Evaluation of the method proposed by Western Power to calculate service standard benchmarks and targets for the fourth access arrangement period

Report to the ERA

14th September 2018

Client:

Economic Regulatory Authority Western Australia

Project:

0027_ERA_2018

Consultant:

Rohan Sadler

Pink Lake analytics

mobile: 0433 192 600
email: rohan.sadler@pinklake.com.au
ACN: 611 093 120
ABN: 60 611 093 120

Table of Contents

Executive Summary

The Economic Regulation Authority (ERA) is currently reviewing proposed changes to the methodology for determining Service Standard Benchmarks (SSBs) for Western Power's fourth access arrangement (AA4). In making a decision the Economic Regulation Authority must determine whether a proposed access arrangement meets the Code objective and the specific requirements set out in Chapter 5 of the Code.

The Proponent (Western Power) proposes two key methodological changes. The first change is to lift the SSB quantile from the 97.5th quantile, as it was in the previous access arrangement, to the 99th percentile. Their key argument for raising the SSB is to counter the multiple trial problem where multiple service provision indicators are applied in determining a penalty if service provision failures exceed the respective SSBs of the indicators. In statistical terms, the multiple trial problem increases the Type I error rate of a false positive declaration of a service breach, i.e., by simple randomness the yearly service provision indicator exceeds the SSB without there being any actual decline in service provision. For a 97.5th percentile the Type I error rate is nominally 0.025. However, if five independent service provision indicators are considered, and a service breach declared as soon as at least one of the indicators exceed their respective SSBs, then the compounded Type I error rate will be 0.112. For a 99th percentile with a nominal 0.01 Type I error rate the compounded Type I error rate associated with five independent service provision indicators is 0.0498. The effect of the multiple trial problem compounds with more trials (i.e., standard service provision).

The second change is a move away from a single best model fitted by the AIC criterion towards model averaging. In theory, model averaging is seen to reduce model selection bias which can have significant influence on the estimated value of a parameter such as an SSB quantile. Moreover, reducing model selection bias can reduce uncertainty around parameter estimates. The Proponent proposes a committee method of weighting derived from the Burnham and Anderson information-theoretic approach (BAITA).

To evaluate these different assertions then a two-stage Monte Carlo simulation was applied that assumed a mixture model estimated from daily customer interruption data as a hypothetical 'true' distribution. From this model a sampling distribution of yearly SAIFI and SAIDI values was constructed. This sampling distribution could then be compared to sampling distributions of the SSBs estimated from the simulated data for the SAIFI and SAIDI service performance indicators. The approach is flexible in that it allows consideration of different levels of data aggregation (yearly data, monthly 12-month rolling averages, and daily data), different SSB quantiles (97.5th and 99th) and different options for model averaging. A non-parametric method of SSB quantile estimation was also considered as this important class of models was omitted from the basket of parametric distribution models included in the model averaging.

The Monte Carlo study addresses a key weakness in the Proponent's proposal to date, whereby claims of best practice and superiority of method are not well supported by measures of statistical performance of those methods applied to the data at hand. We introduce several statistical performance measures (including standard error, bias, and prediction error) that address in part the requirement for evaluating the accuracy, replicability, consistency, and robustness of the SSB quantile estimates of the different estimation procedures.

The key findings are that:

- The Proponent's proposed model averaging does not significantly improve upon the single best model or standard BAITA model averaging. The standard BAITA outperforms (marginally) the Proponent's method in half of the data scenarios considered.
  - There is high correlation among component parametric models (e.g., the three parameter Weibull model is a generalization of the two parameter Weibull model).
  - The data have been aggregated as a monthly 12 month rolling average. It may be speculated that under the central limit theorem then regular distributions fit reasonably well to the data, and so little improvement in performance is observed when model averaging is applied.
- Modelling the daily data returns produced better performing estimates of the SSB quantiles than the monthly 12-month rolling average or yearly data, in general.
- Non-parametric kernel density estimation (KDE) applied to the daily data outperformed the BAITA-based model averaging options. Other non-parametric estimators are available that have not yet been explored so further improvements in SSB quantile estimation can, in principle, be achieved.
- The proposal to raise the quantile (from 97.5^{th} to 99^{th}) on which the SSB estimate is based is fraught.
  - Estimates of 99^{th} quantiles have much higher associated uncertainty than estimates of 97.5^{th} quantiles, introducing a greater element of risk in determining service standard breaches.
  - The Proponent highlights only the upper bound of the Type I error rate of multiple trials. This error rate will likely be confounded by correlation among the different service provision indicators. Actual Type I error rates are likely to be lower due to correlation between indicators. Moreover, only the nominal Type I error rate is highlighted, while the actual estimated Type I error rate of each individual SSB is ignored.
  - The Proponent's argument completely ignores Type II errors (i.e., false negatives), where a decline in service provision is not detected by the SSB. An effort to reduce Type I errors, that incur a cost to the Proponent, will increase Type II errors that incur a cost to the consumer community. Importantly, Type II error rates are more sensitive to changes in the SSB quantile than Type I error rates and can increase dramatically with the choice of quantile.
  - Ideally, the costs of Type I errors and Type II errors should be known. If these costs were known then an SSB quantile minimizing inappropriate monetary transfers between market actors could potentially be identified that accounts for the multiple trial problem. However, determining such costs and deciding upon a model of Type I and Type II error required for such an approach would likely be contentious.

It is recommended that the proposed model averaging approach be rejected as in practice it does not achieve the benefits over the single best fit model that model averaging should deliver in theory for the service provision indicators tested here.

Similarly, it is recommended that the proposal to raise the quantile by which an SSB is defined be rejected, as costs associated with Type II errors have been ignored, and Type I errors not correctly accounted for.

It is recommended also that the statistical performance of proposed methodological changes be clearly demonstrated in future proposals so that superiority of one method over another be readily reviewed.

Terms of Reference

1. Pink Lake was invited by the Economic Regulation Authority to review Western Power's methodology for determining Service Standard Benchmarks (SSBs) that has been proposed for Western Power's fourth access arrangement (AA4).
2. The terms of reference were to provide advice and supporting analysis on the multi-model averaging method proposed by Western Power to derive service standard benchmarks for the fourth access arrangement period. The assignment will be conducted in two stages:
    1. Addressing the multi-model averaging approach, assess whether the service standard benchmark derived at the nominated quantile proposed by Western Power:
        i. is more accurate than that derived by using a single probability distribution of best fit
        ii. may be objectively replicated
        iii. is more consistent over time than the single probability distribution of best fit
        iv. is more statistically robust than the single probability distribution of best fit
        v. is biased or open to manipulation to the detriment of customers
        vi. is applicable to Western Power's performance data?
3. Comment and analysis is also sought on:
    1. whether the construction of the data set (60 * 12-month rolling averages) results in biased quantile estimates
    2. the relative stability of the 99th quantile estimate against an alternative, such as the 97.5th quantile.

In forming an opinion, the following documents were principally referred to:

- Western Power, Fitting Distributions for AA4 Service Standard KPIs – Setting the Service Standard benchmark (SSB) and Service Standard Target (SST), Attachment 6.2 – Access Arrangement Information, 2nd October 2017.
- Analytics + Data Science, Review of service standards methodology, A report prepared for Western Power as 'Attachment 6.1 – Access Arrangement Information', 18th September 2017.
- Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018.

Declaration

1. This report has been prepared by Rohan Sadler of Pink Lake Analytics Pty Ltd.
2. As the author of this report I have read, understood and complied with the Expert Witness Guidelines entitled Expert Witnesses in Proceedings in the Federal Court of Australia (as defined in the Federal Court of Australia's Practice Note CM 7; attached as Appendix B). As the author I have made all the inquiries that I believe are desirable and appropriate and that no matters of significance that I regard as relevant have, to my knowledge, been withheld from this report.
3. A curriculum vitae for the consultant has been provided as Appendix C.

Introduction

4. The Economic Regulation Authority (ERA) is currently reviewing proposed changes to the methodology for determining Service Standard Benchmarks (SSBs) for Western Power's fourth access arrangement (AA4).

5. In making a decision the Economic Regulation Authority must determine whether a proposed access arrangement meets the Code objective and the specific requirements set out in Chapter 5 of the Access Code. Section 5.6 of the Access Code requires service standard benchmarks to be:

    a. reasonable; and

    b. sufficiently detailed and complete to enable a user or applicant to determine the value represented by the reference service at the reference tariff.

Methodology proposed by Western Power

6. The methodology proposed by Western Power for setting service standard benchmarks is described by both Western Power[1] and Analytics + Data Science (A+DA)[2]. In summary, the methodology:

    a. Selects five years of monthly data computed as a 12 month rolling average, thereby providing 60 data points, for each performance measure (Table 1).
    b. Fits several candidate statistical (theoretical) distributions to a performance measure's data (Table 2).
    c. Manually examines each fitted distribution using quantile-quantile (Q-Q) and quantile-quantile (P-P) plots;
    d. Determined the theoretical distributions' goodness-of-fit using the Anderson-Darling test.[3]
    e. A distributional model is discarded from further evaluation if it is rejected at a 5% significance level of the Anderson-Darling test.
    f. The AIC (Akaike Information Criterion) is calculated for each fitted distributional model.
    g. Those distributional models that are within 1% of the AIC of the best fitting distribution (i.e., lowest AIC) are selected to form a model average and are assigned equal weight.
    h. The SSB of the performance measure is then the equally weighted average applied to the 99th quantiles of each of the selected distributions.

7. The proposed approach differs from that applied in the third access arrangement period in which the service standard benchmark was determined at the 97.5th quantile of the single distribution of best fit.

8. The SSBs for distribution reference services are expressed in terms of System Average Interruption Duration Index (SAIDI), System Average Interruption Frequency Index (SAIFI) and call centre performance.

[1] Western Power, *Fitting Distributions for AA4 Service Standard KPIs – Setting the Service Standard benchmark (SSB) and Service Standard Target (SST)*, Attachment 6.2 – Access Arrangement Information, 2nd October 2017.
[2] Analytics + Data Science, *Review of service standards methodology*, A report prepared for Western Power as 'Attachment 6.1 – Access Arrangement Information', 18th September 2017.
[3] Anderson, T.W. and Darling, D.A. (1954). "A Test of Goodness-of-Fit". *Journal of the American Statistical Association*. Vol. 49: Pages 765–769

a. SAIDI is the sum of the duration of each sustained (greater than 1 minute) distribution customer interruption (in minutes) attributable to the distribution system (after exclusions) divided by the number of distribution customers served, over a 12 month period.

b. SAIFI is the number of sustained (greater than 1 minute) distribution customer interruptions (number) attributable to the distribution system (after exclusions) divided by the number of distribution customers served, over a 12 month period.

c. Call centre performance measures interruptions and life threatening emergencies over a 12 month period as the percentage of calls responded to in 30 seconds or less (after exclusions).

9. SSBs are determined also for transmission services (system minutes interrupted, loss of supply event frequency and average outage duration) and street lighting services (street lighting repair time).

Burnham and Anderson Information-Theoretic Approach (BAITA)

10. A+DA refer to Burnham and Anderson's Information-Theoretic Approach (BAITA)[4] for modelling the SSBs in their latest report. Specifically, they state both:[5]

"The specification of one particular model as the "best" model for determining SSB/SST quantile values is inconsistent with the standard approach used by peer reviewed studies into statistical inference. In a practical manner, the "best" model is likely to vary from data set to data set, even if replicate data is captured from the same underlying process (Burnham & Anderson, 2002, p.151). The effect is not limited only to problems with small sample sizes. With data sets of even a moderate size, a slight change in the data may lead to the selection of a different model (Zou & Yang, 2004, p.70)."

and

"Burnham & Anderson (2004) provide a more complete discussion of the conclusions from multiple studies that demonstrate that a multimodel averaging approach is superior to the methodology in which parameter estimates are obtained from only the single "best" model."

11. In practice, the number of statistical modelling solutions for data are numerous. Moreover, with limited data several plausible models may be reasonably fitted. Hence, in many practical cases it is not possible to clearly identify a single most-appropriate model.

12. Multimodel averaging attempts to solve this dilemma by providing a mix of plausible models whose weighted average minimizes the averaged model prediction error (i.e., the sum of the bias and variance of predictions).

13. A key consideration is that BAITA is but one of several different model averaging techniques. Moreover, the method of weighting that the Proponent has employed is not strictly one of possible methods that BAITA have suggested. In effect, a committee method of model averaging is suggested

[4] Burnham, K.P. and D. R. Anderson, *Model Selection and Multimodel Inference: A practical information-theoretic approach*, 2002, Springer-Verlaag, New York, Second Edition, 515 pp.
Burnham, K.P. and D.R. Anderson. "Multimodel inference: Understanding AIC and BIC in model selection", *Sociological Methods Research*, 2004, pp. 261-304.

