Model Diagnostics in Integrated Stock Assessments

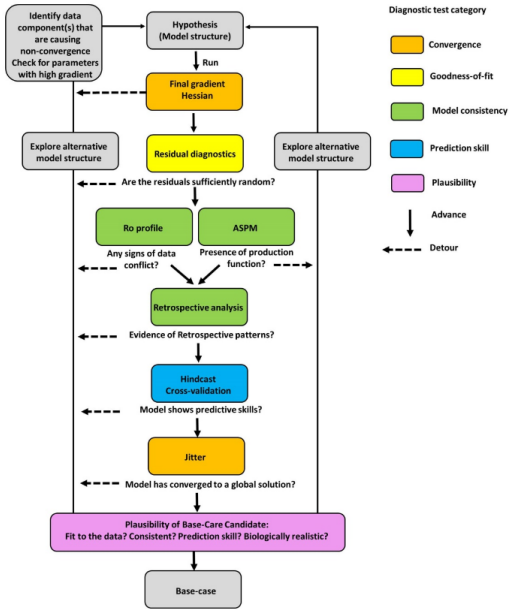
Hindcasting

Laurence Kell, Henning Winker, Massimiliano Cardinale, Rishi Sharma, Iago Mosqueira, Toshihide Kitakado

Jan 31-Feb 3 2022

What diagnostics should be defaults in assessment reports?

Cookbook Carvalho et al. (2021)



ICES

- Working Groups assess multiple stocks using a single model per stock

ICCAT & IOTC

- Working groups assess a single stock using multiple modelling frameworks and scenarios

