Scottish Government
Riaghaltas na h-Alba
gov.scot

Scoping Study - Regional Population Viability Analysis for Key Bird Species CR/2016/16

Scottish Marine and Freshwater Science Vol 11 No 10

K Searle, A Butler, M Bogdanova and F Daunt

marinescotland

Scoping Study - Regional Population Viability Analysis for Key Bird Species CR/2016/16

Scottish Marine and Freshwater Science Vol 11 No 10

Kate Searle, Adam Butler, Maria Bogdanova and Francis Daunt

Published by Marine Scotland Science
ISSN: 2043-7722
DOI: 10.7489/12327-1

Marine Scotland is the directorate of the Scottish Government responsible for the integrated management of Scotland's seas. Marine Scotland Science (formerly Fisheries Research Services) provides expert scientific and technical advice on marine and fisheries issues. Scottish Marine and Freshwater Science is a series of reports that publishes results of research and monitoring carried out by Marine Scotland Science. It also publishes the results of marine and freshwater scientific work that has been carried out for Marine Scotland under external commission. These reports are not subject to formal external peer-review.

This report presents the results of marine and freshwater scientific work carried out for Marine Scotland under external commission.

Scoping Study - Regional Population Viability Analysis for Key Bird Species CR/2016/16

January 2020 – Final report

Funded by the Scottish Government's Contract Research Fund

Kate Searle[1], Adam Butler[2] Maria Bogdanova[1] and Francis Daunt[1]

[1]Centre for Ecology and Hydrology

[2]Biomathematics and Statistics Scotland

Contents

Executive summary

1) The Scottish Government has set a target of 100% of Scottish demand for electricity to be met by renewable sources by 2020. The marine environment offers considerable potential with respect to harvesting renewable energy, through wind, wave and tidal stream energy generators. However, offshore renewable developments have the potential to impact on seabird populations.

2) Population Viability Analysis (PVA) is considered best practice in order to understand the population-level consequences of predicted impacts from renewable energy developments on seabirds. If inappropriate PVA methods are used, then there is a risk that assessments become protracted, standards of assessment vary, and decision-making uncertainty increases; this in turn can increase risk and reduce confidence that consent decisions are made in a timely manner. There is, therefore, considerable need to provide clear guidance on which PVA approaches should be used in each circumstance.

3) Within this project, we evaluated and compared the performance of a range of different modelling methods for PVAs that have been used in practice. We evaluated the performance of the methods in producing accurate predictions of future “baseline” abundance – i.e., abundance in the absence of an offshore wind farm. PVAs are typically summarised, in the context of offshore renewables, in terms of metrics that compare scenarios of impact against a “baseline” projection of abundance. Caution should, therefore, be taken in relating the results of our evaluation (which is concerned with absolute abundance) directly to the ability of methods to produce accurate values of PVA metrics (which are often concerned with comparing relative abundance under different scenarios). However, we nonetheless expect the results of our evaluation to provide a useful qualitative guide to the relative strengths and limitations of different methods.

4) At the national scale, we compared deterministic and stochastic Leslie matrix models (which are usually used in practice for PVAs) against each other, and against a range of simple time series growth models. We applied these methods to data on abundance, breeding success and survival for 15 seabird species, for breeding colonies throughout the British Isles. We evaluated performance by applying the methods to abundance and demographic data, with these split in to a training period and a subsequent “test” period, assessing whether the predictions that the methods generated for the “test” period were consistent with the observed counts of abundance for that period.

We considered four possible definitions of the test period – 1998-2017, 2003-2017, 2008-2017 and 2013-2017 – and in each case considered the training period to be all years (with suitable available population data) prior to this.

5) In one region (Forth/Tay), for five species, we also used a five year test period (2013-2017) to compared this suite of approaches against the Semi-Integrated Population Models (SIPMs) produced by Freeman *et al*. (2014) and Jitlal *et al*. (2017) (termed Bayesian State Space Models in those reports).

6) We assessed performance of methods by looking at: a) whether the method was possible to apply; b) how frequently it yielded "highly implausible" results (i.e., results that are more than 100 times larger or smaller than the actual abundance); c) whether it produced systematically biased results (i.e. over-estimated or under-estimated actual abundance); d) how much error the methods had, on average, in predicting the observed count; e) whether the method provided an accurate quantification of uncertainty; f) the level of uncertainty associated with each method and g) the computational time required to implement the method.

7) The results of the comparisons are inconsistent, suggesting there is no simple hierarchy of performance between different models – the results vary depending upon the species and colony being considered, and vary between the different possible criteria that can be used for assessing performance. Some consistent patterns do emerge, however. One key finding to emerge from our comparisons is that deterministic and stochastic Leslie matrix approaches frequently perform relatively poorly, compared to time series models (and, where evaluated, SIPMs), in terms of the accuracy of the predictions produced. This is likely to reflect known differences in data coverage in the UK, as well as inherent differences between the methods. Data on adult and immature survival rates are very limited, whereas data on abundance are much more widely available. This makes it is unsurprising that methods that use abundance data (time series models, SIPMs) frequently produce more accurate predictions of absolute abundance than methods that do not (Leslie matrix approaches).

8) The second key finding to emerge from our comparisons is that stochastic Leslie matrix approaches that only use demographic data systematically underestimate uncertainty. The frequency with which 95% confidence intervals contained observed counts was consistently much lower than 95% for these approaches. This is a more surprising finding, but we think this result

is likely to occur because the stochastic Leslie matrix models that are currently used in PVAs relating to seabirds make a biologically implausible assumption of independence of demographic rates (between different rates and different years), and also usually fail to account for uncertainty in rates, even when they do account for environmental variation in rates.

