AN EVALUATION OF BIAS AND PREDICTION SKILL FOR THE EAST ATLANTIC BLUEFIN STOCK ASSESSMENT

L.T. Kell[1], A. Ben Mhamed[2], T. Rouyer[3] and A. Kimoto[4]

SUMMARY

Stock assessment models are vulnerable to abnormal observations, which may result in biased estimates of parameters, underestimation of uncertainty, and poor prediction skill. This is especially true when the number of observations are relatively small since there are fewer cases to counter abnormalities. It is therefore advisable to identify influential points, explore their impact, and to try and find reasons for their occurrence, e.g. are they due to miscodes, exclusion of important explanatory variables, incorrect model structure, fisher behaviour, management or non-stationarity in biological processes? In this paper we use regression diagnostics, the jackknife and crossvalidation to evaluate the influence of individual observations from the catch per unit effort series used to calibrate the East Atlantic and Mediterranean bluefin Virtual Population Analysis assessment.

RÉSUMÉ

Les modèles d'évaluation des stocks sont vulnérables aux observations anormales (valeurs atypiques), ce qui peut entraîner des estimations biaisées des paramètres, une sous-estimation de l'incertitude et une mauvaise capacité de prédiction. Ceci est particulièrement vrai lorsque le nombre d'observations est relativement faible, puisqu'il y a moins de cas pour prévenir les anomalies. Il est donc conseillé d'identifier les points influents, d'analyser leur impact et d'essayer de trouver les raisons de leur présence, en se demandant par exemple s'ils sont dus à des erreurs de codage, à l'exclusion d'importantes variables explicatives, à la structure incorrecte du modèle, au comportement du pêcheur, à la gestion ou à la non stationnarité dans les processus biologiques ? Dans cet article, nous utilisons des diagnostics de régression, la procédure par eustachage et la validation par recoupement pour évaluer l'influence des observations individuelles de la série de capture par unité d'effort utilisée pour calibrer l'évaluation de l'analyse de population virtuelle du thon rouge de l'Atlantique Est et de la Méditerranée.

RESUMEN

Los modelos de evaluación de stock son vulnerables a observaciones anormales, que podrían dar lugar a estimaciones sesgadas de los parámetros, a una subestimación de la incertidumbre y a una capacidad pobre de predicción. Esto es especialmente cierto cuando el número de observaciones es relativamente pequeño dado que existen menos casos para contar las anormalidades. Por tanto, es aconsejable identificar puntos influyentes, explorar su impacto e intentar hallar razones para que se produzcan, por ejemplo, ¿se deben a códigos erróneos, a la exclusión de importantes variables explicativas, a una estructura incorrecta del modelo, al comportamiento de los pescadores, a la ordenación o a la no estacionalidad en los procesos biológicos? En este documento se utilizan diagnósticos de regresión, el jacknife y la verificación cruzada para evaluar la influencia de las observaciones individuales a partir de las series de capturas por unidad de esfuerzo utilizadas para calibrar la evaluación del análisis de población virtual del atún rojo del Atlántico este y el Mediterráneo.

[1] ICCAT Secretariat, C/Corazón de María, 8. 28002 Madrid, Spain; Laurie.Kell@iccat.int; Phone: +34 914 165 600; Fax: +34 914 152 612.
[2] INRH, Boulevard Sidi Abderrahman Ain Diab 2; a.benmhamed@mail.com
[3] Ifremer Sète, UMR MARBEC, Avenue Jean Monnet 34200 Sète
[4] NRIFSF;5-7-1 Orido Shimizu Shizuoka Japan

KEYWORDS

Bias, Cross Validation, Diagnostics, Jackknife,
Prediction Skill, Stock Assessment, Residuals

Introduction

Many stock assessment models use linear regression, e.g. when using catch per unit effort (CPUE) series used to calibrate a Virtual Population Analysis (VPA), and the evaluation of goodness of fit often relies on inspecting model residuals (i.e. the observed minus the fitted values) looking for patterns, such as outliers, correlations with extra covariates, nonlinearities, correlations between the residuals, heteroscedasticity, and departure from normality.

When fitting models to data a few abnormal data points can skew distributions and have a large influence on results. This is especially true when sample sizes are small since there are fewer cases to counter abnormal points. It is therefore important to identify influential points, explore their impact, and to try and find reasons for their occurrence, e.g. are they due to miscodes, incorrect model structure, exclusion of important explanatory variables, fisher behaviour, management or non-stationarity in biological processes? Throwing away inconvenient data when they do not support a hypothesis, however, is never a good idea, rather any points identified as influential should be double checked.

In stock assessment, however, regression is mainly used as part of an iterative procedure to estimate catchability (q) a nuisance parameter. This may means that even if a point is influential in the regression the impact on the parameters of interest may be less. Also model fit statistics are not always a good guide to how well a model will predict, however, as a high R or low root mean square error (RMSE) does not necessarily mean a good model as it is easy to over-fit a model to data. For example, in a simple polynomial regression better fits to the data can be obtained by adding higher order terms but the predictions from the model on new data will usually get worse as higher order terms are added.

We therefore used cross validation is a method for evaluating the predictive error of a model by testing it on a set of data not used in fitting. However, in stock assessment there is often insufficient data to allow some of it to be kept back for testing. A more sophisticated way to create test datasets is, like the jackkife, to leave out one (or more) observation at a time. Cross validation then compares the difference between the fitted and the predicted values, where the predicted residual is then calculated as the observation minus the out-of-sample predicted value. We also calculate DFBETA a measure of the difference in parameters and quantities of interest with and without an observation, using the jackknife, and use the jackknife to estimate bias.

Material and Methods

The dataset used is the 2014 East Atlantic and Mediterranean bluefin Virtual Population Analysis. Two scenarios were considered by the working group in 2014 corresponding to two catch series. These were those reported to ICCAT and a series that was inflated to take into account unreported catches, the reported and inflated scenarios respectively. Seven series of CPUE were used for calibration (**Figure 1**).

Linear regression

Before identifying abnormal values it is necessary to characterise what is normal. This can be done by examining the overall shape of the data looking for important features and departures from assumptions then identifying observations that are far removed from the rest of the data, e.g. using quantile-quantile, scatter and box plots. Outliers are not necessarily influential, however, for example in a regression analysis if the dataset is very large a single outlier may not have a big effect, while even a small outlier may have a big impact if it is at the extreme end of a covariate's range.

In a regression of Y on X therefore abnormal points are those where there is a large discrepancy between the observed and predicted Y values (outlier) or when they have an unusual X-value (leverage).

The influence of a point is a function of discrepancy and leverage, e.g.

Influence = Discrepancy X Leverage

In a linear regression outliers can be identified using studentized residuals and quantile-comparison plots, leverage can be assessed by exploring the hat-values, and influence by calculating Cooks distance or DFBeta (Cook and Weisberg, 1982; Williams, 1987). Data points with large residuals and/or high leverage may distort the outcome and hence accuracy of an analysis and Cook's distance and DFBeta measure the effect of deleting a given observation.

Jackknife

The jackknife is a resampling technique that is a special case of the bootstrap and can be used to identify points with a large influence and bias in statistical estimates. When using the jack knife the i^{th} observation is omitted in turn from the sample dataset and the parameter re-estimated, this allows the influence of individual points to be evaluated. The jackknife can also provide an estimate of sampling bias, for example the estimate of a parameter θ_i, may tend to come out on the high or low side of its true value if a data sample is too small. Thus the estimate θ derived from a fit to N data points may be higher or lower than the true value. When this happens removing a measurement, using the jackknife will increase the bias. The effect can be measured by comparing the mean of the jackknife values θ_i^J, call it $\theta_.^J$ with the result θ_m^{Ji} of fitting the full dataset.

All analysis was conducted using R and FLR and the diags package which provides a set of common methods for reading these data into R, plotting and summarising them.(http://www.flr-project.org/)

Results

Regression Diagnostics

First the CPUE residuals used in the VPA calibration were examined to look for departures from model assumptions, then the jackknife results are summarised before comparing them to the regression diagnostics.

Figure 1, show the CPUE time series, the continuous black line is a lowess smother showing the average trend by area. The fit to the indices can be evaluated by plotting the observed against the predicted values (**Figure 2**), if the index is assumed to be proportional to stock size the points should fall either side of the y = x line. Departures from the assumption that the index is proportional to the stock can be seen by plotting the residuals by time (**Figure 3**). Autocorrelated residuals may mean that the estimated parameters are biased, and can be checked by plotting the residuals against each other with a lag of 1 (**Figure 4**). The error distribution is then checked by plotting the observed and the predicted quantiles for the normal distribution (**Figure 5**), and the variance by plotting residuals against the fitted values (**Figure 6**). Finally the residuals diagnostics are summarised in **Figures 7** and **8** which shows the residuals, sigma, CVs and Cooks Distance for the CPUE series fitted to the two VPA scenarios.

Although the residuals do appear to be normally distributed, the CPUE series do not appear to be good proxies for relative abundance and there are strong residual patterns. The diagnostics are similar for the two scenarios. The influence of the data points are summarised in **Figure 9** which plots each point by leverage and discrepancy, with the size of the bubble corresponding to Cooks distance. The results are similar for Spanish bait boat across the two scenarios, but for the other series there are differences seen in both residual magnitude, leverage and influence.

Jackknife Diagnostics

The jackknife estimates of SSB/B_{MSY} , F/F_{MSY} r and R_t and summarised in **Figures 10, 11, 12** and **13**, by year and CPUE series. Where r is the population growth rate at low abundance levels, and R_t the time to recover to B_{MSY} at F_{MSY}, in **figure 13** negative values indicate stock is above SSB_{MSY} and therefore has to decline to reach the target. The jackknife estimates are then compared with the regression residuals in **Figure 14**. Then the jackknife estimates of SSB/B_{MSY} , F/F_{MSY} r and R_t for the two scenarios are compared in **Figures 15, 16, 17** and **18**. The same data are compared using their ranks in **Figures 19, 20, 21** and **22**. The ranking shows that the most influential points can be identified using Cooks distance for some but not all CPUE series.

The estimates of bias and variance are compared in **Table 1** for the four summary statistics.

Cross Validation

Figures 23, 24, 25 and 26 show the prediction residuals by CPUE series and year for the reported and inflated catch Scenarios**.** The boxplots for the types of residuals are quite similar. When the residuals are plotted by year, however, a significant bias is seen in the recent years which tends to be worse for the prediction residuals.

Discussion

There are two forms of error when assessing model predictions, i.e. error due to bias and error due to variance. The error due to bias is the difference between the predictions from a model and the correct values and error due to variance is the variability of a model predictions regardless of how correct they are. Causes of bias are many and include to non-linearity, non-stationarity, non-normal error distributions, sampling bias, survivorship bias, under or overfitting, missing variable or incorrect model structure.

In the Kobe framework mainly error due to variance is considered, although error due to bias is also important, since the kobe phase plot shows stock status relative to reference points and bias may mean a stock is overfished but the advice is that the stock is being safely exploited. There are many potential causes of bias in stock assessment, however, it is difficult to compare model predictions with actual values. One way to do this is to compare observations with their out-of-sample predictions (Kell *et al.*, 2016). If a model can not predict the observations it is unlikely to be able to predict the underlying dynamics.

A comparison of the jackknife estimates (**Table 1**) showed a large bias for F/F M SY in the Inflated catch scenario. In the jackknife example it is assumed that the model was correct, and that bias was due to small sample size for the CPUE data, an assumption for the jackknife, like the bootstrap, is that the residuals are I.I.D. The plot showed, however, that the CPUE series are not good proxies for stock abundance as there are strong patterns in the residuals, and the residuals are also autocorrelated. This means that the jackknife and the model assumptions are violated, and the parameter estimates and the estimates of variance and bias are likely to be incorrect.

Although CPUE was used as a proxy for relative abundance it has long been recognised that such time series may not accurately reflect trends in population abundance (e.g. Beverton and Holt, 1993; Harley *et al.*, 2001; Maunder *et al.*, 2006; McKechnie *et al.*, 2013; Polacheck, 2006). Particularly since factors that affect changes in the spatial distribution of populations and the allocation of effort in response to management and economic drivers can affect catch and effort independent of stock abundance (e.g. Paloheimo and Dickie, 1964; Tidd *et al.*, 2011). Interactions between changes in oceanographic conditions and exploitation can drive spatial and temporal dynamics (Fromentin *et al.*, 2013), and a variety of factors may affect fish and fisher behaviour so that CPUE is not proportional to stock abundance (Hilborn *et al.*, 1992). For example hyperstability where catches remain high as a population declines (Erisman *et al.*, 2011) and hyperdepletion where catches decline faster than the population (?). Particularly since fishing is an economic activity and catch should be thought of as revenue and effort as costs which are affected by markets, management regulations, labour cost, interest rates and technology.

The jackknife showed that there was substantial bias in the estimate of $F:F_{MSY}$ for the inflated catch scenario; although for other derived statistics such as SSB :B_{MSY}, population growth rate and time to recover to B_{MSY} the bias was less. The cross validation showed that model residuals underestimate the bias in the recent period. It is likely therefore that one-step predictions would show a large bias and indicate poor ability of the models to provide advice on future stock status.

Although performing a jackknife and leave one out cross validation is relatively quick for VPA, which only considers error in the CPUE series, it will become more computer intensive for models which allow for error in the catch-at-age or catch-at-size. Instead a leave one out cross validation could be done for the recent years as this would allow models with poor forward prediction ability to be identified. Stock assessment models are primarily used to provide advice on future status and current management regulations, if model can not predict next years observations then it is unlikely to be able to describe the underlying dynamics. Alternatively blocks of data could be left out.

Conclusions

- The jackknife and cross validation was applied to a model with relatively few parameters and showed problems due to both bias and poor prediction skill. In models with many more parameters, where overfitting is likely to be a problem, step one ahead prediction errors would be a useful tool for model evaluations. This is important when fitting model to provide advice, since if we can not predict actual observations how can we predict the stock dynamics which are hidden from view?