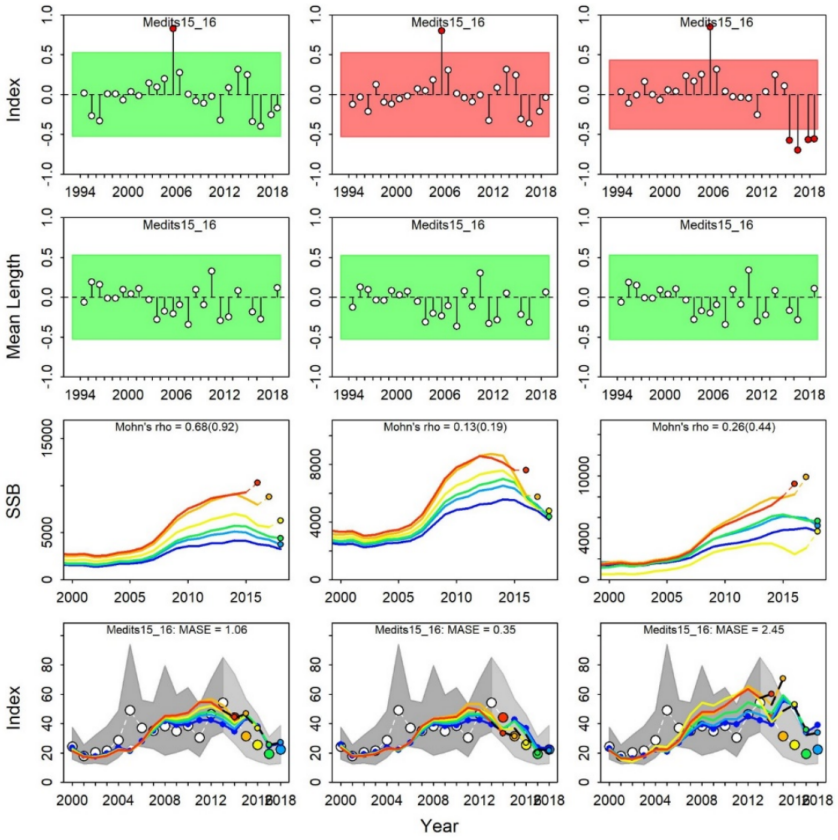
Therefore focus is on

- Residuals
- Retrospective Analysis
- Prediction skill

GCFM Example

Sicilian Hake

Alternative scenarios



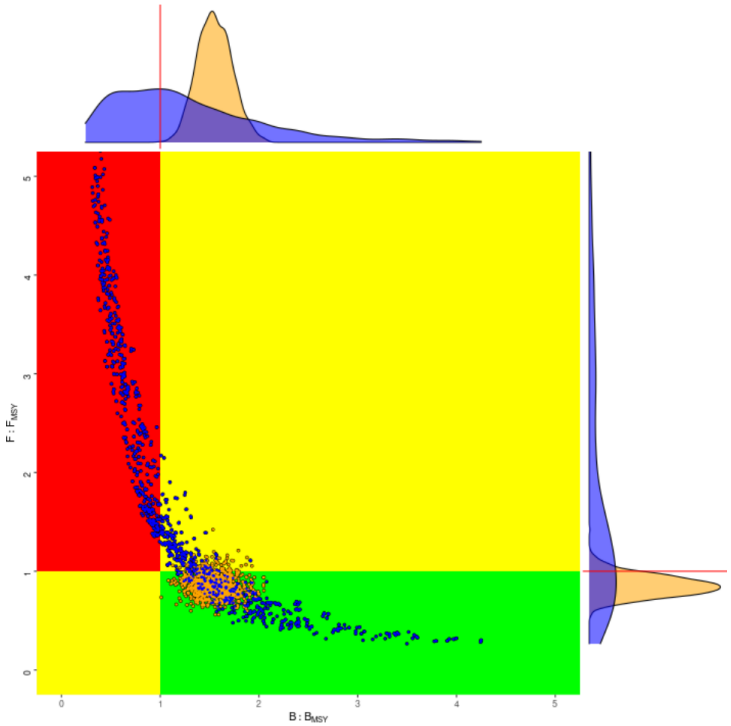
Diagnostics

- Residual Runs Test for
  - *Indices*
  - *length compostions*
- Retrospective Analysis and Mohn's ρ
- Prediction Skill and MASE
- 2^{nd} model passes Retrospective Analysis & Prediction Skill test, but fails Runs Test for index.

IOTC

Albacore

Model Error v Estimation Error

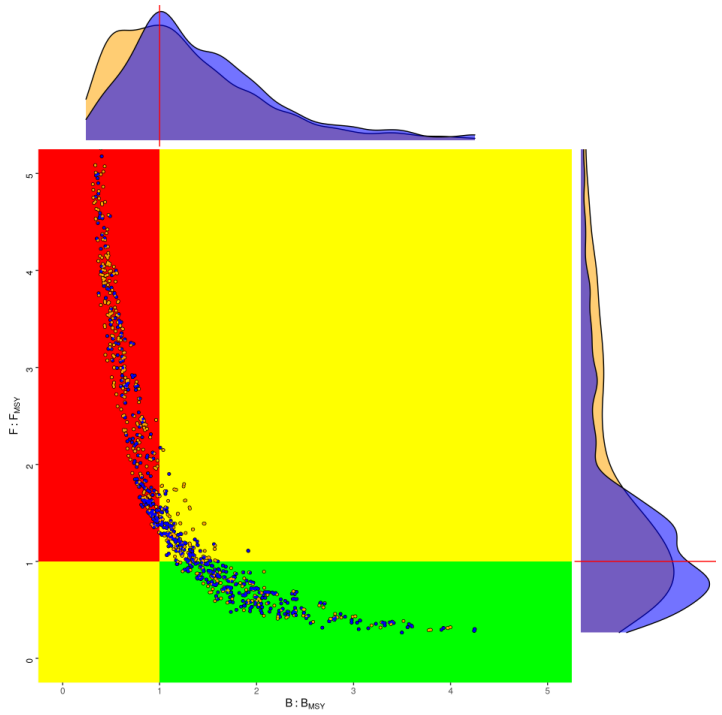


1440 scenarios

- M
- Steepness
- Sigma R
- Effective Sample Size
- Index
  - *CV*
  - *Catchability*
- Selectivity