[5] Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 4.

by the Proponent, where models for the averaging are selected based on a 1% distance from the Akaike Information Criterion (AIC) score[6] of the best performing model (i.e., AIC minimizing model). The selected models are then assigned equal weight. Instead, the BAITA framework recommends that a cut-off for considered models are those with an AIC difference of less than 10 (based largely on likelihoodist principles)[7], where the number of model parameters and candidate models are small (as is the case with the Proponent's proposed methodology):[8]

"The sampling distribution of Δ_p was examined for many situations, and we found that generally, for a small number of candidate models, a value ≥ 10 corresponds to at least the 95th quantile and more often at least the 99th quantile. This supports our contention that an observed $\Delta_i \geq 10$ is strong evidence against model g_i."

where Δ_p is the difference in AIC scores between the theoretical best model and the best performing model (i.e., AIC minimizing model); and,

Δ_i is the difference in AIC scores between model g_i and the best performing model.

14. In contrast, the Proponent applies a 1% AIC cut-off. Given the minimum AIC score reported by the Proponent for their method ranges from 415.06 to 655.9 for SAIDI scores and -185.59 to 86.89 for SAIFI scores across the different feeder categories,[9] then a 1% AIC cut-off equates approximately to $\Delta_i \geq 4.2 - 6.6$ for SAIDI and $\Delta_i \geq 0.3 - 1.9$ for SAIFI. Overall, fewer component distributions are being included in the Proponent's model averaging method than what would nominally be recommended under the BAITA.

15. Significantly, BAITA propose the use of Akaike weightings of each of the selected models to perform the model averaging, that is, a weighted average of the selected models should be applied rather than an equal weighting:[10]

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)}$$

where w_i is an Akaike weight as a measure of the weight of evidence in favour of a model; and,

R is the total number of models applied in the model averaging.

[6] Akaike information criterion (AIC) is an information theory derived measure of the amount of 'information' in a model. Models with greater information, and hence a lower AIC score, will in theory predict future data better than models with less information. Bayesian information criterion (BIC) is determined similarly, but focuses more on model parsimony, and assumes that the 'true' model is considered within the set of fitted models.

[7] Royall, R., Tibshirani, R. (1997). *Statistical Evidence: A likelihood paradigm*, New York: Routledge, 1997

[8] Burnham, K.P. and D. R. Anderson, *Model Selection and Multimodel Inference: A practical information-theoretic approach*, 2002, Springer-Verlaag, New York, Second Edition, pp. 264-265.

[9] Western Power, *Fitting Distributions for AA4 Service Standard KPIs – Setting the Service Standard benchmark (SSB) and Service Standard Target (SST)*, Attachment 6.2 – Access Arrangement Information, 2nd October 2017, pp. 18-32.

[10] Burnham, K.P. and D.R. Anderson. "Multimodel inference: Understanding AIC and BIC in model selection", *Sociological Methods Research*, 2004, p. 272.

Note that bootstrapped values of the weighting could be derived, but the AIC derived weighting is sufficient in most instances given the convergence of the AIC to the model prediction error as sample size increases, while being computationally much simpler.

16. If the Akaike weight for one model out of all candidate models is 1 (or close to 1), and hence all other weights are zero (or close to zero), then a single best model is selected via BAITA. Hence, BAITA model averaging methods include the selection of a single best model as a special case.

17. Under the BAITA the unconditional standard error of parameter estimates may be estimated once the Akaike weights are computed:[11]

$$s.e.\left(\bar{\bar{\theta}}\right) = \widehat{var}\left(\bar{\bar{\theta}}\right)^{1/2}$$

$$\widehat{var}\left(\bar{\bar{\theta}}\right) = \left[\sum_{r=1}^{R} w_i \left[\widehat{var}(\bar{\theta}|g_i) + \left(\bar{\theta}_i - \bar{\bar{\theta}}\right)^2\right]^{1/2}\right]^2 \qquad \text{Eqn. 1}$$

where $\bar{\bar{\theta}} = \sum_{r=1}^{R} w_i \, \bar{\theta}_i$ is the Akaike weighted average of the parameter $\bar{\theta}_i$ estimated by each model g_i.

18. As the Proponent's model averaging method includes fewer component distributions in the average than what is nominally recommended under the BAITA, and an equal weighting is applied to those component distributions that are included in the average. It is therefore hypothesized that $\widehat{var}\left(\bar{\bar{\theta}}\right)$ will be greater than what is recommended under the BAITA. This is primarily due to proportionally more weight being assigned to the second-best or third-best component distribution of the proponent's method than under the BAITA. That is, more weight is assigned to poorer fitting distributions as measured by the AIC score. Moreover, more model terms are expected to be included under BAITA given the $\Delta_i \leq 10$ criterion, hence there is more hedging against out-of-sample 'surprises' than under the Proponent's method when it comes to model prediction.

Choice of 97.5th or 99th Quantile

19. The BAITA standard error estimate has relevance insofar as the performance of different parameters of interest may be compared (i.e., variability in the 97.5th and 99th quantile estimates of each model averaged distribution). Critically, there has been no reference made by the proponent to standard error estimates of their SSB estimates. Instead, A+DS state that:[12]

"We are also not aware of any statistical basis which would suggest the 99th quantile value to be any more or less appropriate than an alternative threshold. Consequently, we concur that the 99th quantile value an appropriate threshold that aligns with Western Power's longer term strategic objectives for AA4."

[11] Burnham, K.P. and D.R. Anderson. "Multimodel inference: Understanding AIC and BIC in model selection", *Sociological Methods Research*, 2004, p. 273.

[12] Analytics + Data Science, *Review of service standards methodology*, A report prepared for Western Power as 'Attachment 6.1 – Access Arrangement Information', 18th September 2017, p. 10.

20. Moreover A+DS state that:[13]

    "Given that objective, it is appropriate to choose a quantile value that does not penalise Western Power for not continuing to improve performance. Choosing a lower threshold value would increase the probability that, in the absence of further investment (at the expense of customers), service standards would not be met and Western Power would be financially penalised."

21. The Code states that:[14]

    "establishes a framework for third party access to electricity transmission and distribution networks with the objective of promoting the economically efficient investment in, and operation and use of, networks and services of networks in Western Australia in order to promote competition in markets upstream and downstream of the networks."

22. While it is true that the Proponent should not be unnecessarily penalized with any setting of the SSB, similarly it is true that the setting of the SSB should not transfer network costs unnecessarily from the service provider to consumers of the service.

23. In this context, a naïve benefit-cost analysis would seek to set the SSB at a level whereby the marginal cost to the Proponent of inappropriate penalties (due to too strict an SSB) equals the marginal cost to the community of consumers of appropriate penalties not being applied whenever the SSB fails to detect insufficient service provision (due to too lax an SSB).

24. Hence, setting a high quantile for the SSB that effectively minimizes the rate of penalization, as A+DS appear to have recommended, should be viewed with some skepticism, as it does not necessarily balance the requirements of the consumer community for a level playing field.

25. Complicating the issue, the data are limited, and hence any quantile estimate applied as the SSB will be uncertain. High uncertainty in the SSB quantile estimate will deliver less trust in the SSB than if the SSB quantile can be estimated with greater certainty.

26. Asymptotically, the variance of a quantile estimate is inversely proportional to the square of the probability density function evaluated at the quantile. For normally distributed data this means the variance of the 99th quantile is approximately 80% greater than the variance of the 97.5th quantile (i.e., 36.9% greater in standard error terms).[15]

---

[13] Analytics + Data Science, *Review of service standards methodology*, A report prepared for Western Power as 'Attachment 6.1 – Access Arrangement Information', 18th September 2017, p. 10.

[14] Electricity Industry Act 2004, 30th November 2004.

[15] The asymptotic variance of a quantile after Gross, A.M. and V. Clark, *Survival Distributions: Reliability Applications in the Biomedical Sciences*, John Wiley, New York, 1975, is given by:

$$\sigma_P^2 = \frac{P(1-P)}{f(x_P)^2 N}$$

where σ_P^2 is the asymptotic variance for the Pth quantile, $f(x_P)$ is the probability density function (pdf) evaluated at the Pth quantile of the distribution x_P, and N is the sample size.

In practice the asymptotic variance formula should not be used except for very large sample sizes due to instability in the variance estimate (Brown M.B. and R.A. Wolfe, "Estimation of the variance of quantile estimates", *Computational Statistics & Data Analysis*, 1983, pp. 167-174). However, this calculation is sufficient for demonstration of the concept of higher variances being associated with more extreme quantiles. It follows that for a normal distribution $\sigma_{97.5}^2 \approx 0.024375/(0.145^2 \times 60) = 1.159$ and $\sigma_{99}^2 \approx 0.0099/(0.0675^2 \times 60) = 2.173$, giving $\sigma_{99}^2/\sigma_{97.5}^2 = 1.875$

27. It follows that under standard distributional assumptions of log concavity, of which the normal distribution is but one case, the uncertainty associated with estimates of extreme quantile values is greater than that associated with less extreme quantile values. As demonstrated for the normal distribution, this increase in uncertainty with increasing quantile can be significant even when the increase in quantile is small, whenever these quantiles are located at the extremes of a distribution.

28. In effect, the choice of quantile fundamentally affects the rates of Type I error (i.e., false positive declarations of a breach of SSBs) and Type II error (i.e., false negative declarations where no breach is declared although in truth the service provision may have declined in standard).

29. Hence this study seeks to apply a set of statistical performance measures that measure the uncertainty and reliability of the quantile estimates, as well as assess in a preliminary fashion Type I and Type II error rates. A simulation experiment will provide a case study generated from the available data where the statistical performance measures are quantified.

Reliability assessment of the proposed SSB estimation method

30. The terms of reference require that an SSB be reliably estimated. A statistical performance measure may be assigned to each requirement of the Stage 1 terms of reference (Table 1). Overall, an estimator of the SSB that is accurate, reproducible and consistent may be considered as reliable. Hence competing methods for SSB estimation may be evaluated.

31. A Monte Carlo experiment[16] can be designed to evaluate the SSB estimation methods by the defined statistical performance measures (Table 1). Data were provided by Western Power for the SAIDI and SAIFI scores, without disaggregating the data by feeder category. A three-component mixture model was fitted to the data for each SAIDI and SAIFI score. This mixture model may be viewed as a hypothetical true distribution of a SAIDI or SAIFI score for the purposes of the simulation.

32. From the mixture model 10,000 samples of five years' worth of data were randomly drawn. These data then had outliers removed through the beta threshold method applied by the Proponent.[17] A sampling distribution for the SSB and the statistical performance measures (Table 1) may then be constructed by applying the SSB calculation and the statistical performance measures to the Monte Carlo samples.

33. Note that the data provided were aggregated at the daily scale; hence the beta threshold method was applied to the daily data and not to the individual event data as the Proponent applies. The data provided at the daily scale covered only CMI and CI, and hence only SAIDI and SAIFI scores were calculated.

[16] Efron, B. and R. J. Tibshirani. *An Introduction to the Bootstrap*, CRC Press, 1994.

[17] IEEE guide for electric power distribution reliability indices. New York: Institute of Electrical and Electronics Engineers, 2012. P27. An optimal box-cox transform is applied to the data to identify those data points beyond the beta threshold that is applied on the Box-Cox scale. These extreme data points are considered as major event days and are excluded from the SSB calculation: Western Power, *Service Standard Performance Report for the year ended 30 June 2016*, September 2016, p. 25. Note that a Box-Cox transformation is invariably preferred to other transformations of the data, as it provides a maximum likelihood estimate of a transformation parameter that produces an optimally normal distribution.

34. Transmission service indicators such as LOSEF were not examined. In this instance, LOSEF data were not provided at a sufficiently granular level (i.e., at the individual event level such that the LOSEF indicators may be reconstructed from first principles) for an assessment of data aggregation method on the SSB estimates for these indicators.

35. For the purposes of simulation, the number of customers was kept constant when calculating SAIDI and SAIFI SSB estimates.

36. SSB estimation via different methods may then be applied to the simulated data. In addition to the best fitting single distribution model and the Proponent's model averaging model, the following models were included in the comparison:

    a. The method of setting the standard service target by fitting the single best fitting model to the past five yearly estimates of a service provision indicator. A standard deviation is estimated from this model fitting, and an upper quantile is derived as the SSB.[18] This approach can be generalized as a model of single best fit.

    b. Model averaging applying Akaike weights under BAITA.[19]

    c. A 'naïve' kernel density estimate of the SSB quantile that reflects the kernel density estimate at 0 (lower bound of data).[20]

    d. For speculative reasons, the Proponent's method with and without the Anderson-Darling test applied was also compared.

37. The Stage 2 requirements within the terms of reference seek an investigation of both the influence of data construction on the SSB estimates and the choice of 97.5th and 99th quantile. Hence, these methods were applied to both data aggregated as monthly 12-month rolling averages and as daily data. The methods excepting the kernel density estimate were also applied to the yearly data. A fourth option of modelling discrete service failures within each day was not considered here. Sampling distributions for service provision indicators derived from the aggregated monthly 12 month rolling averages and yearly data could readily be constructed from daily simulations to enable comparisons between the different levels of data aggregation.