- Under the GBYP an Operating Model (OM) is being developed and conditioned on stock assessment data. A first step is to select hypotheses about the dynamics of the bluefin resource, then to weight and reject these (Punt and Donovan, 2007). If hypotheses are not supported by the data then there is a danger that the chosen Management Procedure will not be robust to uncertainty. Cross validation is therefore an important tool in selecting OMs.

- A lot of time is spent examining standardised CPUE series and discussing the importance of individual points in stock assessment working groups. Identifying whether a point or series is influential will help focus discussions. Influence can be measured using regression statistics, from fits of CPUE to biomass to estimate q. However, q is a nuisance parameter estimated as part of an iterative procedure so even if a point may influence q it this soed not mean it will influence quantities of interest such as $F/F_{MSY.}$ Instead influence can be estimate using DFBeta, i.e. the effect of removing a point on F/F_{MSY}.

- The jackknife allows bias to be estimated, and is a leave one out procedure and so can also be used to estimate the prediction residuals, by comparing out of sample value with model predictions.

References

Beverton, R. and Holt. S. 1993. On the dynamics of exploited fish populations, volume 11. Springer.

Cook, R. D. and Weisberg, S. 1982. Residuals and influence in regression. New York: Chapman and Hall.

Erisman, B. E., Allen, L. G., Claisse, J. T., Pondella, D. J., Miller, E. F., Murray, J. H. and Walters, C. 201.1The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Can. J. Fish. Aquat. Sci., 68(10):1705–1716.

Fromentin, J.-M., Reygondeau, G., Bonhommeau, S. and Beaugrand, G. 2013. Oceanographic changes and exploitation drive the spatio-temporal dynamics of Atlantic bluefin tuna (*Thunnus thynnus*). Fisheries Oceanography,

Harley, S. J., Myers, R. A. and Dunn, A. 2001. Is catch-per-unit-effort proportional to abundance? Can. J.Fish. Aquat. Sci., 58(9):1760–1772.

Hilborn, R., Walters, C. J. *et al.*1992. Quantitative fisheries stock assessment. Choice, dynamics and uncertainty. Reviews in Fish Biology and Fisheries, 2(2):177–178.

Kell, L. T., Kimoto, A. and Kitakado., T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries Research, 183:119–127.

Maunder, M. N., Sibert, J. R., Fonteneau, A., Hampton, J., Kleiber, P. and Harley, S. J. 2006. Interpreting catch per unit effort data to assess the status of individual stocks and communities. ICES J. Mar. Sci., 63 (8):1373–1385.

McKechnie, S., Hoyle, S. and Harley, S. 2013. Longline CPUE series that account for changes in the spatial extent of fisheries. Technical report, WCPFC-SC9-2013/SA-IP-05.

Paloheimo, J. and Dickie. L. 1964. Abundance and fishing success. Rapports et Proces-Verbaux des Runions du Conseil International pour l'Exploration de la Mer, 155,

Polacheck., T. 2006. Tuna longline catch rates in the Indian Ocean: Did industrial fishing result in a 90% rapid decline in the abundance of large predatory species? Marine Policy, 30(5):470–482.

Punt, A. and Donovan., G. 2007. Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission. ICES J. Mar. Sci., 64(4):603–612.

Tidd, A.A.N., Hutton, T., Kell, L., and Padda. G. 2011. Exit and entry of fishing vessels: an evaluation of factors affecting investment decisions in the North Sea English beam trawl fleet. ICES J. Mar. Sci., 68(5): 961–971.

Williams., D. 1987. Generalized linear model diagnostics using the deviance and single case deletions. Applied Statistics, pages 181–191.

Table 1. Summary of reported catch Scenario summary statistics.

Param	*Fit*	*Mean*	*SE*	*CV*	*Bias*	*Rel.Bias*
$SSB:B_{MSY}$	**1.745**	**1.747**	**0.3144**	**0.18**	**-0.4329**	**-0.2478**
$F:F_{MSY}$	**0.4664**	**0.4657**	**0.1197**	**0.257**	**0.13**	**0.27**
r	**0.38**	**0.3837**	**0.006905**	**0.018**	**0.04802**	**0.1252**
R_t	**-2.98**	**-2.98**	**0.5108**	**-0.171**	**0.81**	**-0.2726**

Table 2. Summary of inflated catch Scenario summary statistics.

Param	*Fit*	*Mean*	*SE*	*CV*	*Bias*	*Rel.Bias*
$SSB:B_{MSY}$	**1.7**	**1.702**	**0.3365**	**0.1977**	**-0.3608**	**-0.212**
$F:F_{MSY}$	**0.4323**	**0.4287**	**0.1336**	**0.3116**	**0.64**	**1.5**
r	**0.38**	**0.3841**	**0.01111**	**0.02891**	**-0.0294**	**-0.07654**
R_t	**-3.01**	**-3.012**	**0.5821**	**-0.1933**	**-0.15**	**-0.04847**

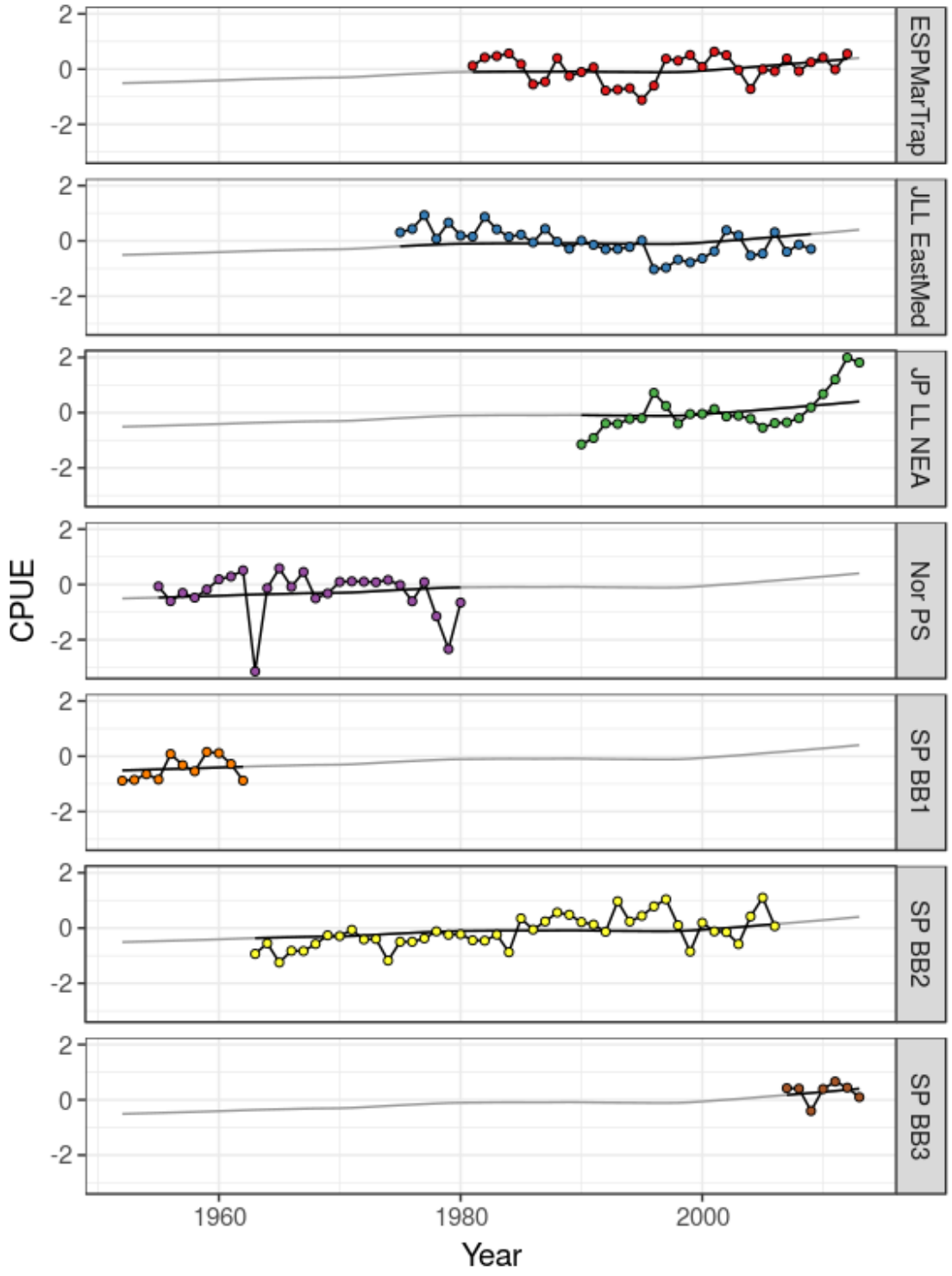

Figure 1. Time series of CPUE indices, continuous black line is a loess smother showing the average trend by area (i.e. fitted to year for each area with series as a factor).

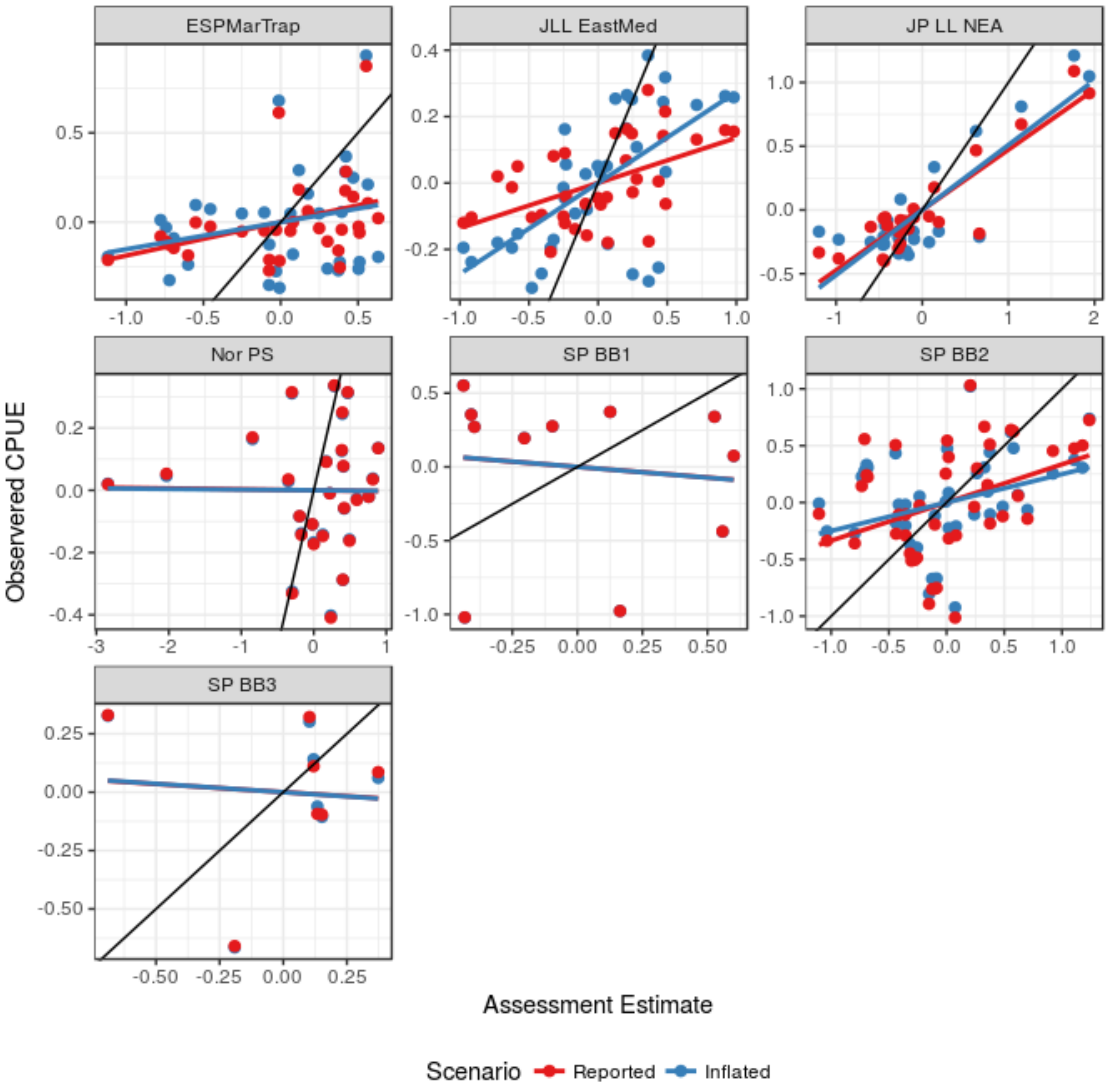

Figure 2. Observed CPUE verses fitted, blue line is a linear regression fitted to points, black the y=x line.