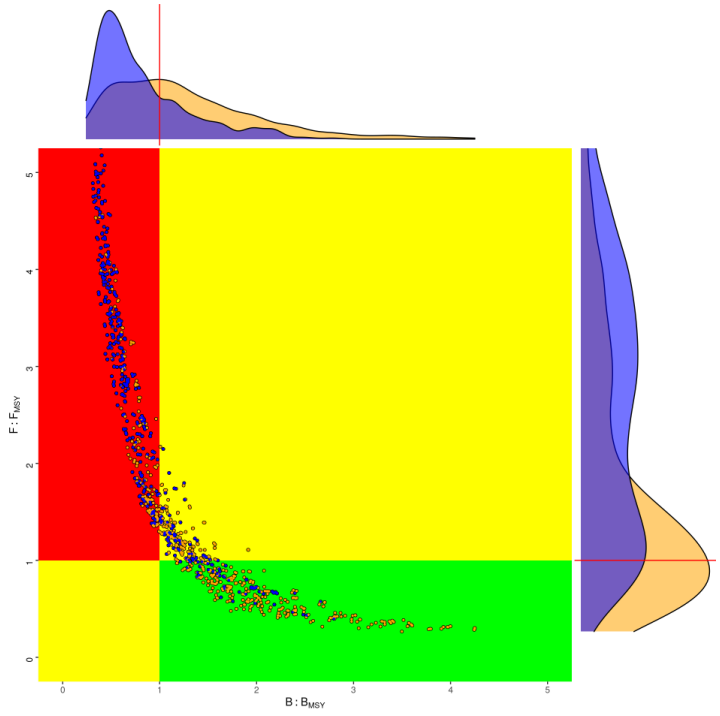
Should diagnostics be used to eliminate models, weight models, or identify and fix model misspecification?

AIC weighting



Should diagnostics be used to eliminate models, weight models, or identify and fix model misspecification?

Mohn's ρ Retrospective Weighting



Comparing Multiple Models

World Conference on Stock Assessment Methods used self- and cross-tests to compare models (Deroba et al. 2015)

However

- *"… modellers are aware that there is a trade-off between the usefulness of a model and the breadth it tries to capture. But many are seduced by the idea of adding complexity in an attempt to capture reality more accurately. As modellers incorporate more phenomena, a model might fit better to the training data, but at a cost. Its predictions typically become less accurate. As more parameters are added, the uncertainty builds up (the uncertainty cascade effect), and the error could increase to the point at which predictions become useless."* (Saltelli et al. 2020)

- *"The primary diagnostics used to compare models are to examine residuals patterns to check goodness-of-fit and to conduct retrospective analysis to check the stability of estimates. However, residual patterns can be removed by adding more parameters than justified by the data, and retrospective patterns removed by ignoring the data. Therefore, neither alone can be used for validation, which requires assessing whether it is plausible that a system identical to the model generated the data."* (Kell et al. 2021)

Cross-validation

Used to determine how well a predictive model will perform in practice.

- A model is fitted using a training dataset, and then predictions made for a test set, not used for fitting.
- An aim is to test the model's ability to predict new data not for fitting, to identify
  - *overfitting*
  - *bias; and*
  - *provide insight on how well the model will perform in practice.*
- Can be performed in a variety of ways e.g.
  - *Peel back years as in a retrospective analysis*
  - *Remove whole series or fleets*
  - *Randomly leave out a year of composition data, e.g. is modelling time varying selectivity important.*
  - *Hindcasting by peeling back individual datasets, or series.*

Hindcast

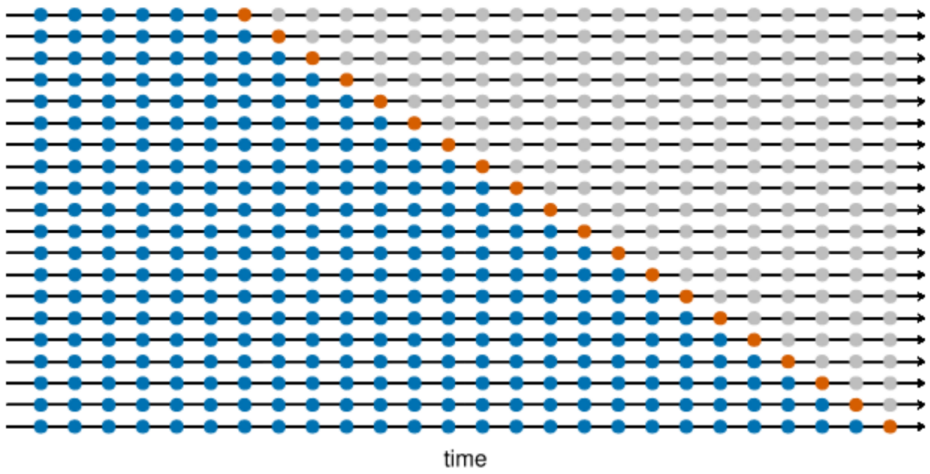
Validation requires that the system is observable and measurable, and so observations should be used, rather than model-based quantities unless these are well known (Hodges, Dewar, and Center 1992).