38. The 'naïve' kernel density estimate (KDE) method proposed here should not be viewed as a favoured method of SSB estimation. Instead, it is considered as a representative of a broader class of non-parametric estimation methods that attempt to 'smooth' the empirical distribution. Non-parametric methods are a valid competitor when estimating quantiles to any parametric method such as the proposed model averaging method. Alternate KDE methods are available, some better suited to extreme value estimation than others.[21] Hence, the inclusion of a non-parametric method guards

[18] For example: Parsons Brickenhoff, *Fitting probability distribution curves to reliability data*, a report to TransGrid, 31st March 2014,

[19] That is, the standard BAITA, as describe in paragraphs 10-18 above.

[20] Bowman, A.W. and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford, 1997.

[21] For a more sophisticated set of kernel density estimators for extreme value distributions then see: Hu, Y. and C. Scarrot, "evmix: An R package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density Estimation", *Journal of Statistical Software,* 84, 2018, pp. 1-28.

against, to a limited extent, the possibility of model misspecification when only parametric models are considered in the model mix.

39. Table 1 outlines the statistical performance measures to be applied to simulations of the different model estimation strategies. These metrics are defined in Table 2 below.

40. For each simulation $n \in \{1,2,\dots,N\}$ then random five yearly data sets are generated from of the mixed s model. Each component distribution for SAIDI and SAIFI is then estimated (from the basket of 12 continuously-valued estimators proposed by the Proponent). To acquire relatively precise estimates of the upper tail quantiles of each component distribution then, for consistency of method, M yearly simulations were independently generated for each distribution. These quantiles of the component distributions were then model averaged using the Akaike weights generated from each simulation n. This method therefore produced N estimates of $\bar{\theta}$ termed here as $\bar{\theta}_{r,n}^{M}$ for each estimation method r given the M second-stage simulations (the Proponent's model averaging with and without the Anderson-Darling test, BAITA, single best AIC model and the kernel density method trialed here).

41. To construct the various statistical performance metrics then M simulations of the hypothetically 'true' mixed s model (indexed as model 0) were generated to provide relatively precise estimates of the 'true' SSB quantiles, termed here as $\hat{\theta}_0^M$.

42. This Monte Carlo (or parametric bootstrap) approach[22] to providing the statistical performance measures may be improved through bias-corrected and acceleration to improve the accuracy of estimation of the statistical performance measures.[23] However, as an estimated mixed s model is assumed to be the hypothetical 'true' distribution, then the uncorrected statistical performance measures are sufficient for method comparison and demonstration of any counterfactual cases that might highlight flaws in the Proponent's proposed methodology.

43. The prediction error was measured as a root mean square error (RMSE) estimate of the estimated SSB quantile. This Monte Carlo estimate of the RMSE is equivalent to the bias corrected estimate of the model averaged $\widehat{var}\left(\bar{\bar{\theta}}\right)$ given in Eqn. 1. Similarly, the RMSE can be decomposed into estimated variance and bias measures of the quantile estimate $\bar{\theta}$.

44. The standard error of a quantile $SE(\bar{\theta}_r^M)$ was calculated with respect to the mean quantile estimate taken across all simulations $\bar{\bar{\theta}}_r^M = \sum_{n=1}^{N} \bar{\theta}_{r,n}^M / N$. As further measures of the uncertainty in the quantile estimate $\bar{\theta}_r^M$ then the 95% confidence band is defined across the N estimates (giving the 2.5% lower and 97.5% upper bounds), as well as the median.

MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M. and G. Russell, "A flexible extreme value mixture model.", *Computational Statistics and Data Analysis*, 55(6), 2011, 2137-2157.

[22] Efron, B. and R. J. Tibshirani. *An Introduction to the Bootstrap*, CRC press, 1994.

[23] Efron, B. and R. J. Tibshirani. *An Introduction to the Bootstrap*, CRC press, 1994.

Table 1. Statistical measures corresponding to each requirement in the terms of reference.

Requirement	Statistical Performance Measure	Reasoning
Assess whether the proposed SSB is more accurate than that derived by using a single probability distribution of best fit.	Prediction error	Prediction error is a sum of model bias and variance in predictions. It measures the out-of-sample accuracy of the model given uncertainty in model parameter estimates.
Assess whether the proposed SSB may be objectively replicated.	Standard error of the SSB	Both replication (i.e., a similar sample of data will be drawn from the population under a given set of conditions, leading to a similar SSB estimate) and reproducibility (i.e., the instructions for producing the SSB estimate are clear, and by following the instructions the same SSB estimate is returned given a sample of data) may be considered. The standard error of the SSB estimate provides information on the uncertainty we have in our SSB estimate (even if biased) from sample to sample (i.e., a weak measure of replicability). Error free code at run-time indicates reproducibility.
Assess whether the proposed SSB is more consistent over time than the single probability distribution of best fit.	Pitman Closeness	The Pitman closeness measure indicates that an estimator is to be preferred to a comparison estimator if the estimated probability of estimator being closer to the true value is greater than 0.5.[24] However, the measure does not exhibit appropriate transivity[25] and it is recommended for it be used cautiously. We apply Pitman closeness here as a simple to compute surrogate for relative efficiency.[26]
Assess whether the proposed SSB is statistically robust to single changes in the underlying data.	Local-shift sensitivity	The local-shift sensitivity measures the effect of perturbing a data point by differing amount.[27] A lower sensitivity is preferred to a higher sensitivity.
Assess whether the proposed SSB is biased or open to manipulation to the detriment of customers	Bias	Bias is a key measure of accuracy in the statistical literature.
Assess whether the proposed SSB is applicable to Western Power's performance data.	-	A normative assessment based on the above measures.

[24] Pitman, E. (1937). The "closest" estimates of statistical parameters. *Mathematical Proceedings of the Cambridge Philosophical Society, 33*(2), 212-222.

[25] Robert, Christian P., et al. "Is Pitman Closeness a Reasonable Criterion?" *Journal of the American Statistical Association*, vol. 88, no. 421, 1993, pp. 57–63

[26] An estimator of the SSB will be relatively efficient compared to another estimator of the SSB if for all true values of the SSB the expected error in estimation of this SSB is lower. A relatively efficient estimator will require a smaller sample size to achieve a specified prediction error.

[27] Hampel, F.R., Ronchetti, E,M., Rousseeuw, P.J. and W.A. Stahel, *Robust statistics,* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, New York: John Wiley & Sons, Inc., 1986.

45. Bias is simply the difference between the hypothesized 'true' mean $\hat{\theta}_0^M$ and the mean quantile estimate $\bar{\bar{\theta}}_r^M$. To compare with the estimated variance of the quantile estimates $\widehat{var}(\bar{\theta}_r^M) = SE(\bar{\theta}_r^M)^2$ then the squared bias is reported.

46. Pitman closeness compares two estimators for the same parameter. Here, each estimator was compared with the single best AIC model estimate of each quantile when the data were aggregated monthly as a yearly rolling average, termed here as $\bar{\theta}_{\boldsymbol{r}^*,n}^M$. Note that function $1(\cdot)$ returns a value of one when the argument of the function inside the brackets is true, and zero otherwise.

Table 2. Derivation of the different statistical performance measures

Statistical Performance Measure	Formula
Prediction Error	$RMSE(\bar{\theta}_r^M) = \sqrt{\frac{\sum_{n=1}^{N}\left(\bar{\theta}_{r,n}^M - \hat{\theta}_0^M\right)^2}{N}} = \sqrt{SE(\bar{\theta}_r^M)^2 + \widehat{Bias}(\bar{\theta}_r^M)^2}$
Standard Error	$SE(\bar{\theta}_r^M) = \sqrt{\frac{\sum_{n=1}^{N}\left(\bar{\theta}_{r,n}^M - \bar{\bar{\theta}}_r^M\right)^2}{N}}$
Bias squared	$\widehat{Bias}(\bar{\theta}_r^M) = \left(\bar{\bar{\theta}}_r^M - \hat{\theta}_0^M\right)^2$
Pitman closeness	$PC(\bar{\theta}_r^M) = \sum_{n=1}^{N} \frac{1\left(\bar{\theta}_{r,n}^M - \hat{\theta}_0^M < \bar{\theta}_{\boldsymbol{r}^*,n}^M - \hat{\theta}_0^M\right)}{N}$
Local-shift sensitivity	$LSS(\bar{\theta}_r) = \max_x \frac{\left\lvert\bar{\theta}_r(x) - \bar{\theta}_r(\bar{x})\right\rvert}{\lvert x - \bar{x}\rvert}$
Type I error	$P\left(\bar{\theta}_r^M < \tilde{Y}_{0,n}\right) = \sum_{n=1}^{N} \frac{1\left(\bar{\theta}_{r,n}^M < \tilde{Y}_{0,n}\right)}{N}$
Type II error with 20% effect size	$P\left(\bar{\theta}_r^M > \tilde{Y}_{0\times 20\%,n}\right) = \sum_{n=1}^{N} \frac{1\left(\bar{\theta}_{r,n}^M > \tilde{Y}_{0\times 20\%,n}\right)}{N}$

47. The local-shift sensitivity implemented here, for ease of computation, perturbs the data value $\bar{x}$ closest to the mean to a hundred possible values x spread evenly over the domain of observed data values. The supremum (i.e., maximum) of the absolute difference in quantile estimates between the perturbed and mean data values is then calculated as the local-shift sensitivity. Further testing of the measure suggested that the measure itself is quite unstable, and hence less weight should be attached to this performance measure in assessing the overall performance of the Proponent's proposed methodology. Note that this performance measure is derived from model estimates of the data, and not from model estimates of simulations generated from a hypothetical 'true' distribution, unlike the other measures

48. Type I error is the frequency by which an observation from the hypothesized 'true' distribution is greater than the quantile estimator, i.e., it measures the rate of false positive declarations of a breach of an SSB when the underlying 'true' distribution of service distribution is below the SSB. False positives arise due to the stochastic nature of gaps in service from year to year. The desired frequency may nominally be set to $1 - \theta$, given the SSB θ being estimated. The frequency is calculated by simulating N yearly SAIDI or SAIFI scores $\tilde{Y}_{0,n}$ from the hypothetical 'true' distribution, and counting the number of times these simulated scores are greater than the simulated estimates of the SSB $\bar{\theta}_{r,n}^M$.

49. Type II error (or false negative rate) is the rate at which a shift in the underlying 'true' distribution is not detected by the SSB derived from a given estimation approach. An effect size needs to be defined before a Type II error rate can be defined, and in this instance the mean parameters of the mixed models were increased by 20%. Alternatively, the Type II error rate of an estimator can also be profiled along a range of different effect sizes, although profiling was not applied here. Nominally, a type II error rate of 20% is, as a rule of thumb, seen as acceptable (equating to a statistical power of 0.8).

50. Simulations were computed within an R environment.[28]

Results and Discussion

Reproducibility of WP estimation method

51. A proposed methodology may work well (i.e., is reproducible) most of the time. When a methodology is fitted to a sample of data that is collected infrequently, as is the case with the SSB estimates on the yearly scale, then the likelihood of detecting a failure in reproducibility is limited.

52. In contrast, the simulated output of the hypothetical mixture model provides a more extensive sand-box within which to test the reproducibility of an analysis as each simulation presents an opportunity to fit a proposed model.

53. In the current study, a tuning of the Proponents proposed methodology was required for the methodology to work across all hypothetical data scenarios (i.e., simulations). This tuning required a move away maximum likelihood methods of model fitting to contrast (or maximum goodness-of-fit) methods based on the Kolmogorov-Smirnov (KS) distance (options *'method="mge", gof="KS"'* for function *fitdist*).[29] We believe that if the data are not close to the fitted distribution then there may be a failure of convergence with the maximum likelihood methods, due to multimodality or extreme values in the data. In turn, the KS statistic provides an L1-norm that is nominally robust to these departures from model assumptions, although at risk of returning biased parameter estimates. The Anderson-Darling distance may instead be preferred as it gives more weight to extreme data values relevant to the estimation of SSB quantiles. However, application of the Anderson-Darling distance did not always lead to model solution and could not be applied reliably across the simulations.

54. Even with this tuning of the distributional fitting procedure both the Weibull and logistic distributions regularly failed to solve across all simulations. Candidate distributions for the basket of models to be considered in the model averaging should ideally be trialed extensively in a sand-box environment; distributions models that have convergence issues at the practical level of model fitting should be excluded from the basket.

55. Apart from the tuning requirements and the lack of identifiability of the Weibull and logistic distributions, the code provided by the Proponent was reproducible insofar that tuning requirements were readily identified and applied to novel data realisations.

[28] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vi enna, Austria, 2018, URL: https://www.R-project.org/.

[29] Delignette-Muller, M.L. and C. Dutang "fitdistrplus: An R Package for Fitting Distributions", *Journal of Statistical Software*, 64, pp. 1-34, 2015.