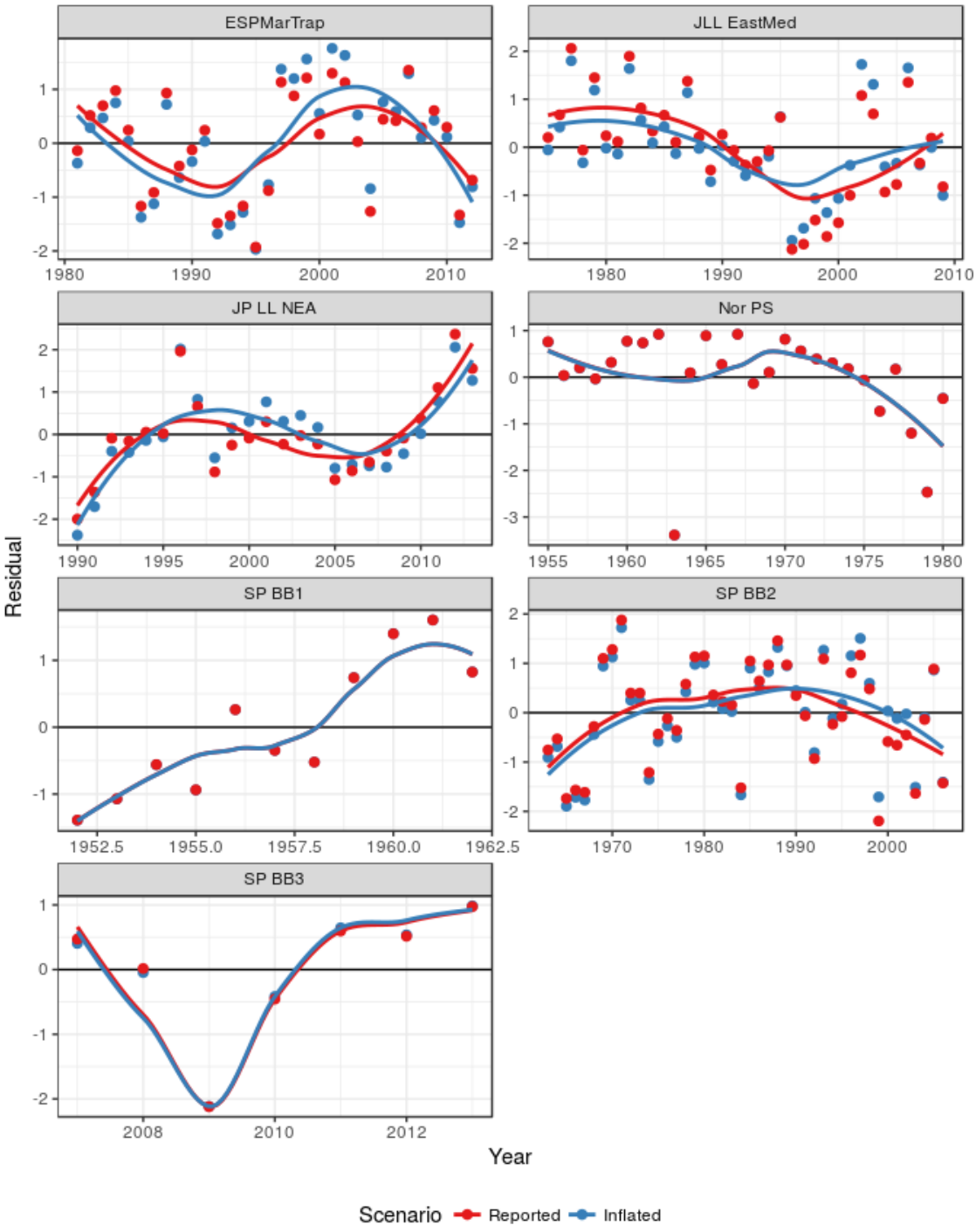

Figure 3. Plot of residuals by year, lines are loess smoothers.

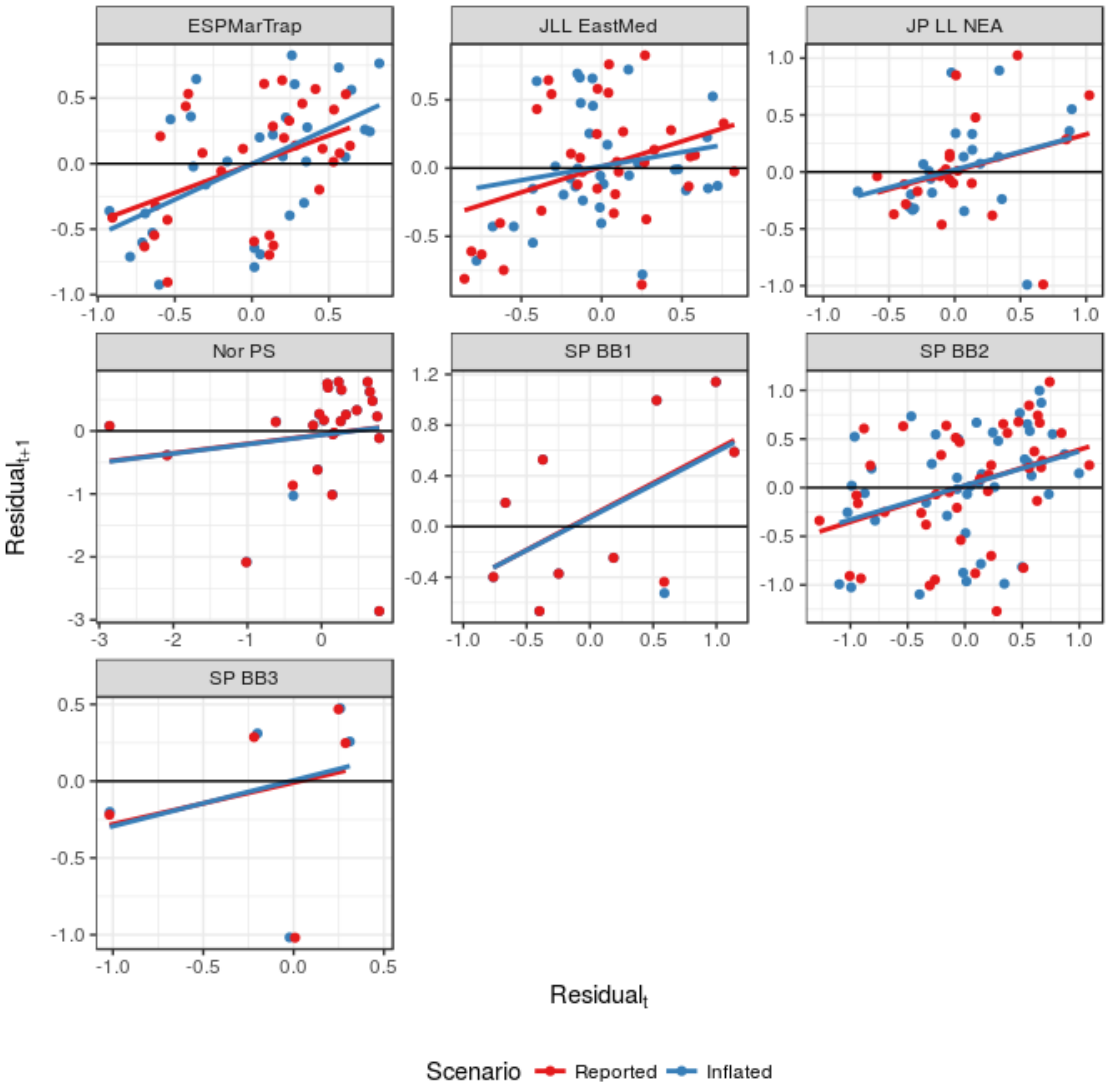

Figure 4. Plot of autocorrelation, i.e. $residual_{t+1}$ versus $residual_t$, line is a linear regression through points.

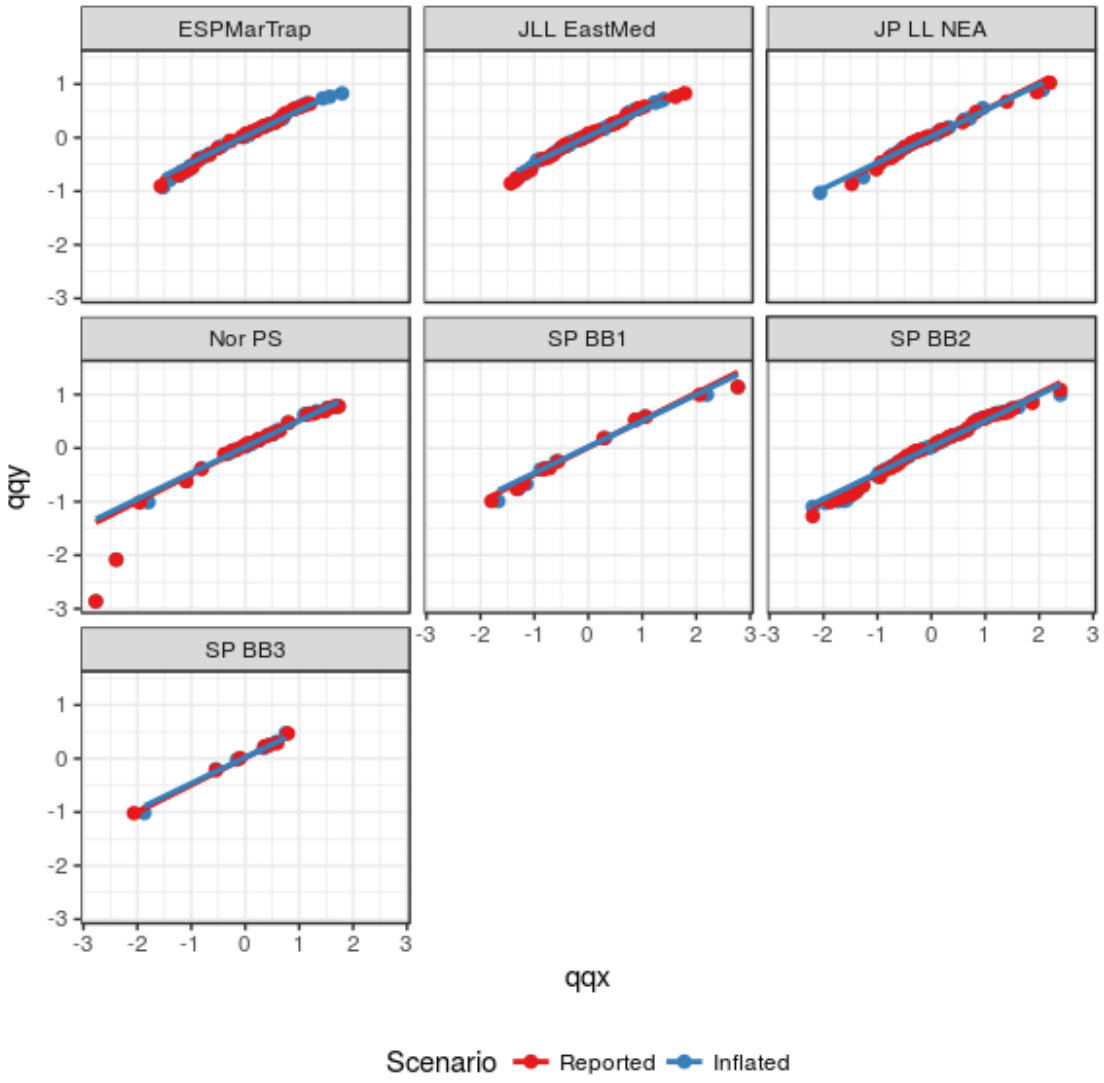

Figure 5. Quantile-quantile plot to compare residual distribution with the normal distribution, line is a linear regression.

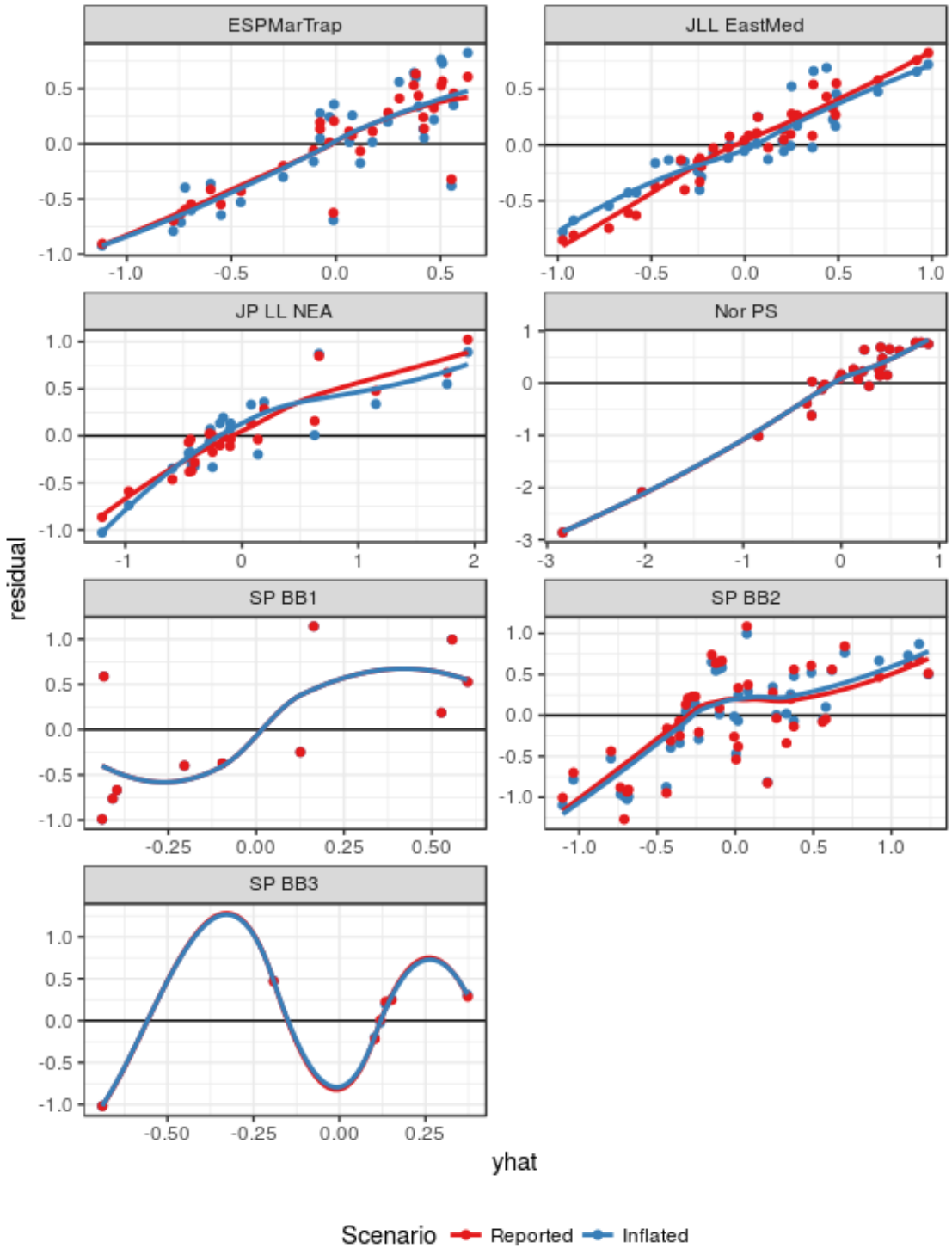

Figure 6. Plot of residuals against fitted value, to check variance relationship, lines are loess smoothers.