- First step is similar to a retrospective analysis where the final year(s) of data are deleted and the model refitted. The fitted model is then projected forward over the omitted years and of missing observations made.
- Prediction skill, a measure of the accuracy of a predicted value unknown by the model relative to its observed value, can then be calculated using the mean absolute scaled error (MASE). This compares a forecast to a naive prediction, i.e. is a forecast better than saying the weather tomorrow is the same as today. A score of 0.5 indicates that the model is twice as good as random walk.
- Hindcasting can be used to validate models that use different data sets and penalty terms allowing different modelling frameworks and scenarios to be compared.
- Model validation serves a purpose complementary to model selection and hypothesis testing. Model selection searches for the most suitable model within a specified family, and hypothesis testing examines if the model structure can be reduced, while validation examines if the model family should be modified or extended Thygesen et al. (2017)

Time series cross-validation.

Training set consists of *observations* that occurred prior to the observation that forms the *testset*. No future observations are used in constructing a forecast.

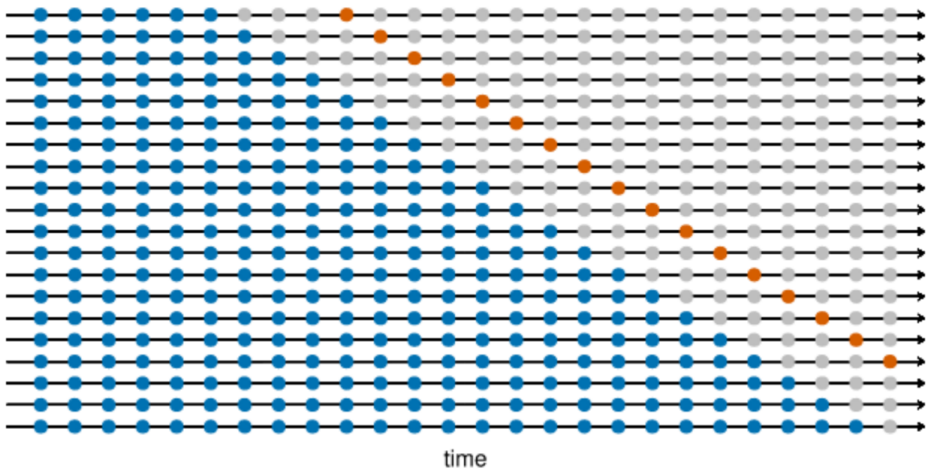
Prediction skill is computed by averaging over the test sets. This procedure is also known as evaluation on a rolling forecasting origin because the origin at which the forecast is based rolls forward in time.



Multi-step forecasts

Multi-step forecasts may be preferreable if assessment advice is for multiple years or benchmarks are conducted every four years.

In this case, the cross-validation procedure based on a rolling forecasting origin can be modified to allow multi-step errors to be used. For 4-step-ahead forecasts the corresponding diagram is