Performance of WP method relative to other methods

56. There was little to distinguish the different model averaging methods considered here and the first best model. BAITA performed marginally better for SAIDI quantile estimates based on the daily data (Tables 4 and 5) and estimates of the SAIFI quantiles based on the monthly data (Tables 10 and 11), based on prediction error. In contrast, the Proponent's method performed marginally better for SAIDI quantiles based on the monthly data (Tables 4 and 5) and SAIFI quantiles based on the daily data (Tables 10 and 11). There was no clear dominance of one method over another for maximising Pitman closeness, minimising local-shift error sensitivity and in minimising Type I error. It may be concluded that the superiority of one model averaging method over another is very much dependent on the data at hand. This finding emphasises the point that the superiority of an estimator must first be demonstrated before general assertions as to the superiority of an estimator are made.

57. The best performing estimator of the 97.5th and 99th quantiles were the kernel density estimates based on daily data. These estimators dominated the model averaging estimates to a large degree, reducing prediction error by approximately a third across the data scenarios considered here, and leading to lower Type II error rates (Tables 4, 5, 9 and 10).[30] Type I error rates were largely consistent with the quantile level (e.g., around 0.025 for the 97.5th quantile estimates and around 0.01 for the 99th quantile estimates). Moreover, the tolerance interval (the 2.5% to 97.5% confidence bounds placed on the quantile estimate) was more heavily skewed towards upper values for the model averaging methods than for the kernel density estimate. This suggests that certain members of the basket of candidate distributions considered by the model averaging will on occasion return unnecessarily high quantile estimates, indicating some instability in the estimation procedure.

Reliability of quantile estimation (97.5th vs 99th quantiles)

58. The 99th quantile was in all cases associated with larger standard errors and greater prediction error than the 97.5th quantiles. This difference in performance between quantiles was most strongly evidenced at the highest level of aggregation when the model averaging was applied to the yearly data (Tables 11 and 12). More significantly, the Type II error rate was much smaller for the 97.5th quantiles than for the 99th quantile estimates. Taken together, the implication is that the 97.5th quantile provides a more reliable estimator across the different statistical performance indicators, primarily because the 97.5th quantile is better able to correctly detect a shift in the underlying distribution if service standards worsen, and without overly penalising good service provision by an unnecessarily high Type I error rate (assuming only a single Bernoulli trial in which the Proponent is penalized only on a single SSB rather than across multiple SSBs). This finding is supported by theory, whereby the estimation of extreme quantiles is associated with greater uncertainty.[31]

[30] Unreasonably high Type I error rates (approaching one) tend to be associated with extremely low Type II error rates. Similarly, unreasonably large Type II error rates tend to be associated with extremely low Type I error rates. A 'good' estimator should provide both low and reasonable Type I and Type II error rates for a given effect size.

[31] *"The asymptotic formula for the variance of a quantile estimate is inversely proportional to the square of the probability density function evaluated at that quantile"*, in Brown, M.B. and R.A. Wolfe, "Estimation of the variance of quantile estimates", *Computational Statistics & Data Analysis*, 1, pp. 167-174, 1983.

Data aggregation

59. Overall, by accessing the finer resolution daily data then sufficient support was available for the kernel density estimates to represent distributional extremes in service failure data. Once the daily data have been modelled then it is a relatively straightforward process to estimate quantiles of the yearly SAIDI and SAIFI distribution from the daily distribution through aggregation of simulations of the kernel density estimate.

60. Aggregating the data to monthly year-long rolling averages provides SSB estimators that are a marked improvement over model averages applied to the yearly data (Tables 12 and 13). For BAITA and the single best model, highly uncertain estimates for an SSB will result when the SSB estimates based on five yearly values of a service provision indicator (as is the current method of the AER) are drawn from an underlying distribution characterised by high variability and/or heavy tails, leading to the fitting of a very 'flat' distributions. This flat distribution will likely over-estimate the extreme quantile values. Such 'non-identifiability' leads to significant instability in the model estimates. Indeed, any improvement provided by the Proponent's method over and above the current method of estimation using five yearly data points may be attributable to data disaggregation, rather than to any superiority of the Proponent's method of model averaging over the single best model. This positive impact of data disaggregation is supported also by the model averaging applied to the daily data being more precise, and generally performing better across all statistical performance measures than when model averaging was applied to the monthly rolling average data.

61. Moreover, it has been shown here that better performing estimators can potentially be developed through using the daily observations explicitly, prior to constructing yearly estimates of SSBs, when applying kernel density estimates. These kernel-based SSB estimates improve further the performance of the estimators significantly across most of the statistical performance measures.

62. That said, kernel density estimates perform poorly for data aggregated at the monthly and yearly scales, as these coarser scales do not provide sufficient support for estimation of extreme quantiles from the empirical density function. For example, 60 months of data are more likely than not to return a maximum valued data point that is located below the 99^{th} quantile of the underlying distribution. In contrast, for the daily data there are 1826 data points (365 days by five years). Thus, 18 data points would be expected to be observed above the 99^{th} quantile on average at the daily level, thereby providing much improved support for the kernel estimates prior to construction of the yearly quantiles (Tables 6, 7, 10 and 11).

63. Exclusion of candidate distributions that are more unstable (i.e., are heavier tailed, such as the GEV) from the model averaging basket may be considered when data are highly aggregated (i.e., at the yearly level).

Autocorrelation

64. Autocorrelation will mean a variance estimate will underestimate the true variance, including both the standard error and prediction error estimates of each SSB estimation procedure. As a consequence Type II error rates will likely be higher and Type I error rates lower, given a positive bias.

65. SAIDI and SAIFI scores derived as monthly 12-month rolling averages have an autocorrelation of close to one when fitting AR(1) models. In contrast, at a daily level SAIDI and SAIFI scores have an

autocorrelation of approximately 0.25. When the daily data are observed at monthly intervals then this autocorrelation falls to approximately 0.02.

66. The bias component of the RMSE prediction error estimates did increase to a noticeable extent with data aggregation; estimates derived from the monthly 12-month rolling average data were more biased than those derived from the daily data (Tables 4-11).

67. Disaggregation of the data to a daily level for model fitting may therefore be a desirable strategy to minimise autocorrelation. Data disaggregation will thus likely reduce Type II error rates of observing falsely no deterioration in the service provision if a deterioration were to occur.

On the use of the Anderson-Darling test

68. The Proponent proposes to apply the Anderson-Darling test prior to the AIC score calculation. The intent of the Anderson-Darling test is to exclude from the model averaging any component distributions that fit the data poorly.

69. To apply both the Anderson-Darling test and the AIC ranking within the same analysis occurs as epistemologically inconsistent. The AIC criterion is in theory sufficient to select from among the competing distributions those to be included in the model averaging. This is because those component distributions failing the Anderson-Darling test will generally have a high AIC score and will most likely be assigned zero weight in the model averaging. As such the poorer fitting distributions will likely fail on both criteria – the AIC and the Anderson-Darling test.

70. This assertion is supported by the Proponent's model averaging that includes the Anderson-Darling test to estimate SSB quantiles. These estimates do not differ from estimates generated from the Proponent's model averaging that does not include the Anderson-Darling test (Tables 4-12). For model averaging purposes the use of the Anderson-Darling test as a filter to exclude poor fitting models should be viewed as largely redundant.

71. There is technically a place for goodness-of-fit testing in identifying when none of the candidate models to be applied in the model fitting have merit (i.e., all models fail the goodness-of-fit test, such as the Anderson-Darling). In the setting of an SSB this has little application as there is likely no feasible instance where the SSB will not be identifiable through fitting a range of competing distributions.

72. The recommendation is to maintain epistemological correctness by excluding prior goodness-of-fit tests from an AIC based SSB estimation procedure.

Summary results

73. In summary:

    a. The Proponent's method is at best marginally more accurate than the single best model and BAITA in only some data scenarios, and marginally less accurate in other scenarios, as measured by RMSE estimate of the prediction error.

    b. The Proponent's method is unstable in its reproducibility when extended over the Monte Carlo experiment to measure the statistical performance of competing statistical methods. This is largely because of the non-zero failure rates associated with fitting the Weibull and logistic distributions to different levels of data aggregation (yearly, monthly 12-month rolling averages,

daily), and the need to tune the estimation method to use different model fitting measures such as the Kolmogorov-Smirnoff distance.

c. As with prediction error, the standard error estimate of precision (taken as a measure of replicability of estimates with different samples) was only marginally different between model averaging methods, including the single best model. No method dominated, with preference for one model averaging method over another varying with data scenario.

d. Statistical consistency implies that with increasing sample size then an estimator will converge to a true parameter value. The sample size effect may be measured as relative efficiency, although we have chosen for simplicity to go with Pitman closeness. In general, Pitman closeness was low for the Proponent's method of model averaging, principally because we measured strict closeness, insofar as an estimate has to be strictly less than the distance from the true parameter value than another estimate. However, as the Proponent's method is a committee method then for most simulations it returned a single best model, rather than an average of multiple models, whenever the single best model was found to have significantly lower AIC. Hence, Pitman closeness is much smaller for the Proponent's method under some of the data scenarios than for the other measures (principally Table 4). Pitman closeness was otherwise varied with data scenario.

e. BAITA may be considered as comparatively more robust than the Proponent's proposed method or the single best method, although the differences in local shift sensitivity were minimal. Moreover, estimates of the 97.5th quantile were observed to be more robust (i.e., have lower local-shift sensitivity) than the 99th quantile (Tables 4-11). The kernel density estimate was found to be more robust than any other of the other estimates, largely because as a locally weighted estimator less weight is generally given to extreme observations than for a distribution whose estimate of the scale and centre may be highly influenced.

f. Bias increased slightly for the 99th quantile estimates compared to the 97.5th quantile estimates. Bias increased to a large degree when data were aggregated as monthly 12-month rolling averages when compared to quantiles generated from the daily data.

g. Overall, 99th quantiles produced less stable estimates than 97.5th quantile estimates, as indicated by poorer performance across most of the statistical performance metrics.

Reasonableness and statistical best practice

74. The Code requires that the service standard benchmark for a reference service must be both reasonable and sufficiently detailed to be applied by other stakeholders in the market.[32]

75. Best practice in statistical methodology includes facilitating reproducibility of estimates (i.e., given a data set then reproduce an estimate). In today's information driven environment all work should be at least reproducible as a basic requirement of an evidence-based approach to decision making. This reproducibility is facilitated by the sharing of code used to generate an estimate.[33] The Proponent has

[32] Section 5.6 b), Electricity Industry Act 2004, 30th November 2004.

[33] For example, Marwick B., Boettiger C. and L. Mullen, "Packaging Data Analytical Work Reproducibly Using R (and Friends)", *The American Statistician*, 72(1), pp. 80-88, 2018.

met satisfactorily the minimum requirement of providing sufficient detail of their SSB estimation methods, although the performance of their methods appears to be on occasion dependent on the data scenario.[34]

76. There are several different statistical methodologies that can be applied to the problem of estimating a quantile. Significantly, the Proponent's method differs from a strict implementation of the BAITA from which it was derived. The Proponent has provided no reasoning for this departure, other than stating that: [35]

"Western Power's methodology for selecting candidate models differs in some respects from those outlined above. However, Western Power's approach also follows the general pattern of using AIC values to establish a ranking of candidate models, and a threshold based on the distance from the lowest observed AIC value."

77. The issue here is that the BAITA was designed to minimize prediction error attributable to model selection bias, that is, BAITA derived estimates perform optimally by some statistical performance measure.[36]

78. The reproducibility of an analysis is not necessarily sufficient for a methodology to be accepted. The analysis also needs to be rigorous: [37]

"Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated, [38] it does not change the fact that problematic research is conducted in the first place."

79. Similarly, a proposed method should not be accepted based solely on an appeal to reasonableness.[39] What is reasonable from a statistical perspective is that a proposed method can demonstrate optimal performance, *vis a vis* some other competing method. Hence, a reasonable method is one that is considered as best-practice, which in turn is optimal according to some objective criterion.

80. Furthermore, what is economically reasonable should also be considered. For example, economic 'reasonableness' potentially involves minimising inappropriate monetary transfers between

[34] Paragraphs 51-55 above.

[35] Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 6.

[36] *"Among the other benefits of this approach, it effectively rules out null hypothesis testing as a basis for model selection because multimodel inference forces a deeper approach to model selection. It means we must have an optimality criterion and selection (weight assignment) theory underlying the approach",* p. 298 in Burnham, K.P. and D.R. Anderson. "Multimodel inference: Understanding AIC and BIC in model selection", *Sociological Methods Research*, 2004, pp. 261-304.

[37] Leek, J.T. and R. D. Peng, "Reproducible research can still be wrong: Adopting a prevention approach", *PNAS,* 112, pp. 1645-1646, 2015.

