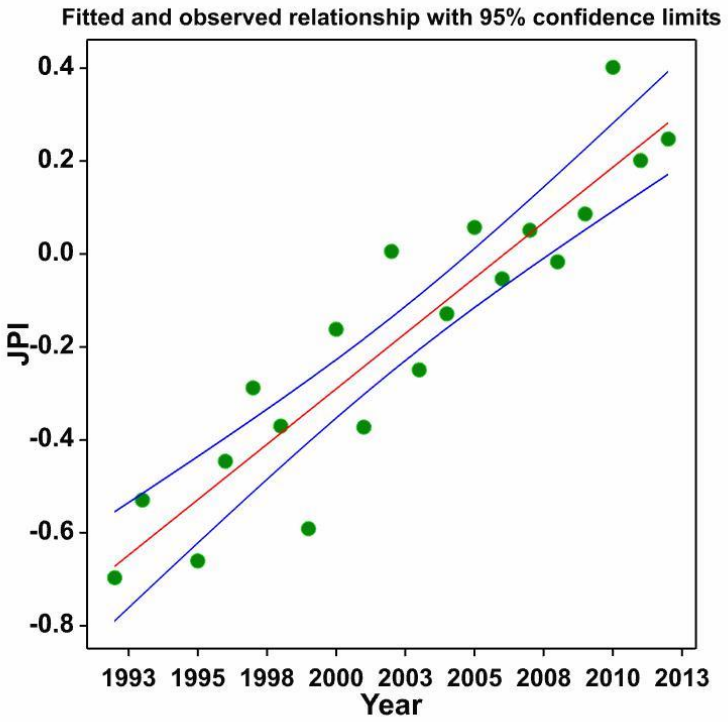
Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

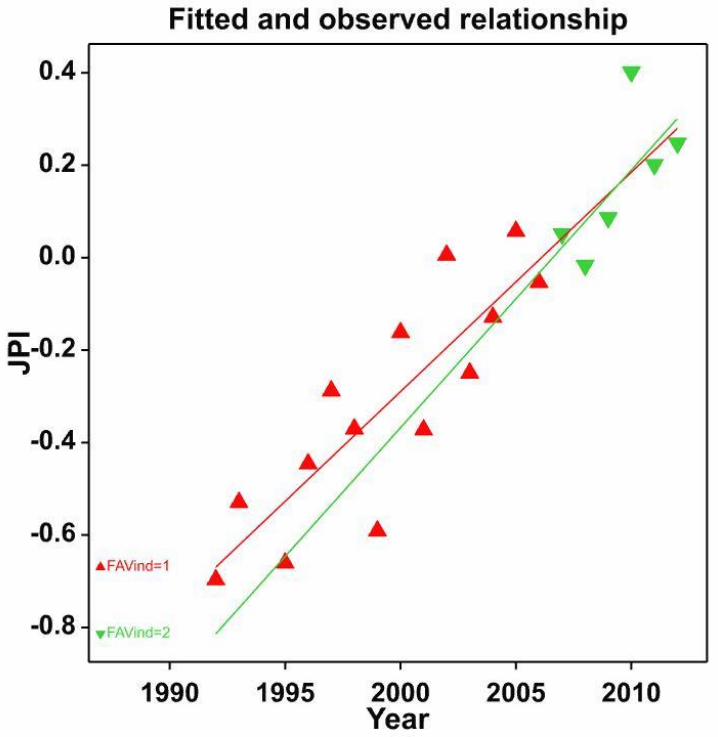
EXAMPLES

Fishery independent survey counts of gastropod abundance during a disease outbreak.

The disease affected different locations at different times. An expert scientific witness in a subsequent litigation incorrectly pooled the data = false inference.

How the data are grouped is critical and in this instance a binary variable was added:
0 = healthy,
1 = diseased





This location, an island 7 km offshore, remained free of disease but the binary variable was added because a substantial amont of fishing effort transferred to this location during the outbreak as infcted areas were closed.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

CLEANSING DATA

Filtering & cleansing

CPUE relies upon the unit of **effort** being identical for all of the catch data. If there is some variability in effort the data can be standardised or perhaps scaled for things such as engine power, tow speed or shot duration. Scaling requires that we know if the relationship is linear or follows a power curve.

Sometimes, we might have information that would suggest the fishers were targeting species other than the one of interest in and hence CPUE will be lower than expected, or some fishers may be inexperienced and less skilled, or their gear may be inconsistent with the rest of the fleet. In these instances, it may be better to omit the data for these fishers or fishing events. This will generally not be a problem when there are many fishers and many events in a dataset.

Transforming

Frequentist statistical analyses often assume the data are from a normal distribution and it is not uncommon to transform the data to reduce heteroscedasticity. In generalised linear models of CPUE, a log-link function is generally specified, whereas with abundance count data a negative binomial or Poisson is chosen due to the dat being zero-inflated i.e. many zero counts among species with spatially heterogeneous distributions within survey areas. It is important to note that transformation reduces or dampens the information (contrast) in the data and back transformation does not recover reality, and should be viewed as a re-scaling process to make it compatible with the raw data scale.

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

CLEANSING DATA

Sugihara recommendations:

- **Do not filter data - filtering can remove some interesting data that has a mid-range signal that is important to the dynamics; although it is common in fisheries and is necessary in standardising gear i.e. units of measurement and there are occasions when it makes sense, more often than not should be avoided.**
- **Need to justify removing outliers - sometimes the largest values are most important**
- **Avoid log transformations as they can obliterate data**

DO NOT OVER CLEAN

Sustainable Inland Fisheries

Nature Research Centre
Akademijos st. 2, LT-08412, Vilnius, Lithuania
www.gamtostyrimai.lt

SUMMARY

Key points -

- EDA is an essential first step prior to analysis and do not forget to plot the raw data
- Consider the biological and fisheries context and assessment objectives prior to making decisions when wrangling data
- Remove outliers only when this can be justified
- Be cognisant of the effects that filtering cleansing, scaling, and transforming data may have on the performance of the intended analyses and their outputs

You should now be confident that you can -

- describe some of the common issues and types of data observed in datasets,
- undertake an exploratory data analysis (EDA),
- identify outliers and decide whether to retain, eliminate or minimise their effects,
- filter and ‘cleanse’ data prior to analysis,
- select an appropriate model statement for a statistical analysis to answer the central question, and
- describe several of the implications of omitting EDA & filtering,
- describe the limitations and pitfalls of this approach.

This will ensure that you can efficiently and effectively prepare datasets for analysis whilst avoiding unnecessarily discarding what could turn out to be important information in the delivery of scientifically defensible advice to fisheries and environmental managers.