Hindcast

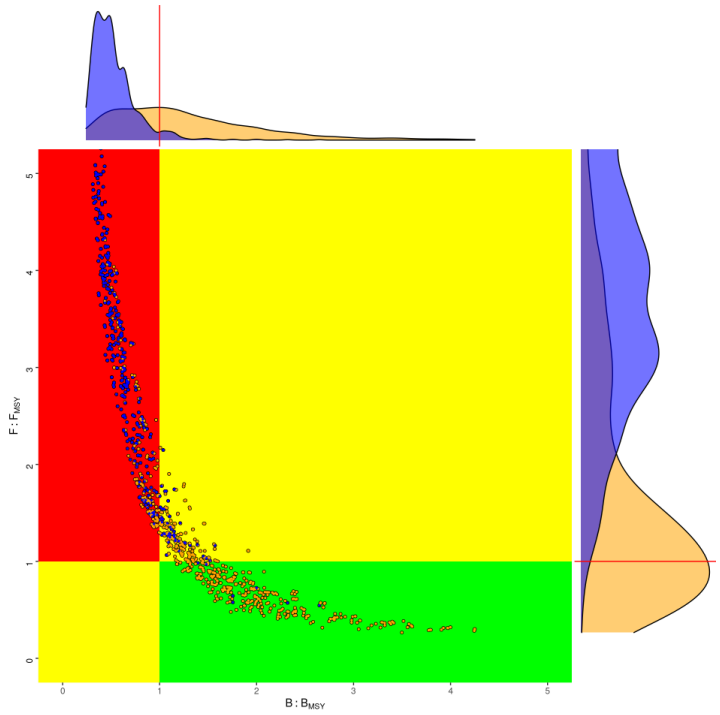
- There two main ways to conduct a hindcast namely
  - *Crossvalidation, using observations; or*
  - *Backtest, using model estimates*
- And three reasons for bothering to do so, namely to
  - *Find a “best assessment”,*
  - *Select and weight models in an ensemble, or*
  - *Condition Operating and Observation Error Models when conducting Management Strategy Evaluation.*
- Multiple ways
  - *As in a retrospective analysis, where all data are removed as years are peeled by*
  - *By data series, e.g. CPUE one-by-one*
  - *By fleet, e.g. CPUE and length compostions*
- And epends on the question
  - *Age-based assessment can predict more steps ahead than biomass-based*
  - *Identify data-conflicts*
  - *Identified where models should be extended or simplified.*

Mean Absolute Scaled Error

- MASE has the desirable properties of scale invariance, so it can compare forecasts across data sets with different scales and has predictable behaviour, symmetry, interpretability and asymptotic normality.
- Unlike relative error, MASE does not skew its distribution even when the observed values are close to zero.
- Easy to interpret as a score of 0.5 indicates that the model forecasts are twice as accurate as a naive baseline prediction.
- The Diebold-Mariano test for one-step forecasts can also be used to test the statistical significance of the difference between two sets of forecasts, i.e. by comparing the prediction $y_t - \hat{y}_t$ to a random walk $y_t - y_{t-1}$.

Simple Skill Weighting

Using MASE



Weighting Metrics

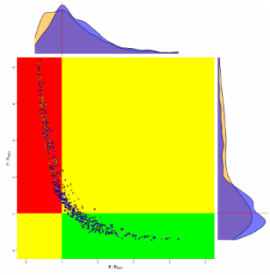
Why can't I just use the AIC?

Because this leads to model overfitting and over-fitted models:

(1)Cannot predict future (advice)
(2)Are highly sensitive to new data points
(3)Likely won't converge when updated next year