[38] Herndon T., Ash M, and R. Pollin, "Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff", *Cambridge Journal of Economics*, 38, pp. 257-279, 2014.

[39] Section 5.6 of the Access Code.

Proponent and customer as in indicator of market efficiency. Consideration of what is economically reasonable is outside of the current terms of reference. However, statistical best practice can be coupled with criteria for economic 'reasonableness'.

81. For example, given the choice of quantile and the uncertainty in those quantile estimates in the setting of the SSBs then costs to each party of committing an error in SSB estimation should be understood. Critically, if an SSB estimate is highly uncertain then the notional conditional value-at-risk[40] above the SSB is greater than for a more certain SSB estimate, assuming both estimates are equal.

82. The statistical parallels for a value-at-risk metric are the Type I and Type II error rates. Type I error rates were observed to be relatively stable across the different data scenarios, and consistent with the significance level represented by the quantile (e.g., 2.5% significance level for a 97.5th quantile). Type II error rates however varied across the data scenarios, with Type II error rates tending to increase with high uncertainty in parameter estimates, right skewness observed in the tolerance interval surrounding each model averaging estimate, and a tendency towards a large bias value with aggregation of the data. Higher uncertainty in parameter estimates, and hence higher Type II error rates, were also observed for the 99th quantile estimates compared to the 97.5th quantile estimates.

83. One could therefore design a risk-minimising SSB simply by choosing a data aggregation method that minimises bias and prediction error (i.e., apply estimation methods to the daily data before computing yearly quantiles), choosing the 97.5th quantile over the 99th quantile, and developing further the use of non-parametric methods of estimation (until such a stage as a better performing estimation method is proposed).

84. In contrast, the system could potentially be gamed by selecting an SSB estimate that has high associated uncertainty. Nominally, this would be an estimator based on a higher level of data aggregation, selecting an estimator that has higher associated uncertainty relative to competing methods, and that is right skewed in such a way as to produce a positively biased estimate of the quantile. In this scenario the Type I error will be relatively small as there would be a low probability of observing a yearly service provision indicator value above the SSB. Hence, fewer inappropriate costs would accrue to the Proponent arising from a false declaration of a breach of the service standard when in fact the service standard has not declined. More significantly though, the Type II error rate could increase significantly as the SSB loses its sensitivity in detecting a breach of the service standard when in truth the service standard has declined. The most significant driver of an increase in Type II error rate observed across the different data scenarios (Tables 4-11) was an increase in the SSB from the 97.5th quantile to the 99th quantile.

[40] The conditional value at risk (CVaR) is the probability weighted sum of all 'losses' beyond a threshold value. As SAIDI and SAIFI values arising from supply interruptions can be said to scale approximately one-to-one with financial losses associated with those supply interruptions (i.e., longer duration of interruptions equates to greater cost), then sum observed SAIDI and SAIFI scores above an SSB can be said to constitute a CVaR metric.

The multiple trial problem

85. The core of the Proponent's argument for increasing the SSB from the 97.5th quantile to the 99th quantile is to avoid the issues surrounding the multiple trial problem, which focuses on the Type I error rate. In this regard advice to the Proponent states that:[41]

    "Assuming stable performance, the sampling of the 97.5th quantile should indicate a 2.5% probability of exceedance per metric. If the current 17 AA3 SSBs were fully independent, this would result in a 34.98% chance of exceeding at least one per year; effectively necessitating performance improvement to ensure compliance. While the metrics are not fully independent, the impact is still valid.

    In AA4, Western Power is proposing network investment to maintain service performance. The proposed network investment aligns closely with customer satisfaction analysis, indicating that customers are satisfied with the current level of performance. As such, Western Power proposes the use of the 99th quantile for setting SSBs. With a 1% probability of exceeding each metric, the total result is a 15.7% probability of exceeding at least one per year. The reduced probability better aligns with the goal of maintaining performance and the proposed investment."

86. In the multiple trial problem a service provision indicator may exceed the SSB simply through randomness, without there being an actual worsening in the service provision itself (i.e., a false positive). This is not necessarily a problem if only a single SSB is being considered. However, if multiple SSBs are considered, and the Proponent is penalised whenever at least one SSB is breached, then the odds of at least one false positive occurring across all of the SSBs begin to compound. Hence, when a market actor can get penalized on one of many indicators the false positive rate (Type I error), and hence the expected penalty, can be significantly higher than the nominal Type I error rate defined for a single SSB.

87. For example, for a 97.5th quantile the nominal Type I error rate is 0.025 (or one-in-40-years). If there are five SSBs, and the Type I error rate is consistent and independent across those SSBs, then a Binomial probability will state that the Type I error rate across all five SSBs will be:

$$P(N(X \geq SSB) \geq 1) = 1 - P(N(X \geq SSB) = 0) = 1 - \binom{T}{0} P(X \geq SSB)^0 \big(1 - P(X \geq SSB)\big)^T$$
$$= 1 - \big(1 - P(X \geq SSB)\big)^T$$

    where $N(X \geq SSB)$ is the number of false positive breaches where the service provision indicator X is above the SSB, and $P(\cdot)$ is a probability measure.

88. Assuming a single trial false positive rate of $P(X \geq SSB) = 0.025$ results in a multiple trial Type I error rate for $T = 5$ trials of $P(N(X \geq SSB_{0.975}) \geq 1) = 0.119$.

89. Equivalently, for a 99th quantile (or one-in-100-year event), the multiple trial probability of a Type I error across five trials will be $P(N(X \geq SSB_{0.99}) \geq 1) = 0.049$.

90. However, these calculations naively assume independence among the different service provision indicators. The service provision indicators are likely correlated (e.g., higher duration of service

[41] Western Power, *Fitting Distributions for AA4 Service Standard KPIs – Setting the Service Standard benchmark (SSB) and Service Standard Target (SST)*, Attachment 6.2 – Access Arrangement Information, 2nd October 2017, p.11.

interruptions is likely correlated with more service interruption events, which in turn correlates with the number of call centre events). If correlation among these indicators were high then the Type I error rate of multiple indicators can be significantly less than what multiple independent trials would suggest. As such the Type I error rate for multiple service indicators when independence is assumed presents only an upper bound which will likely not be realised in practice. More work on estimating the multiple trial Type I error rate would need to be undertaken, with correlation among the indicators well modelled, before evidence may be presented to sufficiently inform a decision maker that a change in nominal Type I error rates (i.e., the quantile SSB) is required.

91. Furthermore, although the nominal Type I error rate is implicitly defined by the setting of the SSB quantile, the actual Type I error rate may differ depending on which service provision indicator and data support is selected for the estimation of the SSB. For instance, the Type I error rate was estimated to range from 0.0076 to 0.0491 for the 97.5th quantile across the SAIDI and SAIFI indicators (Tables 4, 6, 8 and 10). Attention should be applied to estimated Type I error rates when considering the Type I error rate of multiple trials.

92. Again, the solution as to what an appropriate multiple trial Type I error rate should be is largely economic, as Type I errors will have to be traded off against Type II errors (i.e., one would have to incur a higher Type I error rate before reducing the Type II error rate). If costs of Type I and Type II errors were known then it would potentially be feasible to optimally designing the multiple trial problem to minimize inappropriate monetary transfers between market actors.

93. If the costs associated with a higher, multiple test Type I error rate were greater than the associated costs of a high Type II error rate then one could also adjust the requirement of needing at least one SSB to be in breach before declaring that a penalty need be imposed on the Proponent. In this scenario $P(N(X \geq SSB) \geq t)$ would be modelled, where $t \geq 2$. This would reduce the Type I error rate of five trials significantly $(P(N(X \geq SSB_{0.975}) \geq 2) = 0.0285$ and $(P(N(X \geq SSB_{0.975}) \geq 2) = 0.0105$, back to the nominal single trial Type I error rates. This could well be a preferable option to simple increasing the quantile by which the SSB is measured, given greater risk can be associated with the estimation of more extreme quantiles.

94. Regardless, deciding on a way forward in solving the multiple trial problem is non-trivial. In no way has the Proponent put forward a convincing argument that relaxing the SSB quantile is an appropriate step to take, or that their method is at all superior to alternate methods of quantile estimation.

Is model averaging of parametric models better than non-parametric estimation?

95. In particular, the evidence demonstrating the superiority of the Proponent's proposed method is weak for the following reasons:

    a. The superiority of the their BAITA derived method is not demonstrated to be superior over the BAITA specified by its authors in any manner. The overarching conclusion with regard to model averaging is that the Proponent's assertion that their method is best-practice cannot be applied generically across all data scenarios. Indeed, the single best model outperforms the Proponent's method under some data scenarios. Moreover, applying an unequal weighting (as

implemented under the BAITA) will typically be superior to equal weights (as implemented by the Proponent).[42]

b. The Proponent's methodology is mis-specified insofar as BAITA specifies both a $\Delta_i < 10$ cut-off for component models to be included in the model averaging, and for an Akaike weight to be calculated from each Δ_i value.[43] Instead, the Proponent states:[44]

"Alternatively, Richards (2005) proposes that any Δ_i less than 10 be considered an acceptable model. Critically, there is no single threshold value for which there is uniform agreement across all authors.

A more complex methodology is set out in Symonds & Moussalli (2011) in which a weight is calculated from each Δ_i value."

This discussion in the broader literature regarding appropriate cut-off's is largely for convenience given component models may be computationally burdensome to estimate. In BAITA theory, any model with a non-zero Akaike weight may be included in the model averaging under BAITA. For the most part, arbitrary cut-offs such as $\Delta_i > 10$ exclude those component models with zero or near zero Akaike weights. If there is uncertainty around these cut-off values[45] then the default position should arguably be to not apply cut-offs. Regardless, it is preferable to weight component models by their Akaike weights, rather than equal weights that the Proponent applies. The strict BAITA has been designed in part to minimise model selection bias, and issues with equal weighting are well known in theory.

c. The Proponent provides a positive example of where their 1% threshold coincides with a 0.95 cut-off applied to the Akaike weights w_i.[46] This example is co-incidental. Note that the AIC scores vary from SSB indicator to SSB indicator (ranging from 415.06 to 655.9 for SAIDI scores

[42] Dormann, C.F., Calabrese J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barto, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schroder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wuest, R.O., and F. Hartig, "Model averaging in ecology: a review of Bayesian, information-theoretic and tactical approaches for predictive inference", *Ecological Monographs*, in press, URL: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1309.

[43] Burnham, K.P. and D. R. Anderson, *Model Selection and Multimodel Inference: A practical information-theoretic approach*, 2002, Springer-Verlaag, New York, Second Edition, 515 pp.

[44] Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 5.
Moussalli, A. and M.R.E. Symonds, "A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion", *Behavioural Ecology and Sociobiology*, 65, pp.13-21, 2011.
Richards, S.A., "Testing ecological theory using information-theoretic approaches: examples and cautionary results", *Ecology,* 86, pp. 2804-2814, 2005.

[45] "*All methods depend on the use of a threshold value, for which there is no uniquely agreed value in the peer-reviewed literature*" in Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 7.

[46] Table 1. Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 6.

and -185.59 to 86.89 for SAIFI scores).[47] There is no reason why cumulative Δ_i should scale reliably with AIC score, so a 1% AIC threshold will not always correspond to a 0.95 cut-off applied to the Akaike weights. To demonstrate this point, Figure 1 maps the 0.95 cumulative Akaike weight to the 1% cut-off point for all 10,000 simulations from the hypothetical mixture model. As can be seen, the cumulative Akaike weight is widely distributed relative to the 1% cut-off. As such, providing a single example of confirmation, without objectively testing for examples of non-confirmation, may be categorized as a *'fallacy of the lonely fact'*.