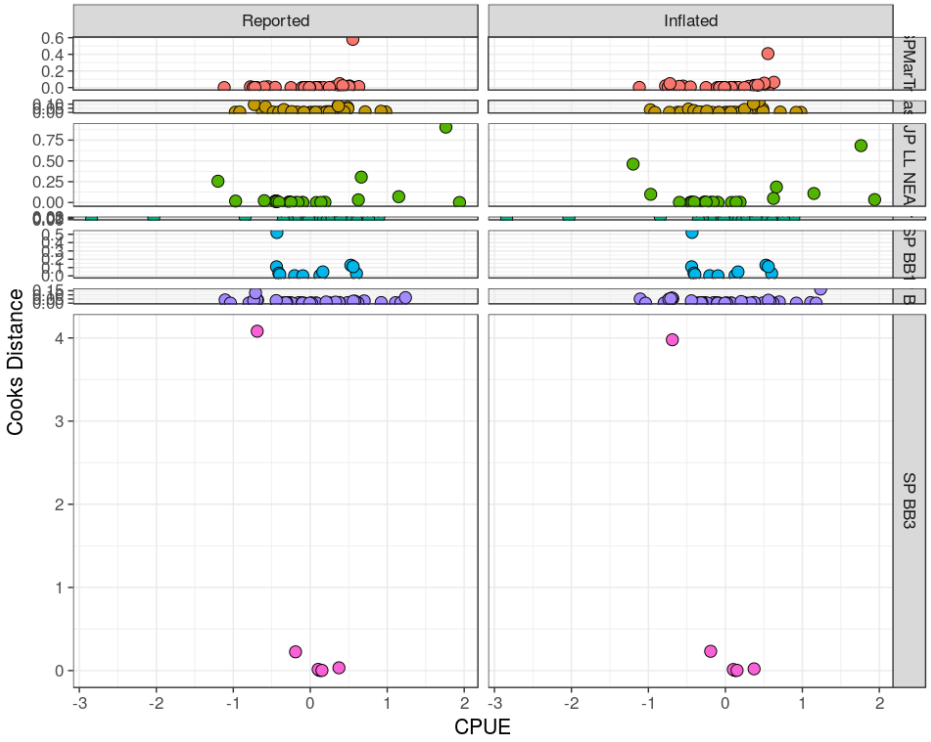

Figure 7. Summary plots of Cooks Distance for the CPUE series fitted to the two VPA Scenarios.

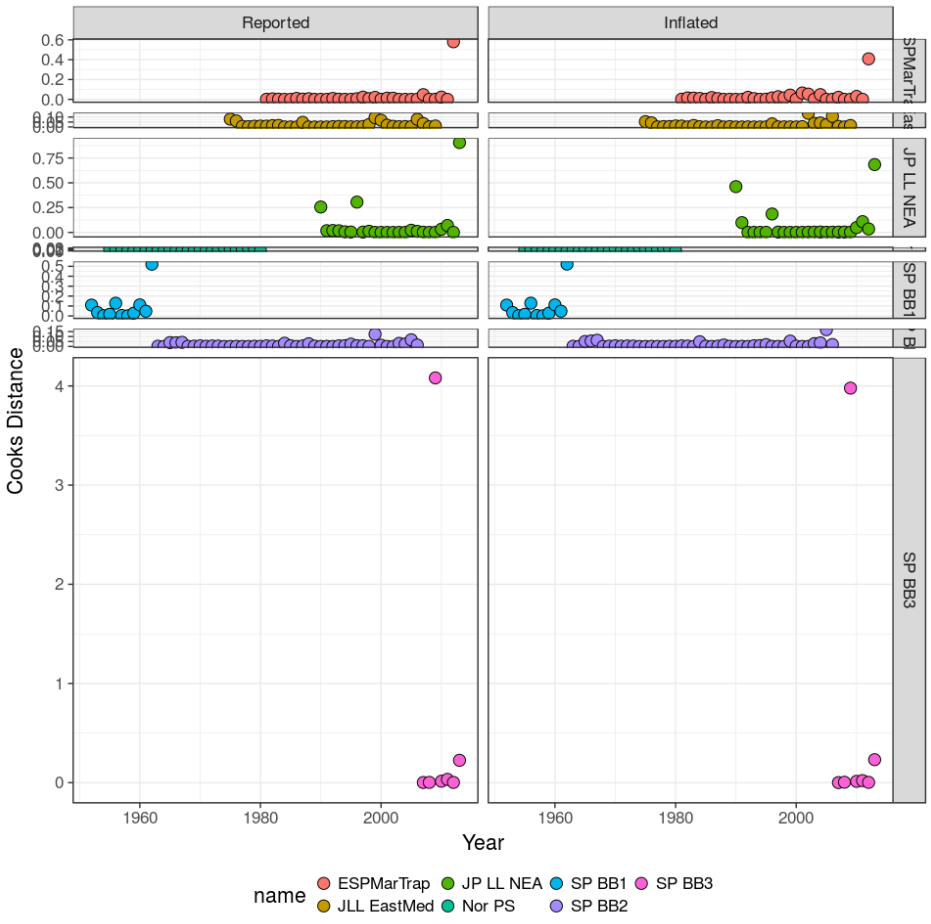

Figure 8. Summary plots of Cooks Distance for the CPUE series by year fitted to the two VPA Scenarios.

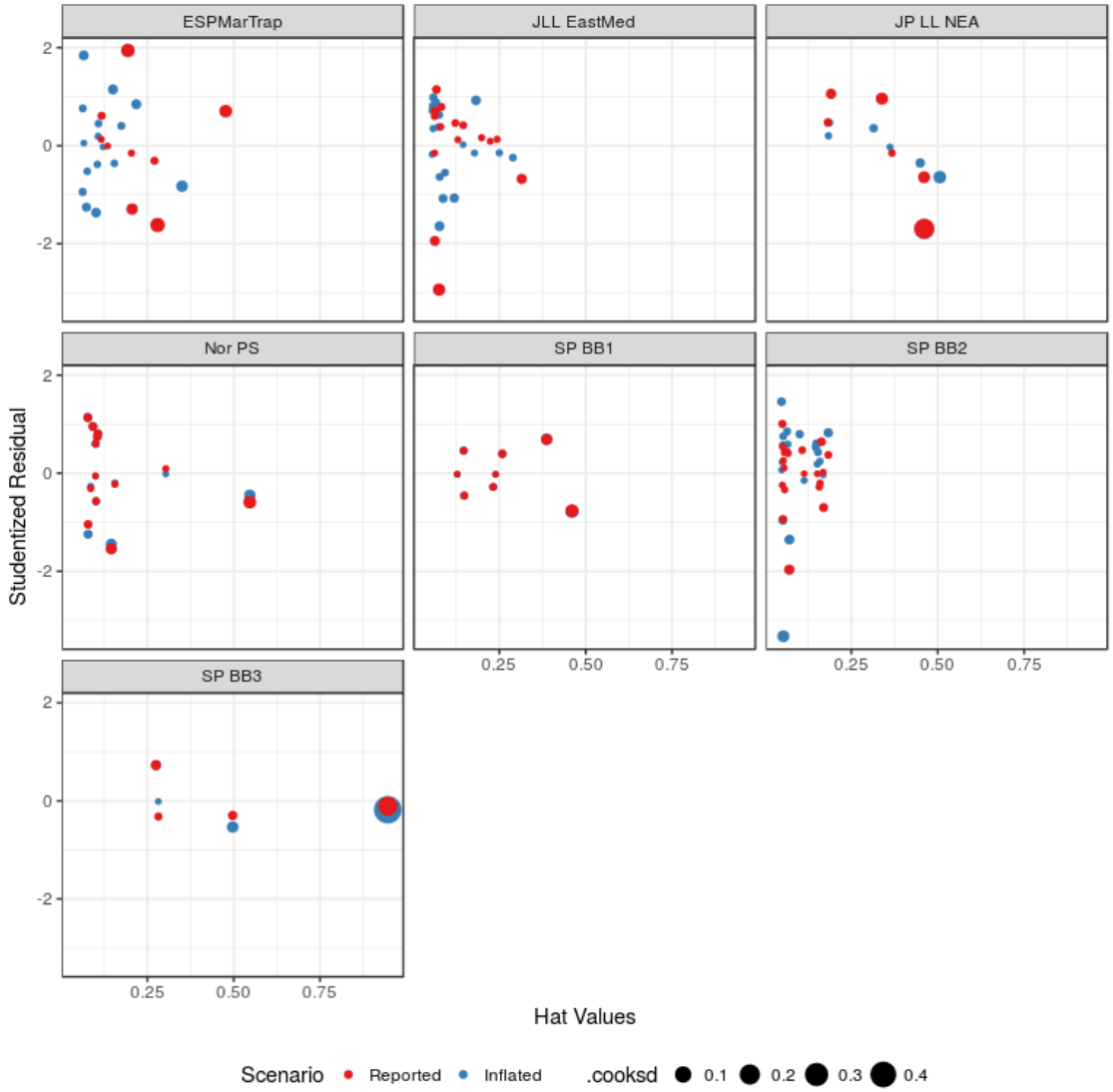

Figure 9. Influence plots plotting residuals against hat values (i.e. leverage), size of bubble show the Cooks distance (i.e. the leverage).

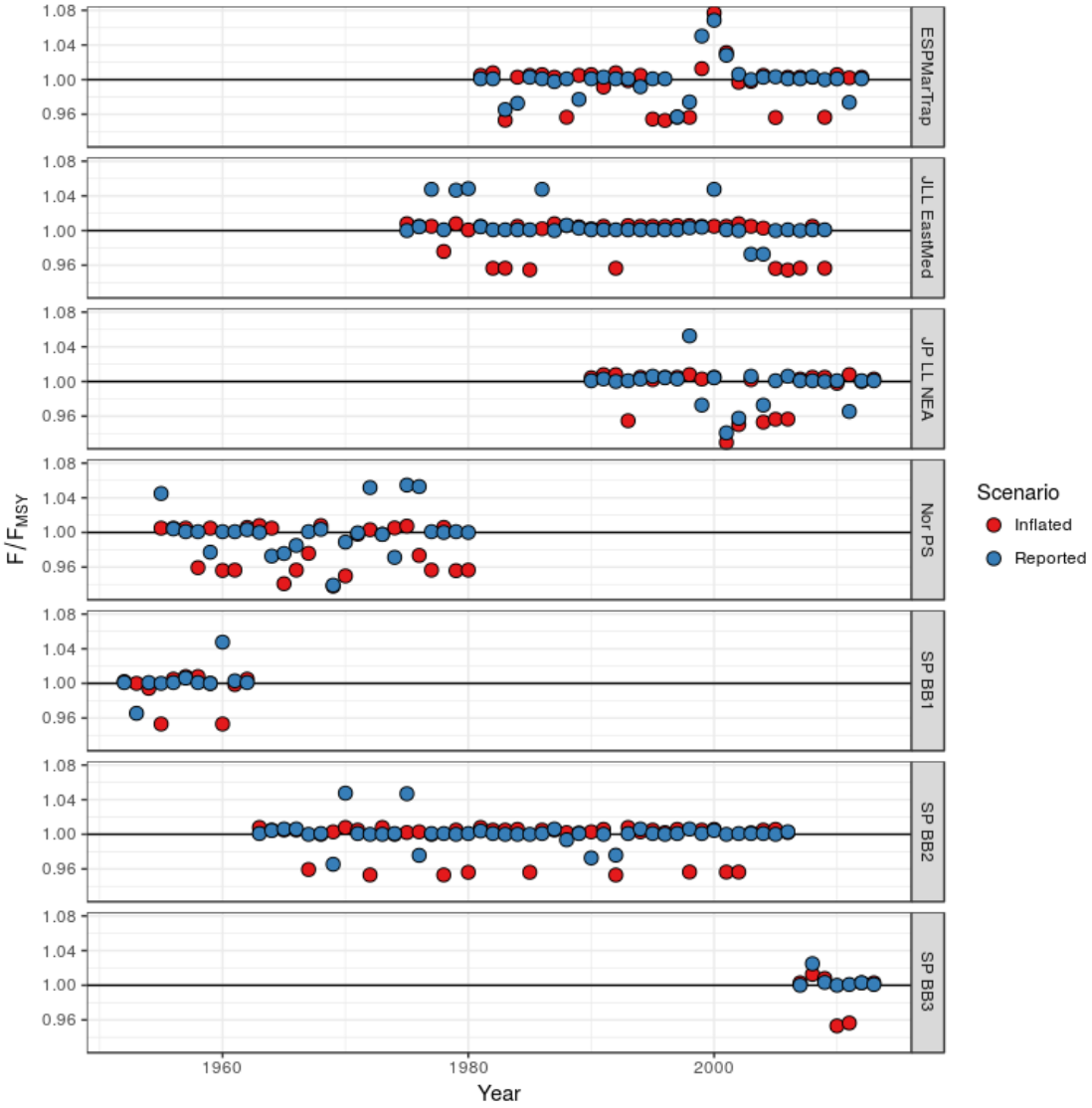

Figure 10. Jackknife estimates of DFBETA for F/F_{MSY} by year and CPUE series.