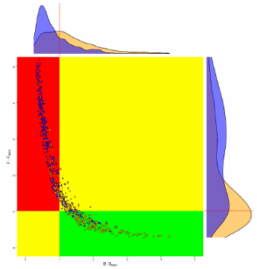
AIC

Fit to the data



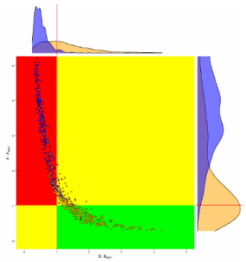
Mohn's ρ

Retrospective Consistency



MASE

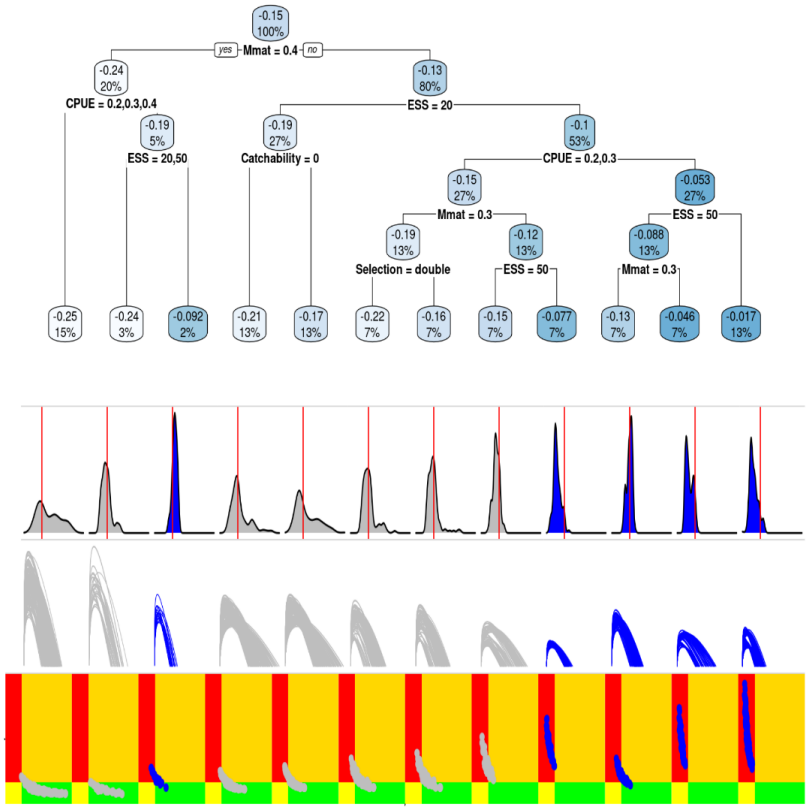
Prediction Skill



IOTC Albacore Stock Synthesis Grid (n = 1440)

European Commission

Albacore Example



Regression Tree for Mohn's ρ

- the darker blue the closer to 0

MASE

- < 1 is better than a random walk

Production functions

Kobe Phase plots

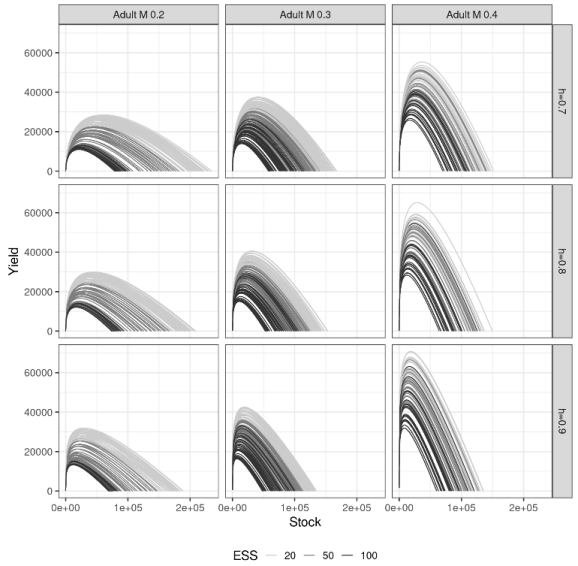
- The better are prediction skill and the retrospective pattern the less productive the stock and the more overfished it is.

Emergent Properties

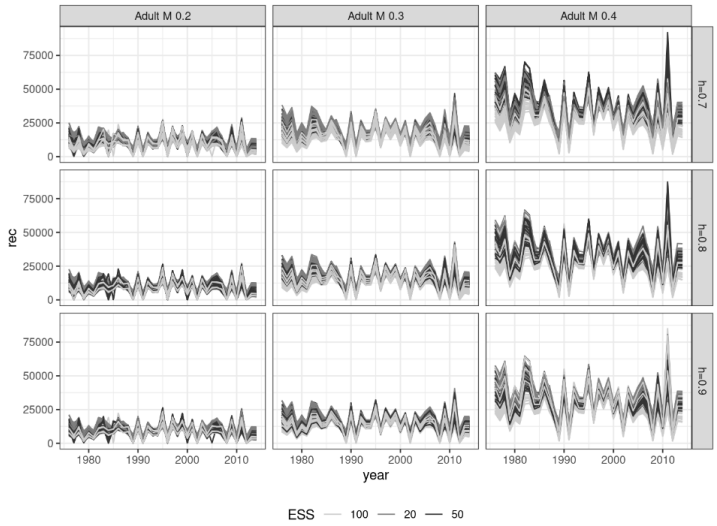
If it looks like a duck, walks like a duck and quacks like a duck, then it is a duck.