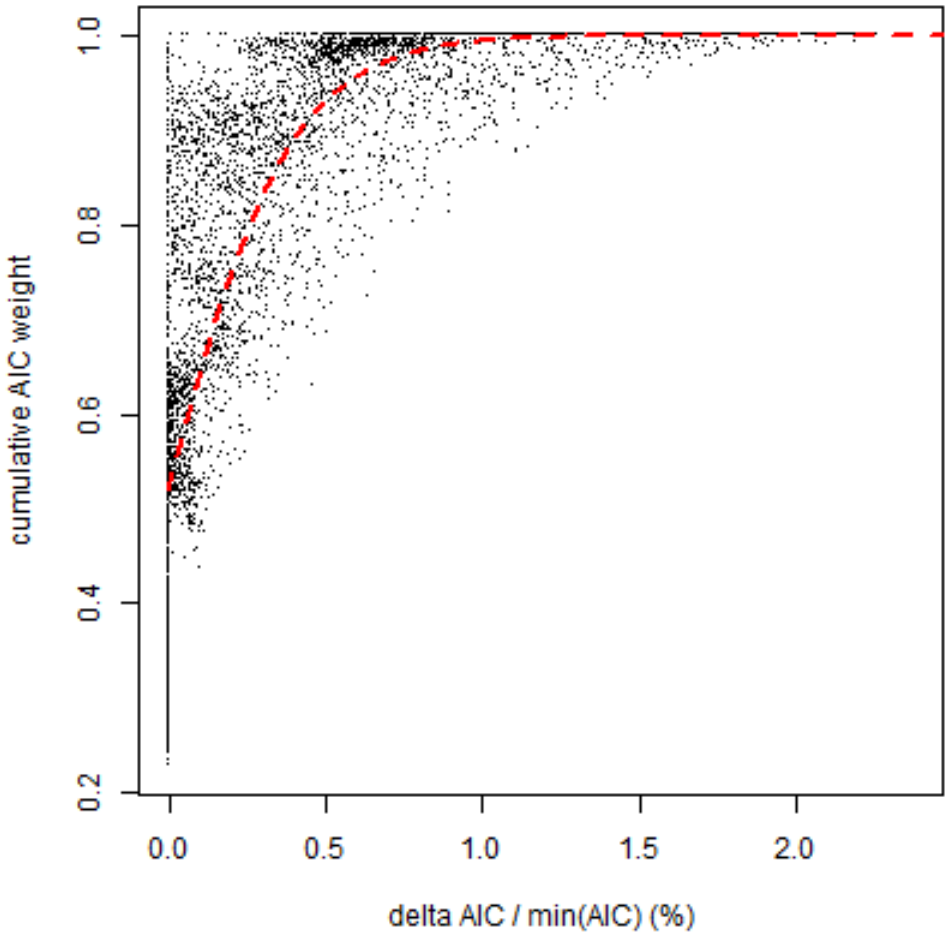

Figure 1. Correlation between the cumulative sum of AIC weights and the Proponent's 1% cut-off. A cut-off to exclude models in the averaging set at $\Delta_i/min(AIC_i) > 0.01$ equates to a cumulative sum of the Akaike weights of 0.994, not 0.95, given a quasibinomial generalized linear model fit of the data (red line). In contrast, a 0.95 cut-off applied to the cumulative sum of Akaike weights equates to a 0.57% cut-off. Note that there is significant variability about the trend line, which illustrates that a single coincidental 'lonely fact' residing within the variability about the trend line is not sufficient evidence to justify a choice of cut-off.

[47] WP guidelines

96. There is some validity to advice provided to the Proponent:[48]

> *In paragraph 1018(2) of the Draft Decision, the ERA notes that "the composition and number of distributions selected within the threshold value are likely to vary with time, introducing volatility and uncertainty". The ERA's observations are valid. The selection of candidate statistical models may change over time when using Western Power's methodology.*
>
> *However, the alternative solution of selecting a single statistical model will only serve to exacerbate this source of variability. A change in the composition of which models are selected will have less of an effect on the quantile estimates than shifting entirely from one single distribution to another single distribution. If intertemporal consistency is indeed a priority, then the preference should be for Western Power's averaging methodology over the selection of a single distribution.*

97. In general, some form of model averaging will outperform a single best fit distribution. Model averaging is advantageous in that model selection bias is reduced, and consequently prediction error of an estimator is also reduced.[49]

98. Model averaging is particularly useful if the predictive error of contributing model predictions is dominated by variance, and if the covariance between models is low.[50] This is likely not the case with the current set of candidate models considered by the proponent, as many of the models are similar. For instance, there are two variants of the log-logistic model, with one a generalization of the other, and this is also true of the Weibull, log normal, and gamma distributions. Hence, we do not see any significant improvement in performance between the different AIC methods; in fact preference for one method over another would be dependent on both the service provision indicator and the data construct. This emphasizes the point that the efficacy of model averaging is highly context dependent, and that for a given context some work needs to be done to demonstrate the superiority of model averaging over other methods.

99. The Proponent suggests also that the large volatility in AIC in response to small changes in data is justification for applying model averaging.[51] While this statement has some legitimacy, it does not consider the sensitivity of the underlying distribution generating the data to small changes in the data. This is the key reason why an invariant 'true' distribution (as in the hypothesized mixture model) is applied within a simulation experiment, and from there the performance of a proposed SSB estimator examined.

---

[48] Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018, p. 7.

[49] Hastie, T., Tibshirani, R. and J. Friedman, *The Elements of Statistical Learning: Data mining, inference and prediction,* Springer-Verlaag: New York, 2nd Edition, p. 289.

[50] Dormann, C.F., Calabrese J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barto, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schroder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wuest, R.O., and F. Hartig, "Model averaging in ecology: a review of Bayesian, information-theoretic and tactical approaches for predictive inference", *Ecological Monographs*, in press, URL: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1309.

[51] Western Power, *Fitting Distributions for AA4 Service Standard KPIs – Setting the Service Standard benchmark (SSB) and Service Standard Target (SST)*, Attachment 6.2 – Access Arrangement Information, 2nd October 2017, p. 13.

100. Clearly, if two competing models have different SSB estimates, and the first best model repeatably switches between these two models depending on yearly random shifts in the data, then there is going to be high volatility in the SSB estimate. However, this can be argued to be a reflection more of volatility in the underlying distribution of service events (even after data have been smoothed with a 5-year rolling average) than any inherent volatility introduced by a switching first-best model estimate of the SSB. This argument is supported by the minimal observed difference between the single best AIC model and the Proponent's AIC model averaging method, as measured by the statistical performance measures (Tables 4-11).

101. Moreover, other statistical methods for estimating distributional quantities may be considered. A key class of estimators not considered by the Proponent are non-parametric kernel density estimators (KDEs).[52] KDEs are applicable whenever the assumptions of parametric distributions fail. These failures of assumption typically occur when the underlying distribution generating the data is multi-modal and/or heavily skewed.

102. More success could feasibly be achieved through a model averaging of the single best AIC model based on monthly data and the kernel density estimate applied to the daily data. This is perhaps best achieved through model stacking[53] rather than the BAITA approach. Moreover, validation-based methods of model averaging have been recommended over BAITA based approaches. The main reason for having semi-independent test data is that: [54]

"Statistical models, which aim to describe the data to which they are fitted, will often have correlated parameters and fits; process models may overlap in modelled processes. Having highly similar models in the model set will inflate the cumulative weight given to them."

Conclusions

103. The Proponent's model averaging brings only minimal improvements over that of the single-best model in estimating the 97.5th and 99th SSB quantiles, and does not demonstrate any dominance over the standard BAITA approach. This is because:

a. There is high correlation among component models (e.g., the three parameter Weibull model is a generalization of the two parameter Weibull model).

b. The data have been aggregated as a monthly 12 month rolling average. It may be speculated that under the central limit theorem then regular distributions fit reasonably well to the data, and so little improvement in performance is observed when model averaging is applied.

104. The Proponent's methodology is flawed from a statistical perspective insofar as the Proponent provides no measure of uncertainty associated with their SSB estimate. The need for statistical performance measures to be associated with SSB estimates may be deemed as 'reasonable' under

[52] Wand, M.P and M.C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC: London, 1995.

[53] Hastie, T., Tibshirani, R. and J. Friedman, *The Elements of Statistical Learning: Data mining, inference and prediction,* Springer-Verlaag: New York, 2nd Edition, p. 290.

[54] Dormann, C.F., Calabrese J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barto, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schroder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wuest, R.O., and F. Hartig, "Model averaging in ecology: a review of Bayesian, information-theoretic and tactical approaches for predictive inference", *Ecological Monographs*, in press, URL: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1309.

the Act, and as part of statistical best practice. It is also noted that past practice has not associated standard error measures with the SSB estimates. However, to demonstrate the statistical 'superiority' of a proposed estimator over and above the current methodology one must necessarily employ measures of statistical performance.

105. The Proponent's methodology does not consider non-parametric estimators of the SSB quantiles. Under simulations from a hypothetical mixture model fitted to the daily data then a kernel density estimate of the SSB quantiles was shown to be distinctly superior to the Proponent's methodology and the single best AIC model, as measured by five different statistical performance measures (standard error, root mean square error, Pitman closeness and approximate local-shift sensitivity).

106. The method of data aggregation applied by the Proponent is to be preferred to the method of fitting distributions to five yearly data points for SAIDI, SAIFI or other service performance indicators. However, the monthly 12-month rolling average of the data were autocorrelated, more so than the daily data, leading in theory to an underestimation of the uncertainty in the SSB estimates (when these standard error estimates are provided). Moreover, fitting either the single best model or the BAITA model average to the daily data to then provide yearly SSB estimates performed better than when these models (including the Proponent's method) were fitted to the monthly rolling average data.

107. Estimation of the 97.5th quantile as the definition of the SSB is to be preferred to the 99th quantile because:

a. Estimates of the 99th quantile were more uncertain than estimates of the 97.5th quantile (evidenced by larger standard errors and prediction error). Uncertainty implies risk, which in turn may reduce the efficient operation of a regulated market as stakeholders may seek to hedge against the risk of inappropriate monetary transfers between market actors.

b. Advice to the Proponent that lower SSBs do not guarantee improved services for all customers is specious, principally because it was not demonstrated that higher SSBs guarantee improved services for all customers.[55] If anything, a relaxation of the SSB through employing a more extreme quantile will diminish the provision of services to customers simply because less will need to be invested to satisfy the SSB (as indicated by higher Type II error rates for the 99th quantile). The Proponent has considered only Type I error (false positives) and not Type II error (false negatives).

108. The Proponent is correct in that implementing a lower quantile will lead to a higher rate of not satisfying the collar SSB through chance variation. This is because there are multiple service performance indicators at play. In this instance, the chance of at least one performance indicator breaching its respective SSB (this chance increases with the number of performance indicators applied) can be much greater than the chance of a single performance indicator breaching its SSB. [1]

109. The solution here is not necessarily to increase the SSB to adjust for the higher false positive rate of multiple indicators (i.e., the service performance indictor is greater than the SSB simply through

[55] Analytics + Data Science, *Methodology for setting the service standard benchmarks and targets – expert report, Report prepared for Western Power* as 'Attachment 13.1 – revised proposed access arrangement information', 6 June 2018.

chance).[56] Instead, setting a target for the false positive rate by allowing multiple indicators to breach their respective SSBs may be considered. For example, applying the collar when at least two service performance indicators breach their SSBs may be preferable to applying the collar when at least one service performance indictor breaches an SSB that is set at a higher level. This is because the Type II error rate (i.e., false negatives) may be unacceptably high with a higher SSB quantile. The proponent considers only Type I errors under the multiple trial problem, and does not consider costs associated with Type II errors.

110. Critically, the costs of a false positive may be asymmetric when compared to those of a false negative (i.e., there has been a decline in service provision, but this is not detected by the service performance indicator exceeding the SSB, i.e., a Type II error). However, an analysis based on costs will likely be contentious as relevant costings will need to be agreed upon among stakeholders. Hence, arbitrarily changing the false positive rate by increasing the quantile on which the SSB is based appears fraught, especially as the costs associated with false positive and false negative rates are largely unknown, and a higher level of uncertainty is associated with the more extreme SSB quantile.

111. Overall, the candidate models proposed for the model averaging may be viewed as highly correlated. Hence, the Proponent's implementation of model averaging does not achieve the benefits over the single best fit model that model averaging should deliver in theory. Importantly, another class of SSB estimators derived from kernel density estimates has been shown to outperform the Proponent's methodology in simulation trials, when applied at the even more granular level of the daily data. Moreover, any proposal to increase the SSB to mitigate Type I error associated with the multiple trial problem should be viewed with skepticism until costs associated with Type II errors are better considered.

[56] Note this false positive rate is nominally one-in-40-years for a 97.5th quantile and one-in-100-years for a 99th quantile when only one service performance indicator is considered.

Glossary

ACRONYM	DEFINITION
A+DS	Analysis + Data Science
AER	Australian Energy Regulator
AIC	Akaike Information Criterion
ARIMA	Autoregressive, integrated, moving average
BAITA	Burnham and Anderson information-theoretic approach
KDE	Kernel density estimate
LOSEF	Loss of service
SAIDI	System average interruption duration index
SAIFI	System average interruption frequency index
SSB	Standard service benchmark.

Appendix A: Tables of Simulation Results

SSB values for annual SAIDI and SAIFI scores of hypothetical 'true' mixture models

Table 3. Quantiles of hypothetical 'true' mixture model distribution, filtered using the beta threshold method.