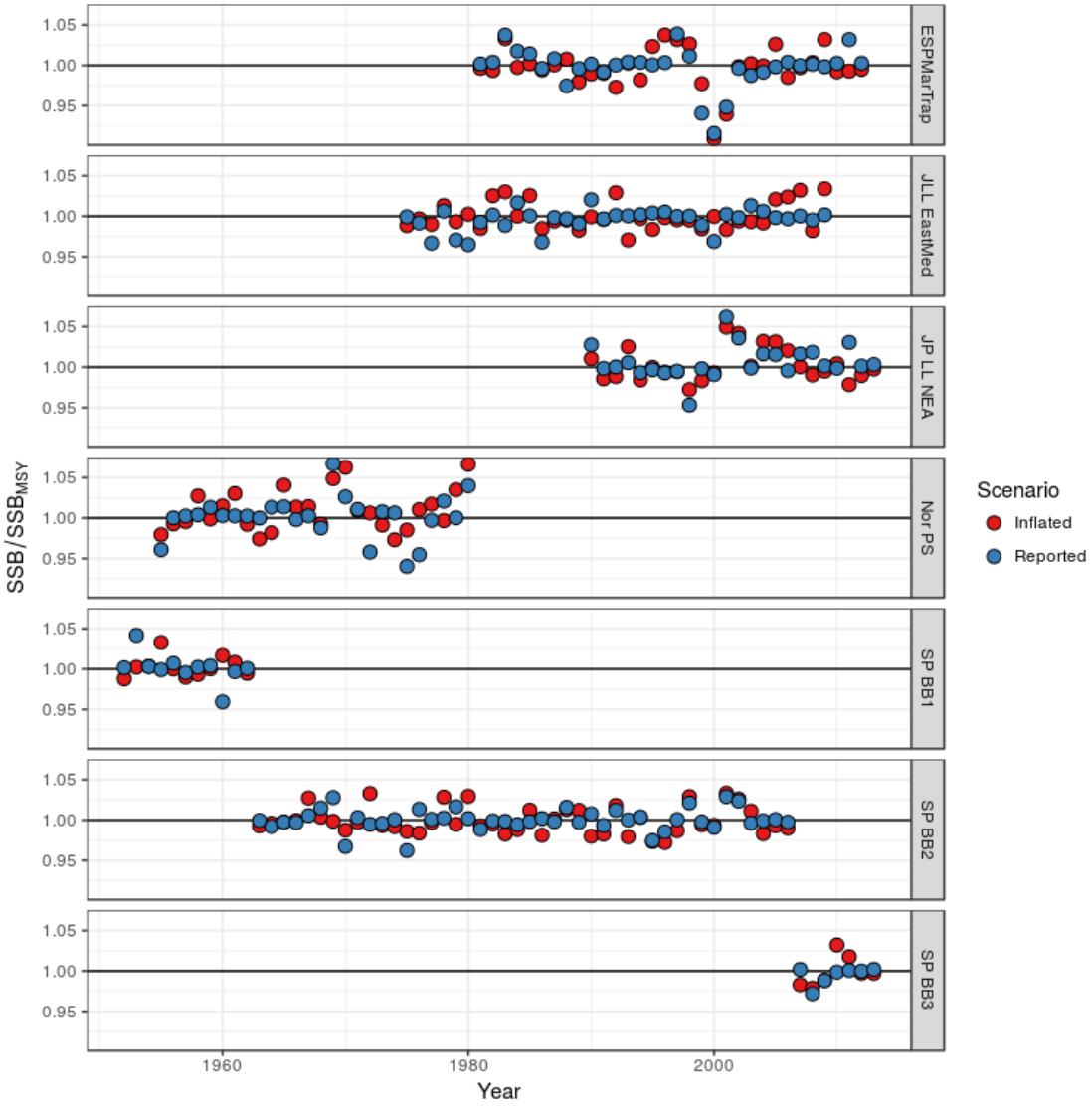

Figure 11. Jackknife estimates of DFBETA for B/ B_{MSY} by year and CPUE series.

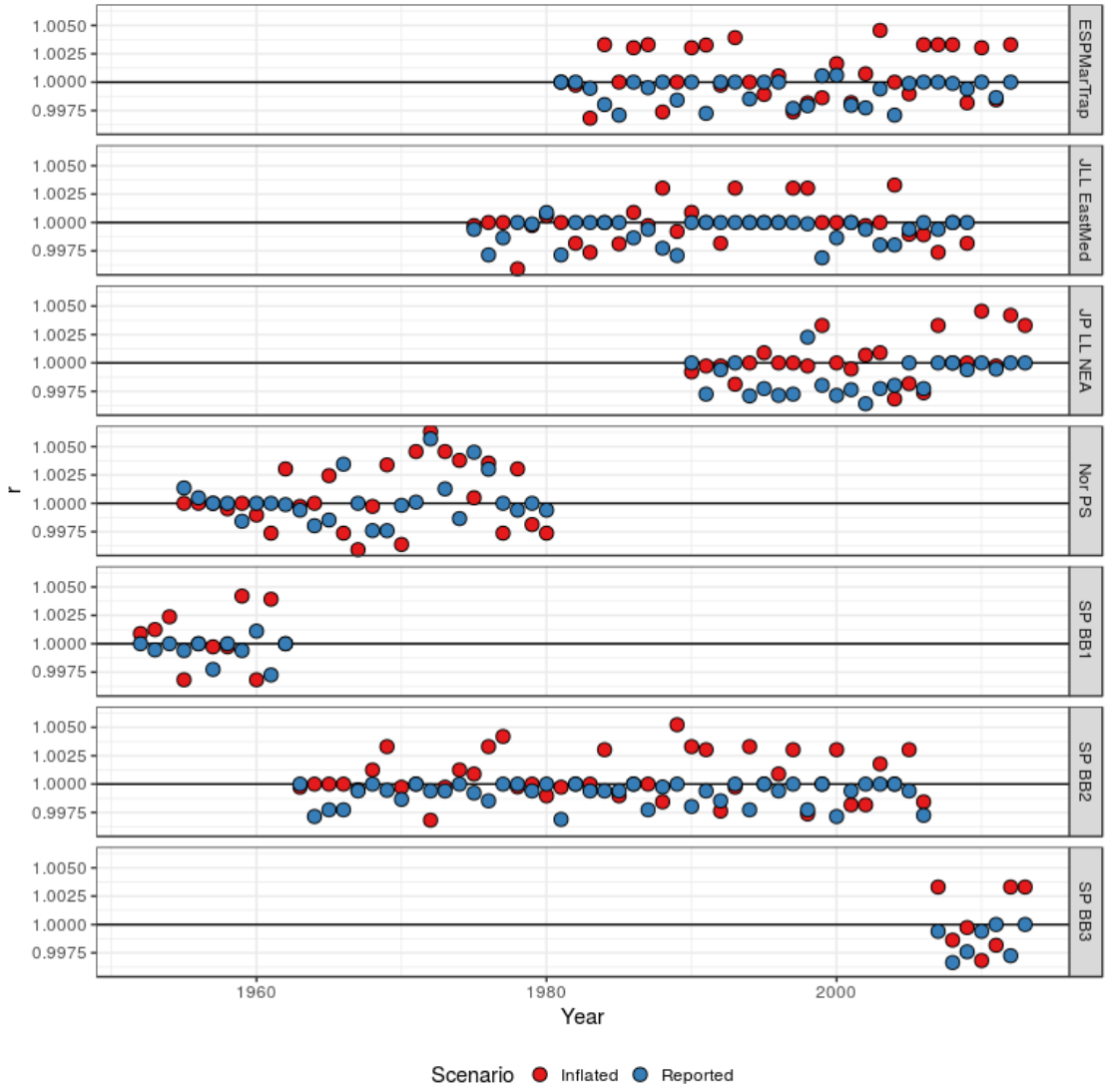

Figure 12. Jackknife estimates of DFBETA for population growth rate *r* by year and CPUE series.

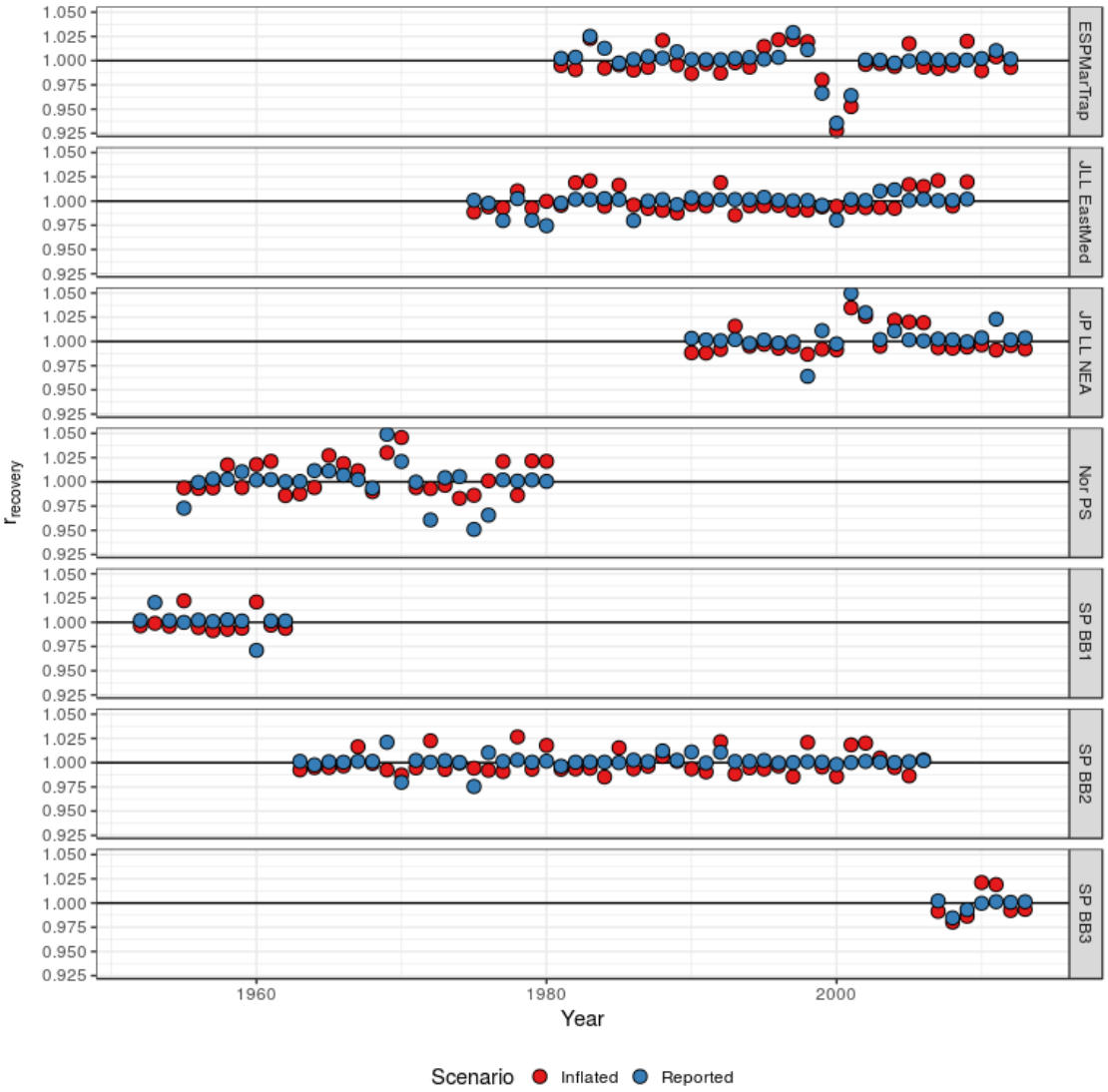

Figure 13. Jackknife estimates of DFBETA for time taken to recover to B_{MSY} (negative values indicate stock is above B_{MSY}) by year and CPUE series.

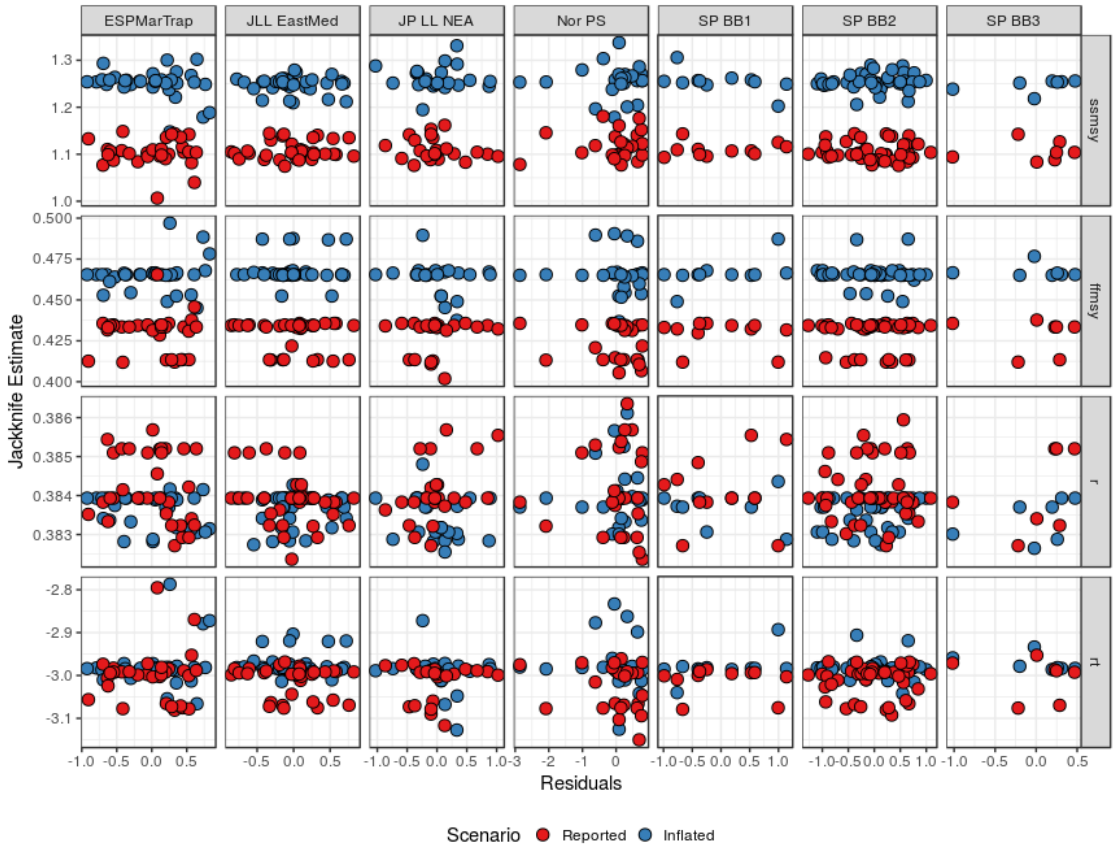

Figure 14. Comparison of jackknifed estimates with residuals.