- When conducting Management Strategy Evaluation you care about productivity and the form of fluctuations, i.e.
- **Emergent Properties** that manifest themselves as the result of the various system components working together; not by a property of an individual component or model choice in isolation.

Production functions



Process Error: Recruitment

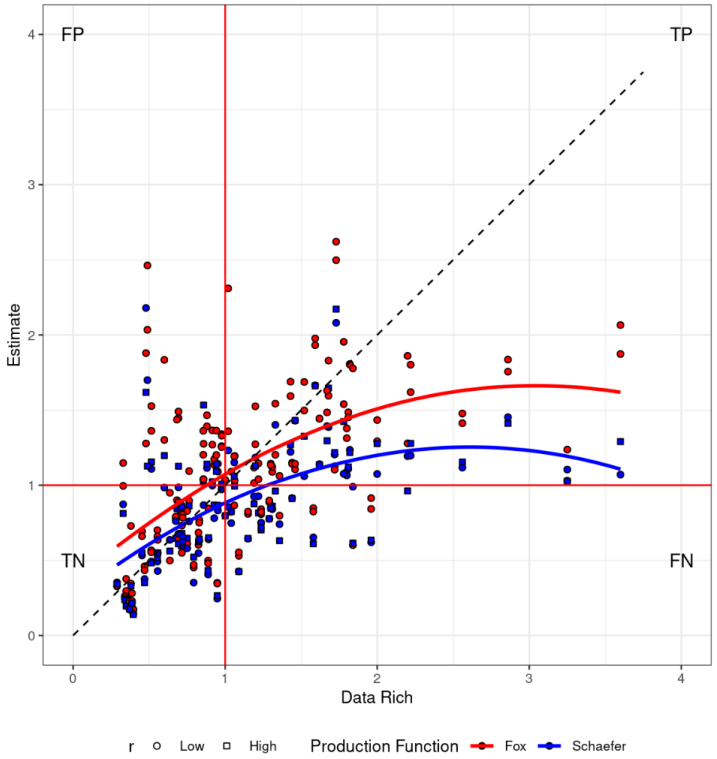


Management Strategy Evaluation

- Can simple skill weighting be used to
  - *Weight and select Operating Models?*
  - *Conditioning Observation Error Models*
- What if two Operating Models A & B conditioned on two indices of abundance I & II are equally plausible?
- But index I only has prediction skill for OM A and index II for OM B?

Is it possible to automate the acceptance-rejection of models for use with large ensembles?

- Supervised learning, maps an input to an output based on example input-output pairs.
- Conduct a cross-test to compare a data-limted assessment to data-rich assessments which are assumed to be **correct**

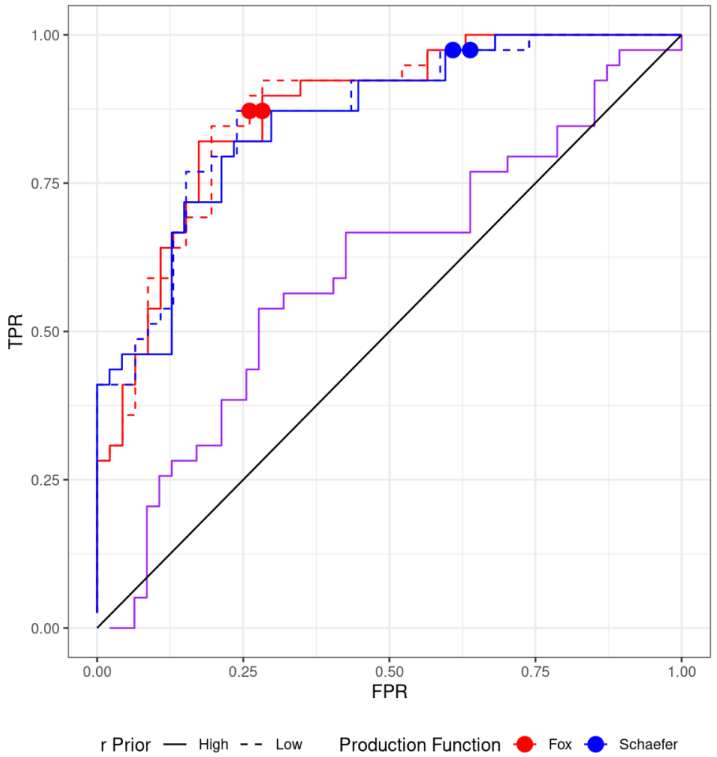


True Skill Score

The true skill score (TSS) is the proportion of true positives less the proportion of false negatives. A perfect prediction would receive a score of 1, random predictions receive a score of 0 and predictions inferior to random ones receive a negative score.