	SAIDI	SAIFI
97.5%	194.98	1.751
99%	200.54	1.786

Estimates of the 97.5th and 99th quantile for SAIDI daily and monthly aggregated data

Table 4. Estimates of the 97.5% quantile generated from fitting distributions to the daily SAIDI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	195.3	195.7	195.4	195.7	193.8
Standard error	12.39	12.57	11.39	12.57	8.52
2.5% lower bound	180.1	180.1	180.1	180.1	177.2
Median	193.9	194.0	194.1	194.0	193.4
97.5% upper bound	214.5	223.0	215.2	223.0	211.7
Bias^2	0.082	0.487	0.154	0.487	1.411
Prediction Error	12.39	12.58	11.39	12.58	8.60
Pitman closeness	-	0.0070	0.0280	0.0070	0.470
Local-shift error sensitivity	0.0033	0.0033	0.0033	0.0033	0.0023
Type I error	0.0241	0.0224	0.0236	0.0224	0.0303
Type II error with 20% effect size	0.4020	0.4119	0.4043	0.4119	0.3636

Table 5. Estimates of the 99% quantile generated from fitting distributions to the daily SAIDI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	201.9	202.7	202.1	202.7	198.8
Standard error	17.42	17.94	15.57	17.94	8.95
2.5% lower bound	185.3	185.3	185.4	185.3	181.3
Median	200.0	200.0	200.1	200.0	198.5
97.5% upper bound	221.6	231.2	222.8	231.2	217.9
Bias^2	1.895	4.486	2.482	4.486	3.036
Prediction Error	17.48	18.06	15.65	18.06	9.12
Pitman closeness	-	0.6800	0.6770	0.6800	0.6350
Local-shift error sensitivity	0.0042	0.0042	0.0042	0.0042	0.0032
Type I error	0.0080	0.0070	0.0077	0.0070	0.0133
Type II error with 20% effect size	0.5717	0.5892	0.5763	0.5892	0.4919

Table 6. Estimates of the 97.5% quantile generated from fitting distributions to the monthly SAIDI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	190.6	190.7	190.6	190.7	171.3
Standard error	13.79	12.28	12.69	12.28	7.14
2.5% lower bound	169.5	169.9	170.0	169.9	157.8
Median	189.0	189.8	189.6	189.8	171.3
97.5% upper bound	219.4	215.5	216.0	215.5	185.6
Bias^2	19.28	18.10	19.05	18.10	562.9
Prediction Error	14.47	13.00	13.42	13.00	24.78
Pitman closeness	-	0.5880	0.6270	0.5880	0.0590
Local-shift error sensitivity	0.0302	0.0293	0.0280	0.0293	0.0054
Type I error	0.0491	0.0478	0.0489	0.0478	0.4050
Type II error with 20% effect size	0.2907	0.2936	0.2911	0.2936	0.0230

Table 7. Estimates of the 99% quantile generated from fitting distributions to the monthly SAIDI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	195.8	195.8	195.7	195.8	171.8
Standard error	19.86	16.58	17.58	16.58	7.20
2.5% lower bound	171.5	172.2	172.4	172.2	158.1
Median	192.9	194.1	193.4	194.1	171.8
97.5% upper bound	237.8	227.8	228.0	227.8	186.3
Bias^2	22.43	22.57	23.09	22.57	826.8
Prediction Error	20.41	17.25	18.22	17.25	29.64
Pitman closeness	-	0.6120	0.6600	0.6120	0.0520
Local-shift error sensitivity	0.0408	0.0400	0.0383	0.0400	0.0070
Type I error	0.0219	0.0219	0.0221	0.0219	0.3900
Type II error with 20% effect size	0.4151	0.4148	0.4131	0.4148	0.0261

Estimates of the 97.5th and 99th quantile for SAIFI daily and monthly aggregated data

Table 8. Estimates of the 97.5% quantile generated from fitting distributions to the daily SAIFI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	1.729	1.795	1.729	1.795	1.760
Standard error	0.1094	0.0508	0.1024	0.0508	0.0483
2.5% lower bound	1.6048	1.698	1.605	1.698	1.670
Median	1.698	1.794	1.702	1.794	1.759
97.5% upper bound	2.029	1.897	2.013	1.897	1.863
Bias^2	0.0005	0.0019	0.0005	0.0019	0.0001
Prediction Error	0.1117	0.0672	0.1047	0.0672	0.0491
Pitman closeness	-	0.6750	0.3270	0.6750	0.7680
Local-shift error sensitivity	0.0033	0.0033	0.0033	0.0033	0.0023
Type I error	0.0433	0.0076	0.0427	0.0076	0.0204
Type II error with 20% effect size	0.1503	0.3551	0.1516	0.3551	0.2378

Table 9. Estimates of the 99% quantile generated from fitting distributions to the daily SAIFI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	1.765	1.841	1.766	1.841	1.794
Standard error	0.1226	0.0528	0.1144	0.0528	0.0495
2.5% lower bound	1.635	1.740	1.634	1.740	1.702
Median	1.730	1.839	1.733	1.839	1.794
97.5% upper bound	2.107	1.948	2.086	1.948	1.898
Bias^2	0.0004	0.0029	0.0004	0.0029	0.0001
Prediction Error	0.1244	0.0756	0.1162	0.0756	0.0501
Pitman closeness	-	0.6650	0.3320	0.6650	0.7880
Local-shift error sensitivity	0.0042	0.0042	0.0042	0.0042	0.0032
Type I error	0.0181	0.0022	0.0178	0.0022	0.0078
Type II error with 20% effect size	0.2515	0.5288	0.2542	0.5288	0.3521

Table 10. Estimates of the 97.5% quantile generated from fitting distributions to the monthly SAIFI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	1.733	1.734	1.735	1.734	1.595
Standard error	0.0768	0.0763	0.0736	0.0763	0.0420
2.5% lower bound	1.609	1.609	1.613	1.609	1.512
Median	1.726	1.727	1.728	1.727	1.594
97.5% upper bound	1.906	1.911	1.892	1.911	1.680
Bias^2	0.0004	0.0004	0.0003	0.0003	0.0245
Prediction Error	0.0790	0.0784	0.0755	0.0784	0.1619
Pitman closeness	-	0.4000	0.5700	0.4000	0.0540
Local-shift error sensitivity	0.0302	0.0293	0.0280	0.0293	0.0054
Type I error	0.0396	0.0389	0.0377	0.0389	0.3906
Type II error with 20% effect size	0.1588	0.1604	0.1645	0.1604	0.0083

Table 11. Estimates of the 99% quantile generated from fitting distributions to the monthly SAIFI data

Statistical Performance Measure	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC Proponent 1% cut-off and AD test	Kernel density estimator
Mean	1.766	1.767	1.769	1.767	1.599
Standard error	0.1007	0.1001	0.0944	0.1001	0.0423
2.5% lower bound	1.627	1.627	1.631	1.627	1.515
Median	1.752	1.753	1.757	1.753	1.600
97.5% upper bound	2.015	2.017	1.988	2.017	1.685
Bias^2	0.0004	0.0004	0.0003	0.0004	0.0352
Prediction Error	0.1028	0.1019	0.0960	0.1019	0.1922
Pitman closeness	-	0.4100	0.6070	0.4100	0.0550
Local-shift error sensitivity	0.0408	0.0400	0.0383	0.0400	0.0070
Type I error	0.0179	0.0176	0.0167	0.0176	0.3735
Type II error with 20% effect size	0.2525	0.2570	0.2630	0.2570	0.0092

Statistical performance measures for yearly SSB data

Table 12. Quantile estimates generated from fitting distributions to the yearly SAIDI data

Statistical Performance Measure	97.5th quantile			99th Quantile		
	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC single best	AIC Proponent 1% cut-off	AIC BAITA
Mean	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞
Standard error	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞
2.5% lower bound	162.8	163.2	164.3	164.4	164.8	165.8
Median	186.4	186.7	188.4	189.8	190.3	193.0
97.5% upper bound	476.9	382.0	313.0	1143.1	884.7	619.0
Bias²	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞
Prediction Error	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞	→ ∞

Table 13. Quantile estimates generated from fitting distributions to the yearly SAIFI data

Statistical Performance Measure	97.5th quantile			99th Quantile		
	AIC single best	AIC Proponent 1% cut-off	AIC BAITA	AIC single best	AIC Proponent 1% cut-off	AIC BAITA
Mean	1.696	1.697	1.712	1.853	1.853	1.897
Standard error	0.4500	0.4499	0.4421	3.757	3.757	3.726
2.5% lower bound	1.529	1.532	1.537	1.533	1.534	1.546
Median	1.657	1.655	1.671	1.677	1.677	1.702
97.5% upper bound	1.991	1.991	2.040	2.192	2.192	2.383
Bias²	0.0030	0.0030	0.0015	0.0044	0.0044	0.0122
Prediction Error	0.4533	0.4532	0.4438	3.758	3.758	3.728

Appendix B: Expert Witnesses in Federal Court Proceedings

FEDERAL COURT OF AUSTRALIA

Practice Note CM 7

EXPERT WITNESSES IN PROCEEDINGS IN THE FEDERAL COURT OF AUSTRALIA

Practice Note CM 7 issued on 1 August 2011 is revoked with effect from midnight on 3 June 2013 and the following Practice Note is substituted.

Commencement

1. This Practice Note commences on 4 June 2013.

Introduction

2. Rule 23.12 of the Federal Court Rules 2011 requires a party to give a copy of the following guidelines to any witness they propose to retain for the purpose of preparing a report or giving evidence in a proceeding as to an opinion held by the witness that is wholly or substantially based on the specialised knowledge of the witness (see **Part 3.3 - Opinion** of the *Evidence Act 1995* (Cth)).

3. The guidelines are not intended to address all aspects of an expert witness's duties, but are intended to facilitate the admission of opinion evidence[57], and to assist experts to understand in general terms what the Court expects of them. Additionally, it is hoped that the guidelines will assist individual expert witnesses to avoid the criticism that is sometimes made (whether rightly or wrongly) that expert witnesses lack objectivity, or have coloured their evidence in favour of the party calling them.

Guidelines

1. General Duty to the Court[58]

1.1 An expert witness has an overriding duty to assist the Court on matters relevant to the expert's area of expertise.

1.2 An expert witness is not an advocate for a party even when giving testimony that is necessarily evaluative rather than inferential.

1.3 An expert witness's paramount duty is to the Court and not to the person retaining the expert.

[57] As to the distinction between expert opinion evidence and expert assistance see *Evans Deakin Pty Ltd v Sebel Furniture Ltd* [2003] FCA 171 per Allsop J at [676].
[58] The "*Ikarian Reefer*" (1993) 20 FSR 563 at 565-566.

2. The Form of the Expert's Report[59]

2.1 An expert's written report must comply with Rule 23.13 and therefore must

(a) be signed by the expert who prepared the report; and

(b) contain an acknowledgement at the beginning of the report that the expert has read, understood and complied with the Practice Note; and

(c) contain particulars of the training, study or experience by which the expert has acquired specialised knowledge; and

(d) identify the questions that the expert was asked to address; and

(e) set out separately each of the factual findings or assumptions on which the expert's opinion is based; and

(f) set out separately from the factual findings or assumptions each of the expert's opinions; and

(g) set out the reasons for each of the expert's opinions; and

(ga) contain an acknowledgment that the expert's opinions are based wholly or substantially on the specialised knowledge mentioned in paragraph (c) above[60]; and

(h) comply with the Practice Note.

2.2 At the end of the report the expert should declare that "[the expert] has *made all the inquiries that* [the expert] *believes are desirable and appropriate and that no matters of significance that* [the expert] *regards as relevant have, to* [the expert's] *knowledge, been withheld from the Court*."

2.3 There should be included in or attached to the report the documents and other materials that the expert has been instructed to consider.

2.4 If, after exchange of reports or at any other stage, an expert witness changes the expert's opinion, having read another expert's report or for any other reason, the change should be communicated as soon as practicable (through the party's lawyers) to each party to whom the expert witness's report has been provided and, when appropriate, to the Court[61].

2.5 If an expert's opinion is not fully researched because the expert considers that insufficient data are available, or for any other reason, this must be stated with an indication that the opinion is no more than a provisional one. Where an expert witness who has prepared a report believes that it may be incomplete or inaccurate without some qualification, that qualification must be stated in the report.

2.6 The expert should make it clear if a particular question or issue falls outside the relevant field of expertise.

2.7 Where an expert's report refers to photographs, plans, calculations, analyses, measurements, survey reports or other extrinsic matter, these must be provided to the opposite party at the same time as the exchange of reports[62].

[59] Rule 23.13.
[60] See also *Dasreef Pty Limited v Nawaf Hawchar* [2011] HCA 21.
[61] The *"Ikarian Reefer"* [1993] 20 FSR 563 at 565
[62] The *"Ikarian Reefer"* [1993] 20 FSR 563 at 565-566. See also Ormrod *"Scientific Evidence in Court"* [1968] Crim LR 240

3. Experts' Conference

3.1 If experts retained by the parties meet at the direction of the Court, it would be improper for an expert to be given, or to accept, instructions not to reach agreement. If, at a meeting directed by the Court, the experts cannot reach agreement about matters of expert opinion, they should specify their reasons for being unable to do so.

J L B ALLSOP
Chief Justice
4 June 2013

Appendix C: Curriculum Vitae of Dr Rohan Sadler

Rohan Sadler

Curriculum Vitae

Profile

Rohan is a professional statistician who is involved in data science, remote sensing, and resource economics with a broad range of clients. With a strong background in the agricultural and environmental domains he has been developing the ecoinformatics capacity of organisations to deliver workflow improvement, data governance, analytics and evidence-based evaluation of management effectiveness.