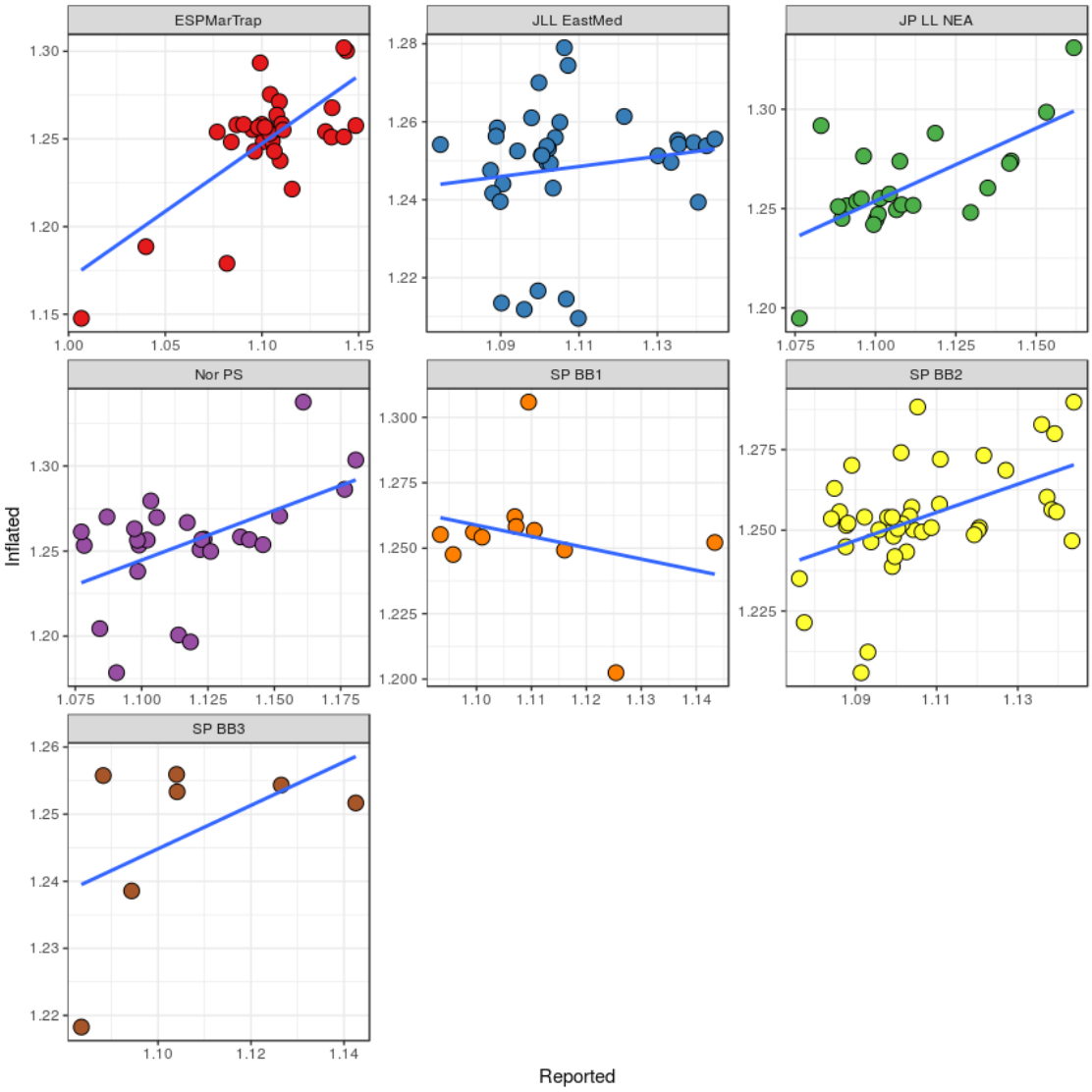

Figure 15. Comparison of jackknife estimates of population growth rate B/B_{MSY} by year and CPUE series.

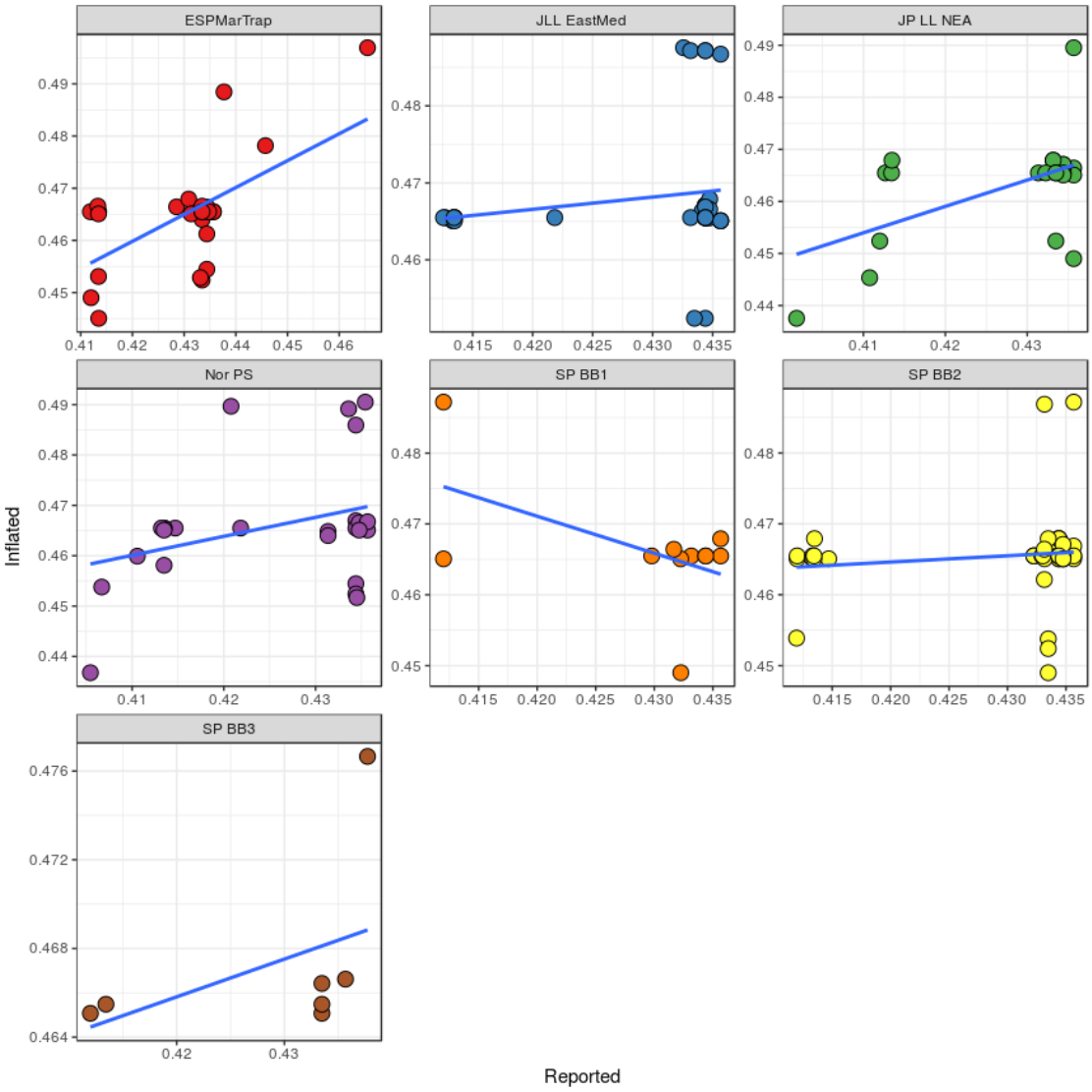

Figure 16. Comparison of jackknife estimates of population growth rate F/F_{MSY} by CPUE series.

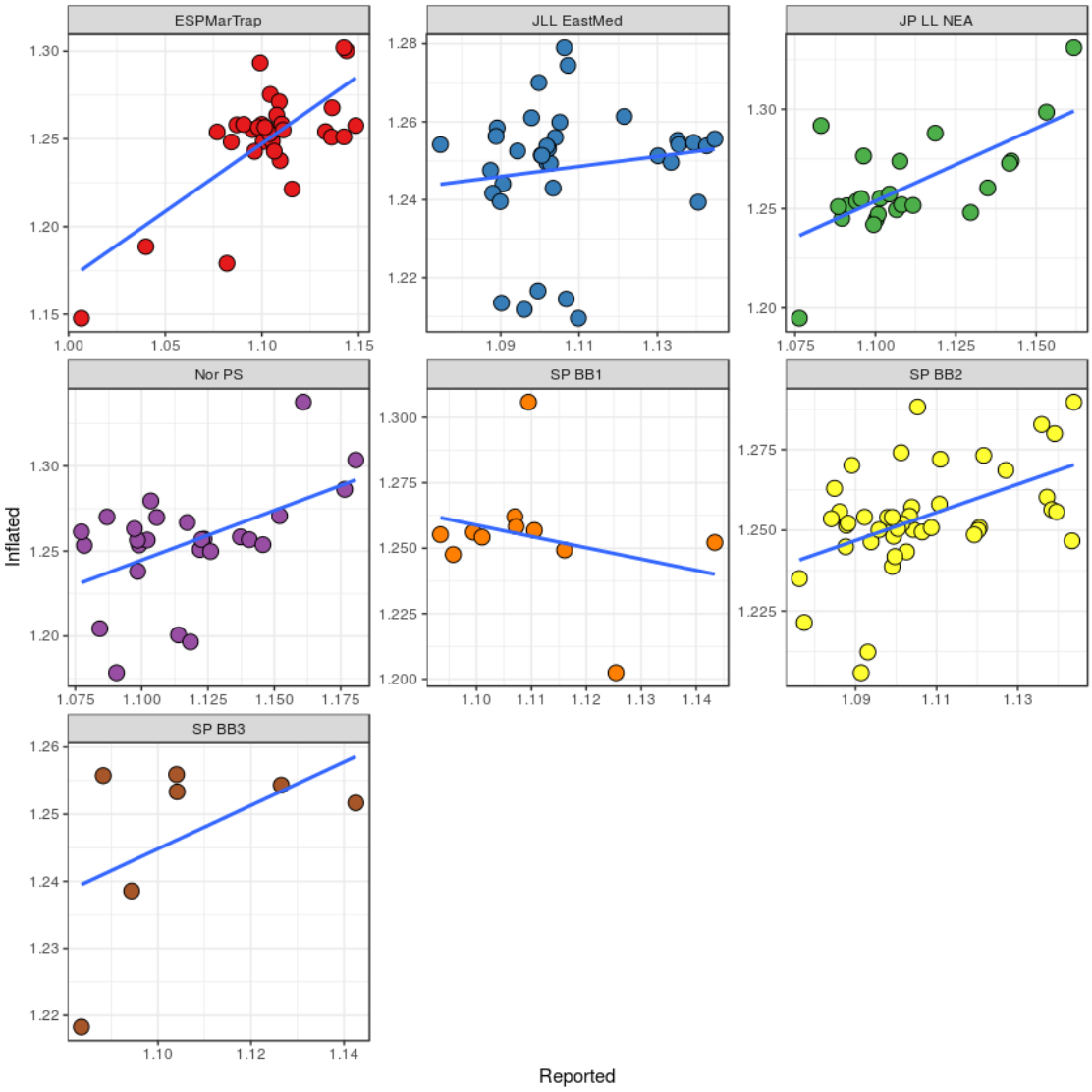

Figure 17. Comparison of jackknife estimates of population growth rate *r* by CPUE series.

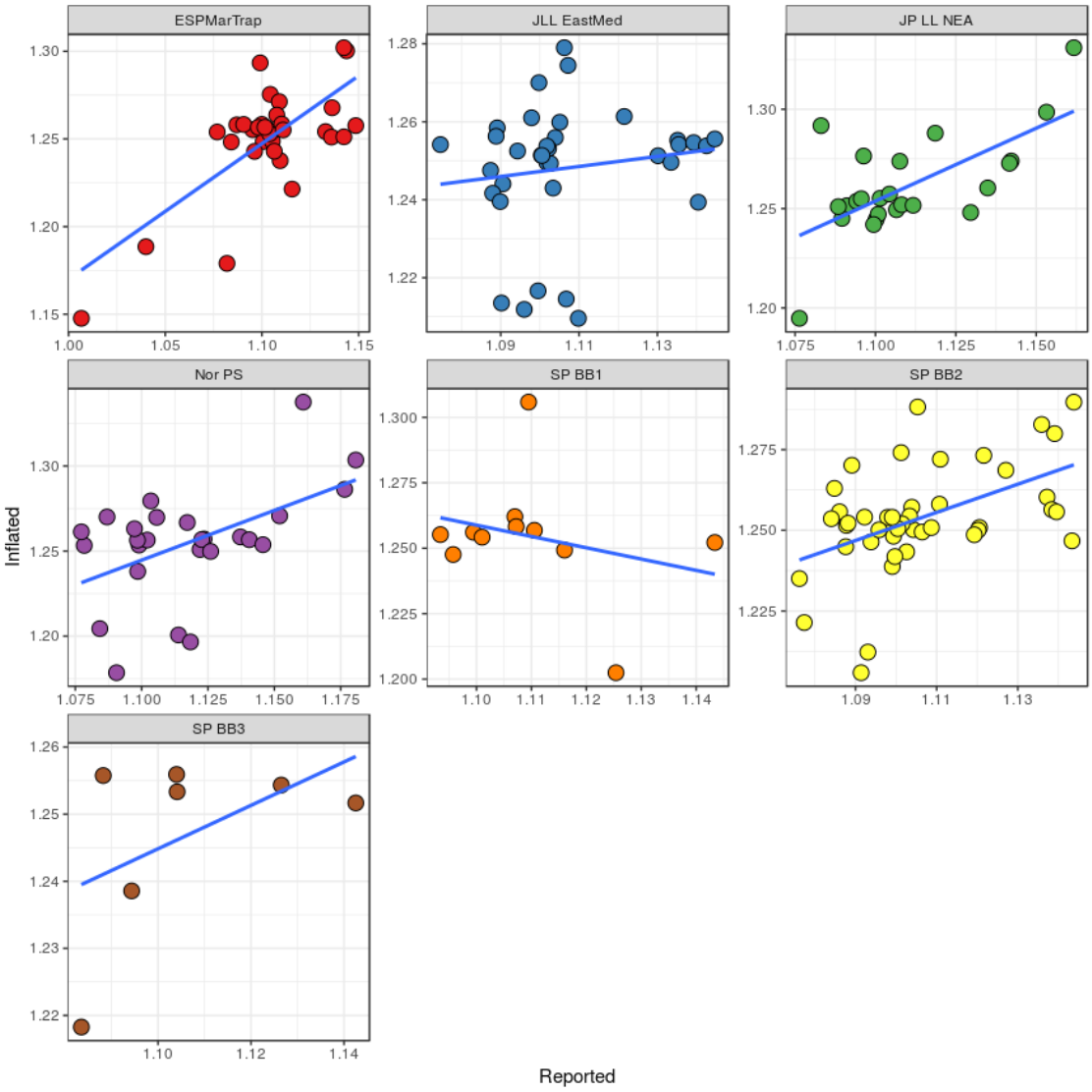

Figure 18. Comparison of jackknife estimates of time taken to recover to B/B_{MSY} (negative values indicate stock is above B_{MSY}) by CPUE series.