Is it possible to automate the acceptance-rejection of models for use with large ensembles?

- Supervised learning, maps an input to an output based on example input-output pairs.
- Conduct a cross-test to compare a data-limted assessment to data-rich assessments which are assumed to be **correct**



Receiver Operating Characteristics

The area under the receiver operating characteristic curve is a performance measure for machine learning algorithms.

ROC curves plot the True Positive Rate againts the False Positive Rate. The ROC curve is a probability curve, and the area under the curve is important for measuring performance. For example, a coin toss would produce a curve that fell along the $y = x$ line and the area under the curve would be equal to 0.5. The closer the area under the curve is to 1 the better an indicator is at ranking.

Is there a management strategy that relates closely to the type of data we observe and can then test predictions for?

Hindcasting evaluates the model's ability to predict observed data in, for example, a one-step ahead approach. This is a very useful if the observed data is directly related to the management objective, but management quantities (e.g. depletion level relative to that associated with MSY) are usually quite different from the observed data (catch, relative indices of abundance, or catch composition). It might be useful to modify management quantities and objectives to be more closely related to the observations. For example, management could be setting catch under a given (e.g. historically observed) effort level or the catch that would increase the relative index by a certain percentage.

- Why predict latent variables?
  - *If you can predict observations then hopefully you can predict latent variables*
  - *However, if you can not predict observations then can you be confident that you can predict latent variables?*
- I can think of one case where you could use management quantities in a hindcast
  - *If you had a length based assessment, say using simple stock synthesis where the objective was to recover mean size to be greater than Lmat. This could be observed, and predicted.*

Conclusions

- As stock assessment methods become more diverse and complex, the need for best-practice guidelines on model diagnostic criteria is increasingly being recognised.
- Need to be able to validate and compare multiple modelling frameworks
- Objective criteria for model plausibility for stock assessments that are intended for management and that these criteria shall be based on best practice in using model diagnostics for evaluating (1) model convergence, (2) fits to the data, (3) model consistency and (4) prediction skill, and (5) as biological plausibility.
- Specificity of diagnostic tests based on common thresholds (e.g. p-values, Mohn's ρ, MASE) as well as causes and implications of failing one criteria or the other largely remain open questions. To address these emerging questions, the need for simulation testing of model diagnostics is discussed.

References

Deroba, JJ, Doug S Butterworth, RD Methot Jr, JAA De Oliveira, C Fernandez, Anders Nielsen, SX Cadrin, et al. 2015. “Simulation Testing the Robustness of Stock Assessment Models to Error: Some Results from the Ices Strategic Initiative on Stock Assessment Methods.” *ICES Journal of Marine Science* 72 (1). Oxford University Press: 19–30.

Hodges, James S, James A Dewar, and Arroyo Center. 1992. *Is It You or Your Model Talking?: A Framework for Model Validation*. Santa Monica, CA: Rand.

Kell, Laurence T, Rishi Sharma, Toshihide Kitakado, Henning Winker, Iago Mosqueira, Massimiliano Cardinale, and Dan Fu. 2021. “Validation of Stock Assessment Methods: Is It Me or My Model Talking?” *ICES Journal of Marine Science*.

Saltelli, A, D Mayo, R Pielke Jr, T Portaluri, TM Porter, A Puy, I Rafols, et al. 2020. “Five Ways to Ensure That Models Serve Society: A Manifesto.” *Nature* 582 (7813). Springer Nature.

Thygesen, Uffe Høgsbro, Christoffer Moesgaard Albertsen, Casper Willestofte Berg, Kasper Kristensen, and Anders Nielsen. 2017. “Validation of Ecological State Space Models Using the Laplace Approximation.” *Environmental and Ecological Statistics* 24 (2). Springer: 317–39.