Education

2006 **PhD**, *The University of Western Australia*, Perth.
Image-based Modelling of Pattern Dynamics in a Semiarid Grassland of the Pilbara, Australia

1993 **B.Sc.Agric.**, *The University of Western Australia*, Perth.

2014- **Diploma of Information Technology**, *TAFE NSW*, Online.

Experience

2016- **Director, Data Scientist**, *Pink Lake Analytics*, Perth.

- Developing leakage survey sample size calculator for Torres Strait biosecurity (Department of Agriculture and Water Resources, ACT)
- Modelling goat production in Australian rangelands (Ausvet, Western Australia)
- Remotely sensed land use and land cover classification and change within an urban municipality (Emerge Associates, Western Australia)
- Prediction of regional milk production for aggregation at the processor plant (milkflow.io, Sydney)
- Population density estimation of an island gecko species (Range to Reef, Western Australia).
- Advice on Estimation of the Market Risk Premium (Economic Regulatory Authority Western Australia, Western Australia).
- Phenotypic factors in germination responses of species suitable for mine-site

restoration (Botanic Gardens and Parks Authority, Western Australia).
- Water potential profiles of native seed germination success (Botanic Gardens and Parks Authority, Western Australia).
- Statistical Advice to the ERA on DBP Submission 56 (Economic Regulatory Authority Western Australia, Western Australia).
- Cost-response and power analysis in BACI-type experimental designs (BMT Oceanica, Western Australia).

2018- **Statistician**, *Ausvet*, Fremantle.
- Risk factors and antibiotic usage in the south American salmon industry.
- Spatial relative risk mapping for foot-and-mouth preparedness in Australia.
- Modelling shipboard mortality of livestock.

2015–2017 **Free Lance Data Scientist**, *Bush Futures*, Perth.
- Estimation of theta in the return on equity (Economic Regulatory Authority Western Australia, Western Australia).
- Empirical testing of theoretical capital asset pricing models and portfolio optimisation (Economic Regulatory Authority Western Australia, Western Australia).
- Cleaning, shaping, databasing and analysis of 30+ years of mammal trapping data for the Otways Region (subcontracted through Barbara Wilson on behalf of Department of Environment, Land, Water and Planning, Victoria).
- Heat mapping of availability of mental health services in Perth (Ray Dunne Public Relations, Western Australia).

2012-2015 **Senior Scientist**, *Astron*, Perth.
- Built Astron's remote sensing capacity and team, spanning various platforms and sensors, including product development and delivering client projects both in and outside of Australia.
- Innovated lidar assessments of landform change, and multispectral assessments of vegetation impacts of altered surface water flows and groundwater abstraction for WA's resource industry.
- Initiated data governance and workflow development within Astron.
- Data Team Leader (Emergency Oil Spill Response for various Oil and Gas clients).
- Statistical project support and population modelling for various clients.

2010-2012 **Research Assistant Professor**, *The University of Western Australia*, Perth.
Cooperative Research Centre for Plant Biosecurity
- Research and development evaluation
- Pest Management Area strategy optimisation

2007-2009 **Post-Doc**, *The University of Western Australia*, Perth.
Design of conservation contracts (DAFF, Market Based Instruments)
Fire behaviour in rehabilitated open forest (ARC Linkage with Worsley Alumina).

2005-2010 **Casual Lecturing and Tutoring**, *The University of Western Australia*, Perth.
Statistics, Decision Tools, GIS

Postgraduate Supervision

2014- **Thayse Nery de Figueiredo**, *PhD Thesis*, UWA, complete.
Optimal land-use change to increase water quality, quantity and biodiversity outcomes.

2014- **Maria Solis Aulestia**, *PhD Thesis*, UWA, in progress.
Land use dynamics in the Chure region of Nepal.

2012 **Hoda Abougamous**, *PhD Thesis*, UWA, complete.
An economic analysis of surveillance and quality assurance as strategies to maintain grain market access.

2011 **Bernard Phillimon**, *Masters Thesis*, UWA, complete.
Assessment of bushfire risk through remote sensing.

Professional Affiliations

Accredited Statistician (AStat), Statistical Society of Australia.

Adjunct Senior Lecturer, School of Agricultural and Resource Economics, The University of Western Australia.

Member, The Institute of Analytics Professionals of Australia (IAPA).

Professional Contributions

2014 **Member**, Statistical Society of Australia
Training Committee, National Branch.

2010 **Chairman**, Statistical Society of Australia
Branch Committee, Western Australia.

2008-2009 **Member**, Statistical Society of Australia
Branch Committee, Western Australia.

Awards

2013 **Innovation Award**, Astron Environmental Services.
2012 **Best Paper**, Australian Journal of Agricultural and Resource Economics

Key Projects

Environmental Policy.

- Agent-based modelling of saline water table management (DAFF)
- Agricultural Land Retirement as an Environmental Policy (LWA)
- Auctions for Landscape Recovery Under Uncertainty (DAFF)

Pest Management.

- Sample size determination for biosecurity monitoring in the Torres Strait (DAWR)
- Optimal Investment in Research and Development for Plant Biosecurity (CRC Biosecurity)
- Long Term Weed Management on Barrow Island (Gorgon)
- Leggadina and Mus Population Dynamics on Thevenard Island (Chevron)

Data Management.

- Otways Long Term Fauna Trapping Data (Parks Victoria)
- Scientific Monitoring for Oil Spill Response (Apache, ROC, VOGA)
- Data Governance: Strategy, Policy and Standards (Astron)
- Optimal Seed Farm Design (BGPA, Saudi Arabia)

Fauna Monitoring.

- Thevenard Island Mouse (Chevron)
- Northern Quoll (Polaris)
- Macropod Population Viability Analysis (Gorgon)

Remote Sensing.

- Remote Sensing of Pre- and Post-Fuel Loads (Worsley)
- Landform Change Detection (Gorgon)
- Vegetation Impacts of Seismic Surveys (Gorgon)
- Vegetation Mapping (RTTI, India)
- Groundwater Drawdown Impacts on Vegetation (BHPBIO)
- Surface Water Flow Impacts on Vegetation (FMG)

Key Products

ePower Toolbox, *BMT Oceanica, Australian Institute of Marine Science, QUT*. Provides power analysis and cost-response curves for the optimal design of beyond BACI (before-after-control-impact) studies.

Landform Change Analysis, *Astron*.

Provides an error budget for identification of statistically significant areas of landform change from LiDAR and photogrammetric DEM (digital elevation model) change assessment.

Vegetation Impacts of Groundwater and Surface Flow Alteration, *Astron*.

Identifies vegetation areas at greatest impact of groundwater drawdown or surface flow modification, as observed from time series of remote-sensed imagery.

Peer Reviewed Publications

Muhammad K Bashir, Steven Schilizzi, Rohan Sadler and Ghaffar Ali, *Vulnerability to food insecurity in rural Punjab, Pakistan*, British Food Journal, *in press.*

Matthias M Boer, Paul Johnston, and Rohan J Sadler, *Neighbourhood rules make or break spatial scale invariance in a classic model of contagious disturbance*, Ecological Complexity **8** (2011), no. 4, 347–356.

Matthias M Boer, Craig Macfarlane, Jaymie Norris, Rohan J Sadler, Jeremy Wallace, and Pauline F Grierson, *Mapping burned areas and burn severity patterns in SW Australian eucalypt forest using remotely-sensed changes in leaf area index*, Remote Sensing of Environment **112** (2008), no. 12, 4358–4369.

Matthias M Boer, Rohan J Sadler, Ross A Bradstock, A Malcolm Gill, and Pauline F

Grierson, *Spatial scale invariance of southern Australian forest fires mirrors the scaling behaviour of fire-driving weather events*, Landscape Ecology **23** (2008), no. 8, 899–913.

Matthias M Boer, Rohan J Sadler, Roy S Wittkuhn, Lachlan McCaw, and Pauline F Grierson, *Long-term impacts of prescribed burning on regional extent and incidence of wildfires—evidence from 50 years of active fire management in SW Australian forests*, Forest Ecology and Management **259** (2009), no. 1, 132–142.

Kerryn A Chia, John M Koch, Rohan J Sadler, and Shane R Turner, *Developmental phenology of Persoonia longifolia (Proteaceae) and the impact of fire on these events*, Australian Journal of Botany **63** (2015), no. 5, 415–425.

_____ , *Re-establishing the mid-storey tree Persoonia longifolia (Proteaceae) in restored forest following bauxite mining in southern Western Australia*, Ecological Research **31** (2016), no. 5, 627-638.

Kerryn A Chia, Rohan J Sadler, Shane R Turner, and Carol C Baskin*, Identification of the seasonal conditions required for dormancy break of Persoonia longifolia (Protecaeae), a species with a woody indehiscent endocarp*, Annals of Botany **118** (2016), no. 2, 331-346.

Veronique Florec, Rohan J Sadler, Ben White, and Bernie C Dominiak, *Choosing the battles: The economics of area wide pest management for Queensland fruit fly*, Food Policy **38** (2013), 203–213.

James J Fogarty and Rohan Sadler, *To save or savour: A review of approaches for measuring wine as an investment*, Journal of Wine Economics **9** (2014), no. 03, 225–48.

Aaron D Gove, Rohan Sadler, Mamoru Matsuki, Robert Archibald, Stuart Pearse, and Mark Garkaklis, *Control charts for improved decisions in environmental management: a case study of catchment water supply in south-west Western Australia*, Ecological Management & Restoration **14** (2013), no. 2, 127–134.

Hoda R Abougamous, Benedict White, and Rohan Sadler, *Contracts for grain biosecurity and grain quality*, Journal of Development and Agricultural Economics **9** (2017), no. 3, pp. 57-65.

Hoda R Abougamos, Rohan Sadler, and Benedict White, *Managing evolving insecticide resistance in stored grain pests within Avon Region*, Western Australia, Journal of Stored Products and Postharvest Research **8** (2017), no. 2, pp. 16-30.

Gavan S McGrath, Rohan Sadler, Kevin Fleming, Paul Tregoning, Christoph Hinz, and Erik J Veneklaas, *Tropical cyclones and the ecohydrology of Australia's recent continental-scale drought*, Geophysical Research Letters **39** (2012), no. 3.

Ram Pandit, Maksym Polyakov, and Rohan Sadler, *Valuing public and private urban tree canopy cover*, Australian Journal of Agricultural and Resource Economics **58** (2014), no. 3, 453–470.

Hazel R Parry, Rohan J Sadler, and Darren J Kriticos, *Practical guidelines for modelling post-entry spread in invasion ecology: Advancing risk assessment models to address climate change, economics and uncertainty*, NeoBiota **18** (2013), 41–66.

Deanna P Rokich, Jack Harma, Shane R Turner, Rohan J Sadler, and Ben H Tan, *Fluazifop-p-butyl herbicide: Implications for germination, emergence and growth of Australian plant species*, Biological Conservation **142** (2009), no. 4, 850–869.

Rohan J Sadler, Veronique Florec, Ben White, and Bernie C Dominiak, *Calibrating a jump-diffusion model of an endemic invasive: Metamodels, statistics and qfly*, 19th International Congress on Modelling and Simulation, Perth, Australia, 2011, pp. 12–16.

Rohan J Sadler, Martin Hazelton, Matthias M Boer, and Pauline F Grierson, *Deriving state-and-transition models from an image series of grassland pattern dynamics*, Ecological Modelling **221** (2010), no. 3, 433–444.

Rohan J Sadler, Douglas B Purser, and Susan Baker, *Hay quality and intake by dairy cows 2. Predicting feed intake with consumer demand models*, Animal Production Science (2018), no. 4, 730-743.

Grezegorz Skrzypek, Rohan J Sadler, and Andrzej Wiśniewski, *Reassessment of recommendations for processing mammal phostphate δ18O data for paleotemperature reconstruction*, Palaeogeography, Palaeoclimatology, Palaeoecology **446** (2016), 162–67.

Grzegorz Skrzypek and Rohan Sadler, *A strategy for selection of reference materials in stable oxygen isotope analyses of solid materials*, Rapid Communications in Mass Spectrometry **25** (2011), no. 11, 1625.

Grzegorz Skrzypek, Rohan Sadler, and Debajyoti Paul, *Error propagation in normalization of stable isotope data: a Monte Carlo analysis*, Rapid Communications in Mass Spectrometry **24** (2010), no. 18, 2697–2705.

Thayse Nery, Rohan Sadler, Maria Solis Aulestia, Ben White and Maksym Polyakov, *Discriminating Native and Plantation Forests in a Landsat Time-Series for Land Use Policy Design*, International Journal of Remote Sensing*, in press*.

Ben White and Rohan Sadler, *Optimal conservation investment for a biodiversity-rich agricultural landscape*, Australian Journal of Agricultural and Resource Economics **56** (2012), no. 1, 1–21.