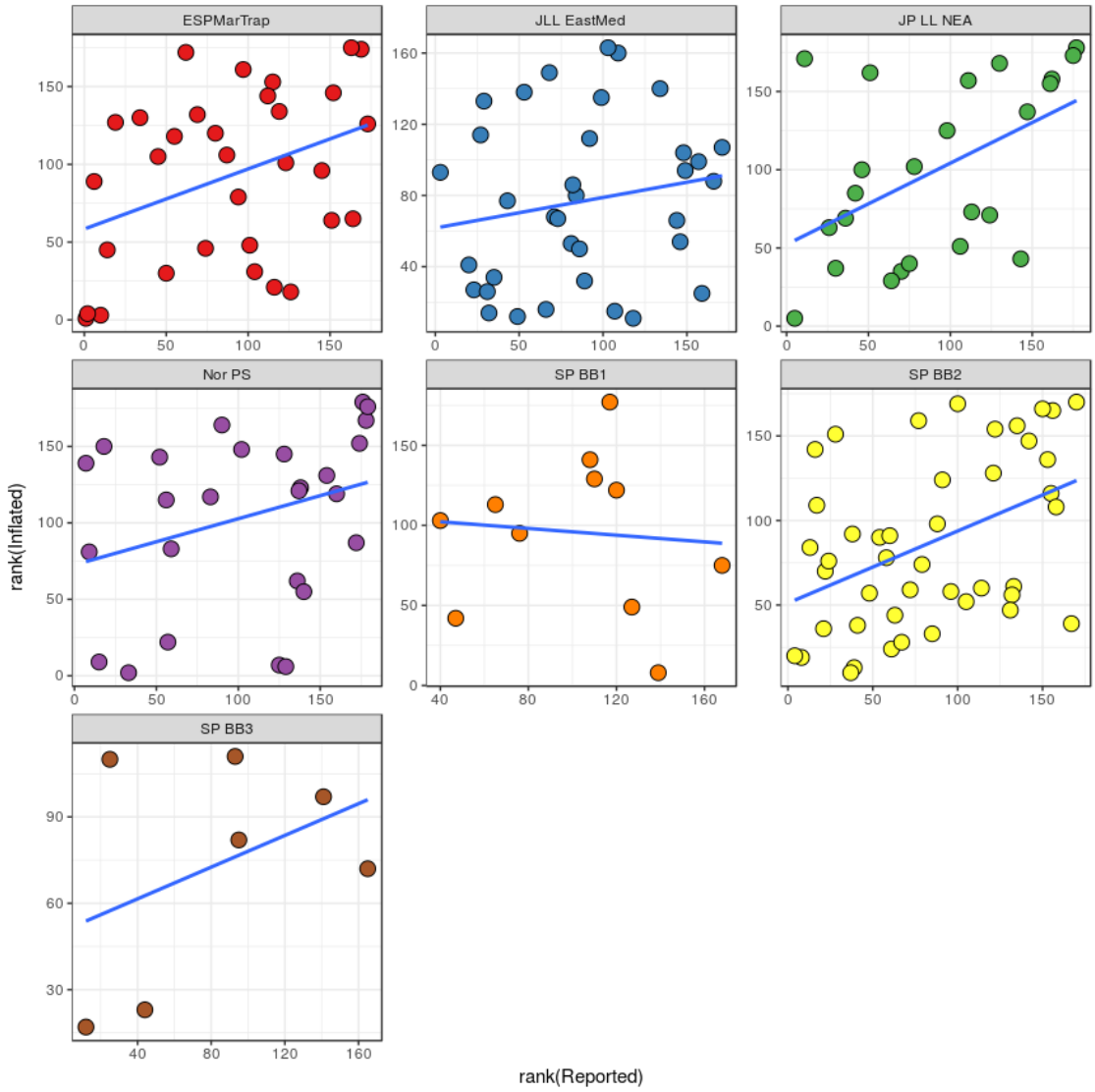

Figure 19. Comparison of jackknife estimate ranks of population growth rate B/B_{MSY} by year and CPUE series.

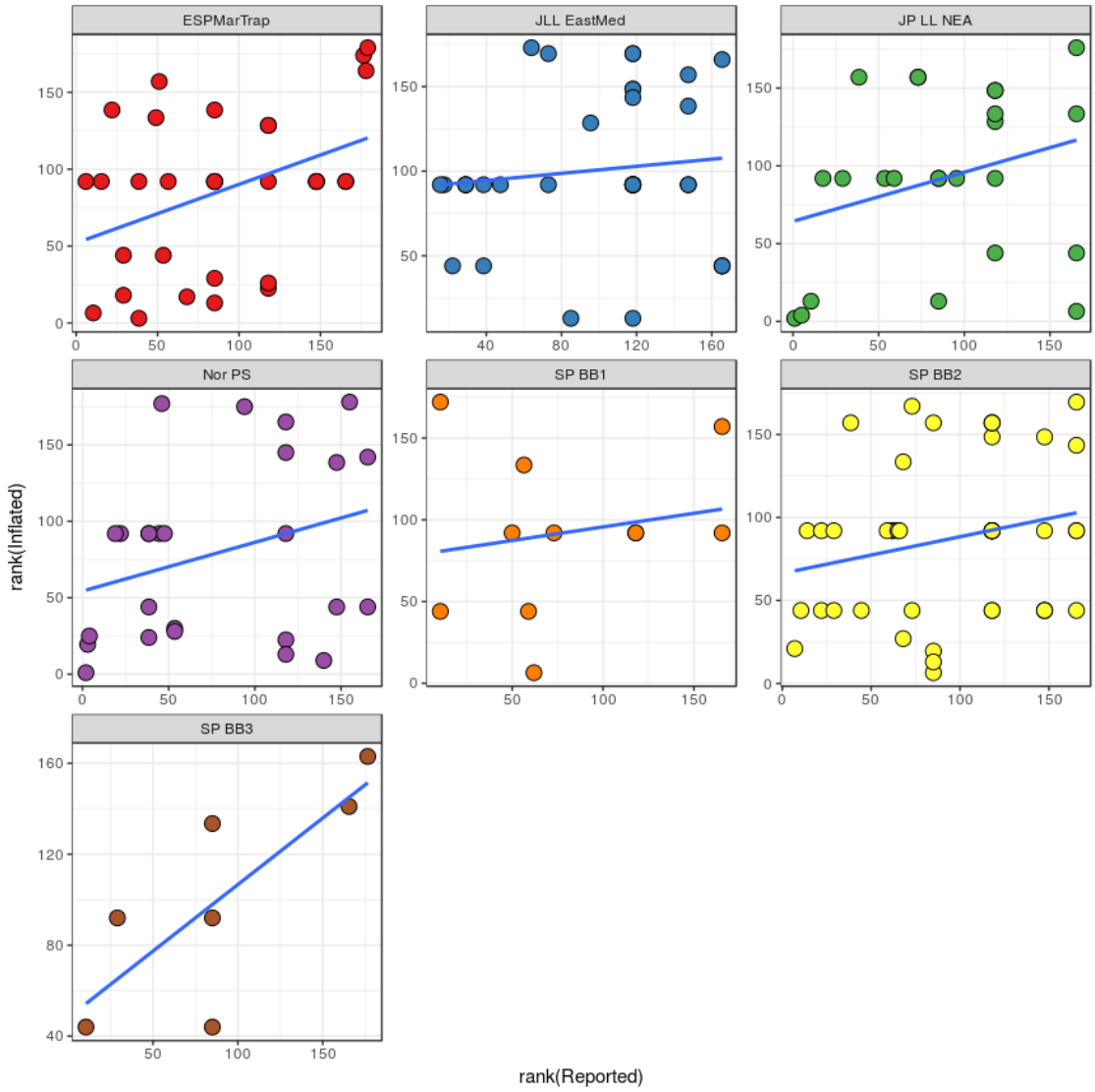

Figure 20. Comparison of jackknife estimate ranks of population growth rate F/F_{MSY} by CPUE series.

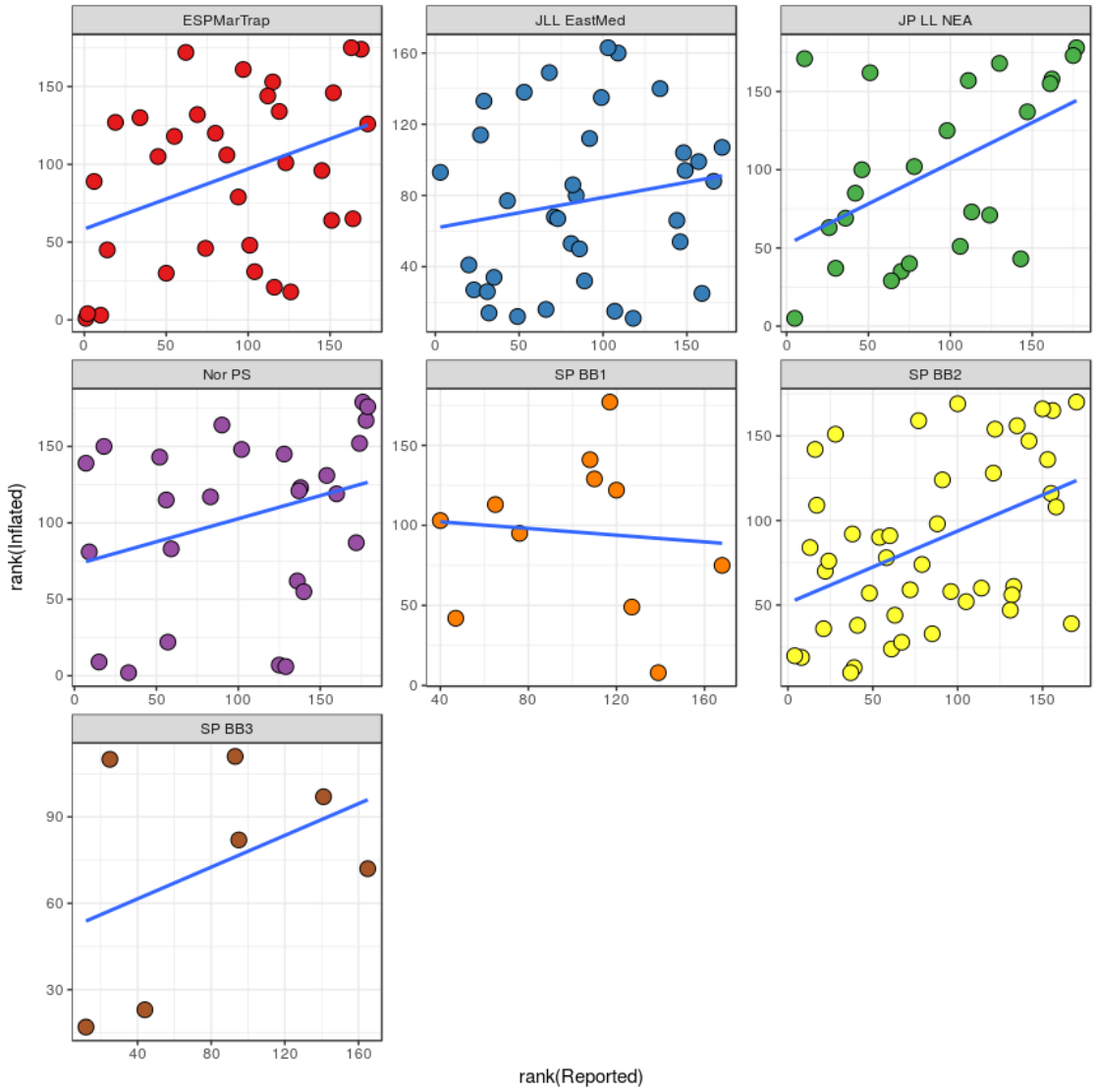

Figure 21. Comparison of jackknife estimate ranks of population growth rate *r* by CPUE series.

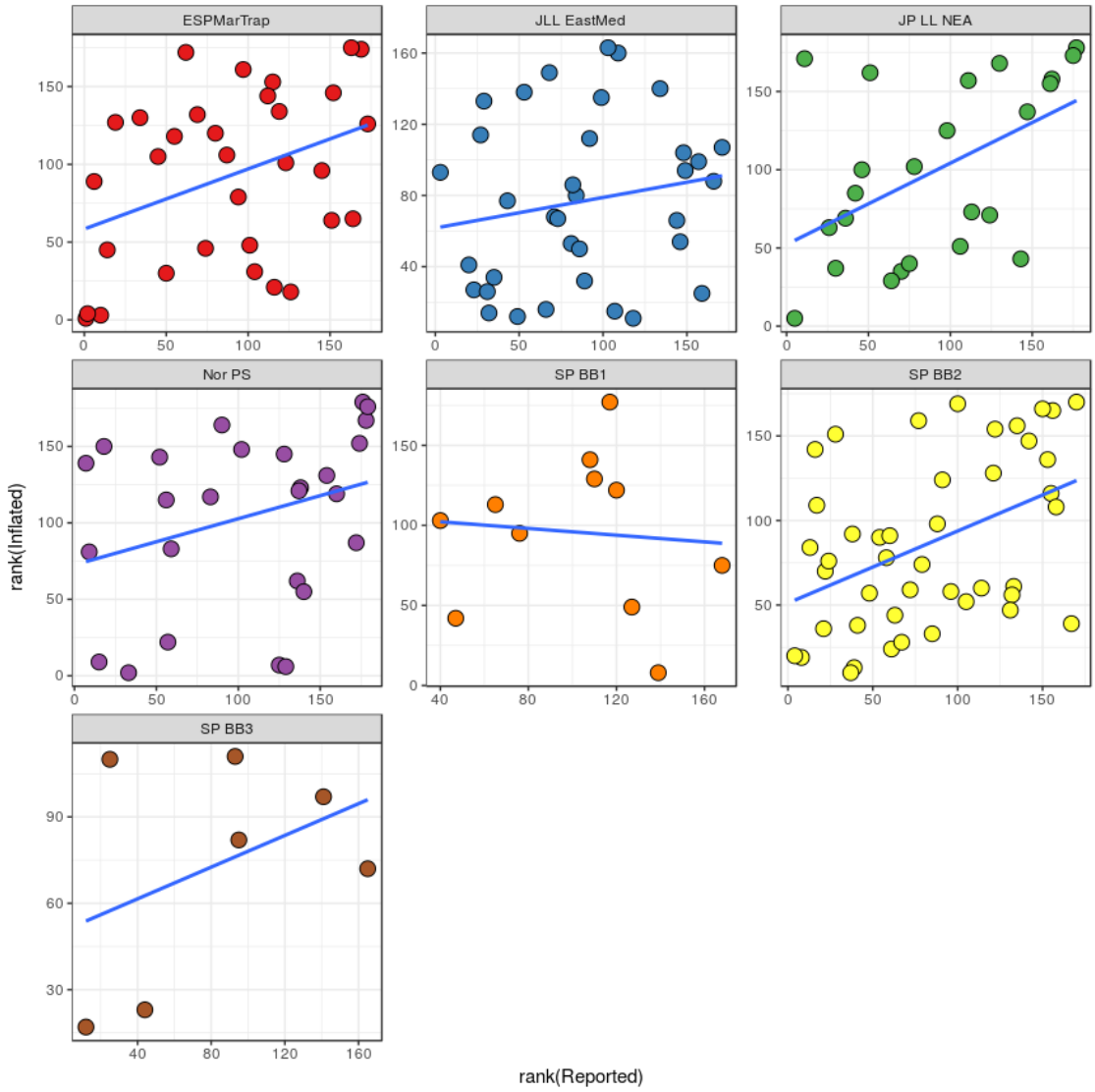

Figure 22. Comparison of jackknife estimate ranks of time taken to recover to B/B_{MSY} (negative values indicate stock is above B_{MSY}) by CPUE series.

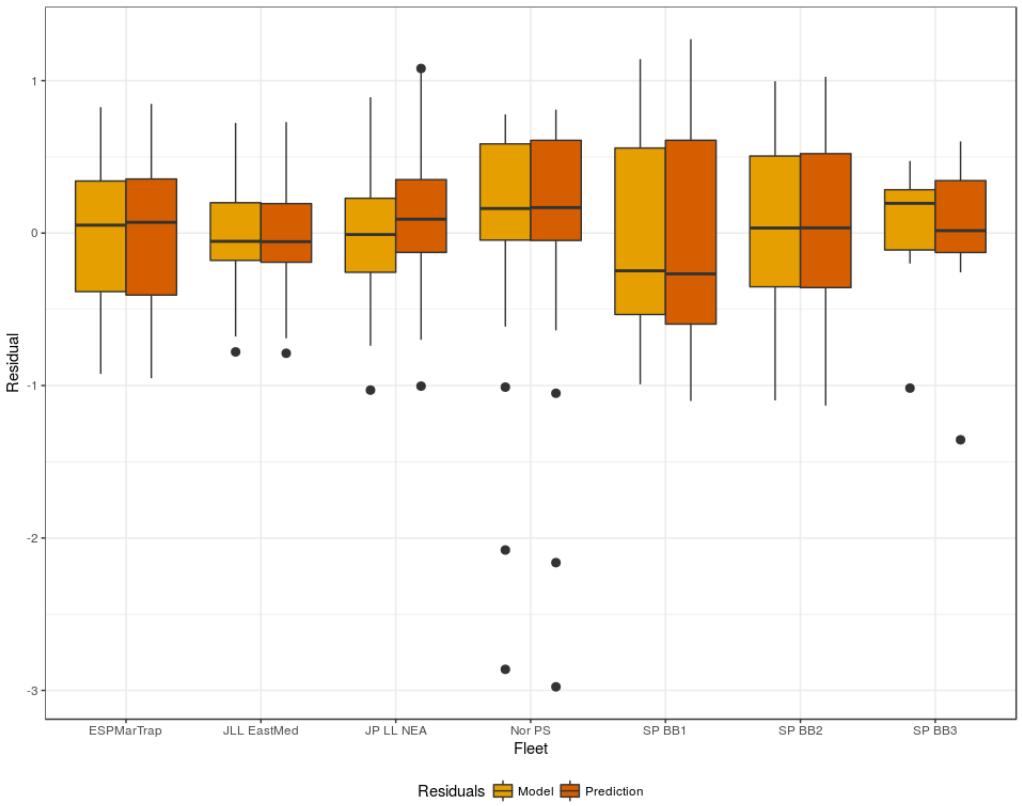

Figure 23. Prediction residuals by CPUE series for the reported catch Scenario.

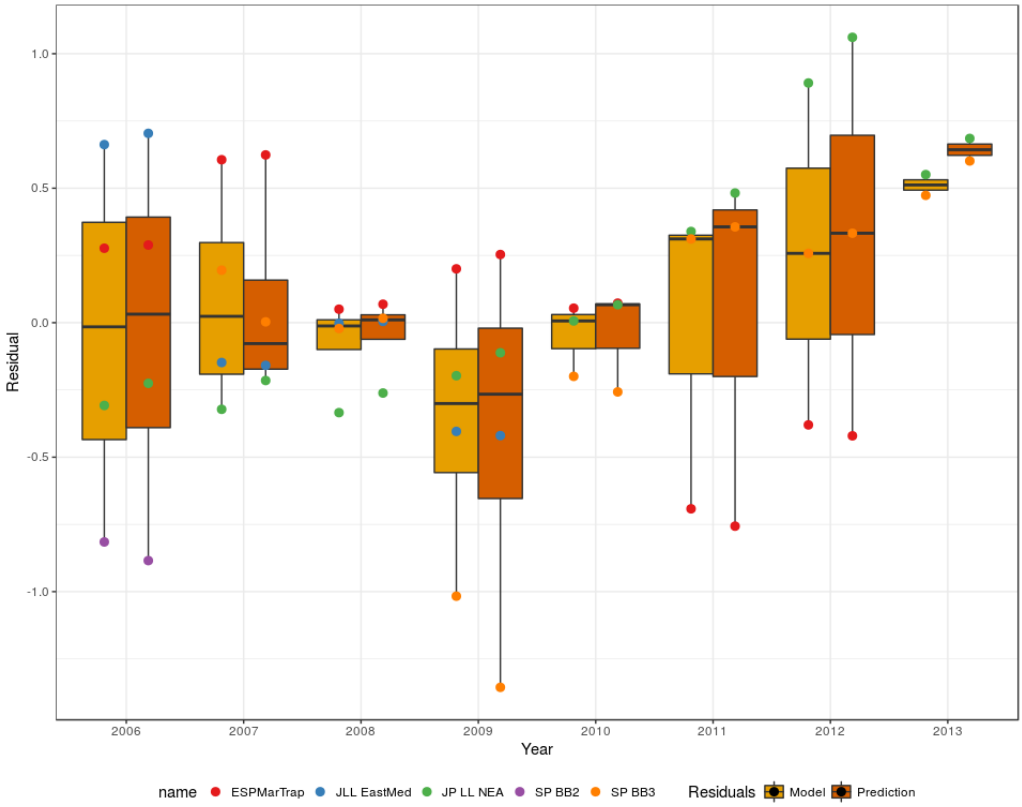

Figure 24. Prediction residuals by year for the reported catch Scenario.

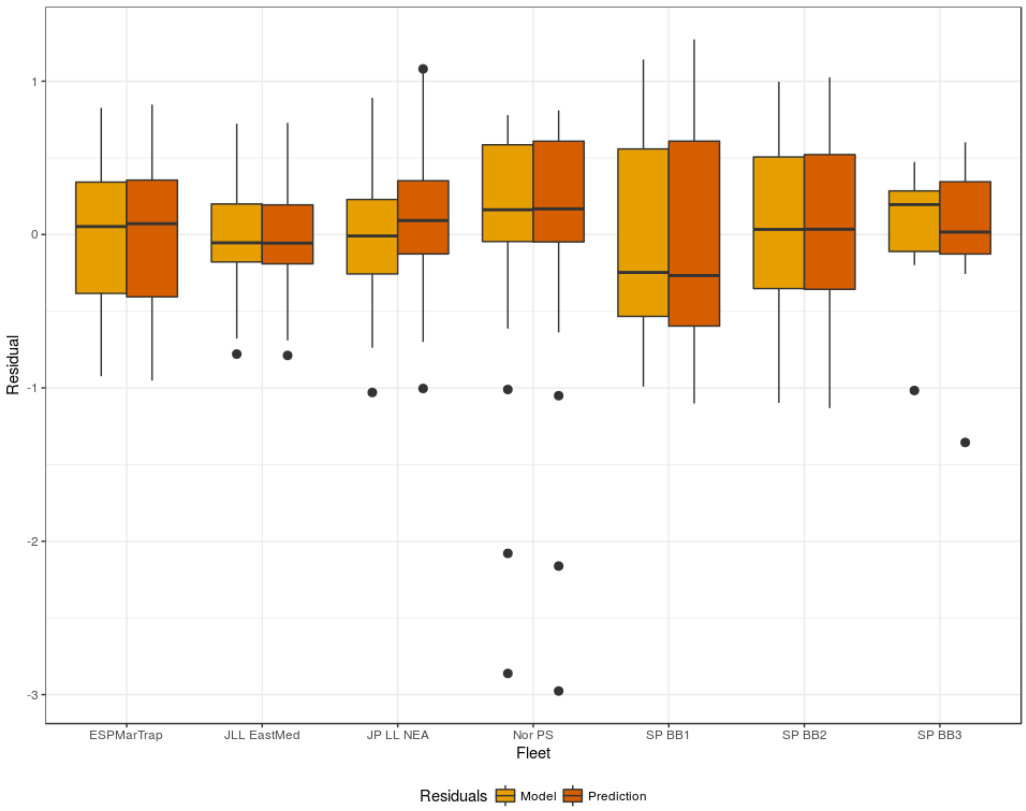

Figure 25. Prediction residuals by CPUE series for the inflated catch Scenario.

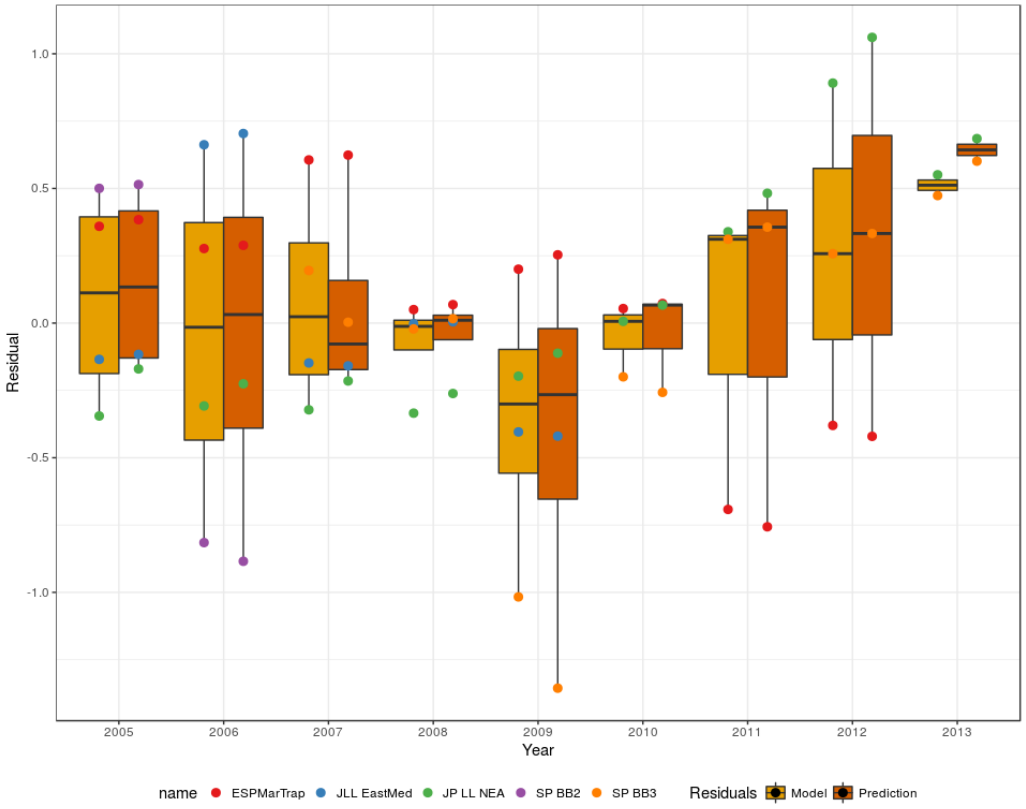

Figure 26. Prediction residuals by year for the inflated catch Scenario.