9) On the basis of these empirical results we recommend that the outputs obtained from Leslie matrix models that only use demographic data should be interpreted with caution. One useful method of improving the performance of Leslie matrix models in situations where abundance data exist is to validate predictions produced by the Leslie matrix approaches against these data. We recommend that "tuning" of this kind is undertaken, using automated approaches. We also recommend, on the basis of our results, that current estimates of uncertainty in absolute abundance from stochastic Leslie matrix models, should be regarded as underestimates, and interpreted accordingly. We believe that work is needed (a) to establish whether the underestimation of uncertainty in absolute abundance also leads to underestimation of uncertainty in PVA metrics, and (b) to improve the representation of uncertainty within Leslie matrix models used in seabird PVAs by quantifying and exploiting correlations within and between demographic rates.

10) We conclude by providing general guidelines on the choice of methodology to use when performing PVAs. These guidelines are based on empirical comparisons undertaken within this project, supplemented by existing knowledge and expert judgement. We generally recommend the use of integrated or semi-integrated population models (IPMs and SIPMs) for PVAs, in situations where sufficient abundance data are available. We note, however, that this recommendation is largely based on existing knowledge (e.g. the arguments outlined in Freeman *et al*., 2014), rather than the results of the comparisons in this project, because only a very limited evaluation of the empirical performance of SIPMs was possible within this project.

1. Introduction

The Scottish Government has set a target of 100% of Scottish demand for electricity to be met by renewable sources by 2020. The marine environment offers considerable potential to harvest renewable energy, through wind, wave and tidal stream energy converters. However, the Scottish Government has a duty to ensure that offshore renewable developments are achieved in a sustainable manner, by protecting the natural environment from adverse impacts in accordance with the requirements of the Marine Strategy Framework Directive (2008/56/EC), the Habitats Directive (92/43/EEC) and the Birds Directive (2009/147/EC).

Offshore renewable energy developments have the potential to impact on seabird populations that are protected by the EC Birds and Habitats Directives, notably from collisions with turbine blades and through displacement from important habitat (Drewitt & Langston 2006; Larsen & Guillemette 2007; Masden *et al.* 2010; Grecian *et al.* 2010; Langston *et al.* 2011; Scottish Government 2011). Other factors of concern are barrier effects to the movement of migrating or commuting birds, direct habitat loss, toxic and non-toxic contamination and negative effects of developments on the distribution and abundance of prey. These potential effects are predicted to be important for breeding seabirds that unlike at other times of the year, or for pre-breeding age classes, are constrained to obtain food within a certain distance from the breeding colony (Daunt *et al.* 2002; Enstipp *et al.* 2006).

Population Viability Analysis (PVA) is considered best practice in order to understand the population-level consequences of predicted effects of renewable energy developments on seabirds. This is because it provides a robust framework that uses demographic rates to forecast future population levels, either under baseline conditions or under scenarios of change resulting from, for example, an offshore energy development (Maclean *et al.* 2007; Freeman *et al.* 2014). PVAs essentially employ mathematical and statistical population models to forecast future population change, and can be undertaken using methods of varying complexity. Within the PVA framework, population models are used to forecast into the future under so-called ‘baseline’ conditions with no impact present, and under ‘scenario’ conditions where an impact is applied to one or more of a set of demographic rates (e.g. survival rate). Comparisons are then made using a range of PVA metrics to assess differences between the baseline and scenario population trajectories. However, criticism has been levied about how the results of such PVAs can be difficult to understand, assess and interpret by stakeholders (Knight *et al.* 2008; Pe'er *et al.* 2013). Moreover, due to uncertainty and variability amongst the input parameters for the population models underpinning PVAs, decision makers may lack confidence in,

and may misinterpret, predictions (Addison *et al.* 2013; Green *et al.* 2016). Thus, it is critically important that steps are made to solve these challenges where possible (Masden *et al.* 2015; Green *et al.* 2016), because PVAs remain one of the most widely used tools for evaluating the predicted impacts of anthropogenic developments, wildlife management or conservation strategies on focal populations.

PVAs are required with increasing frequency in assessments of renewable developments on seabirds, and are also increasingly applied in assessments for the potential cumulative effects arising when a number of licenced activities are considered in combination. This may require the use of different population modelling approaches within PVAs to best address the range of circumstances such cumulative assessments contain; such as small or large effect sizes, sparse or high resolution spatial and temporal data, and varying degrees of data quality for different populations and species. Furthermore, some population modelling approaches are more labour intensive and therefore expensive than others, and there may be a lack of capacity in the UK community of environmental scientists to undertake them. There is, therefore, considerable need to provide clear guidance on which PVA approaches should be used in each circumstance – key elements of this are the form of the population model used to generate population forecasts, and the best framework for sourcing key model parameters such as demographic rates, where these are lacking for a site/population of interest. The choices of both (the population model, and the model parameter values) will be dependent on the quality of the data and the region involved. Without this clarity, there is a risk that assessments become protracted, standards of assessment vary, and decision-making uncertainty increases. This in turn can increase risk and reduce confidence that a consent decision will be made in a timely manner.

Several reviews of appropriate Population Viability Analysis (PVA) model structure and parameter specification for seabirds have been conducted (e.g. Maclean *et al.* 2007; Cook & Robinson 2010; Freeman *et al.* 2014; Horswill & Robinson 2015; Trinder & Furness 2015; Cook & Robinson 2016). These reviews revealed a wide range of methods, where the appropriate PVA model structure was primarily determined by the specific question, the life history characteristics of the species, and the availability of data at the colony of interest. However, the principal structure of the majority of PVAs used in the context of seabirds and marine renewables employed a matrix population model (Caswell 2001). In these models, the population abundance at a point in time is estimated by the abundance in the previous time step, subject to a set of equations governing the form and parameterisation of population demographic processes and their relationship to other ecological processes such as density dependence, immigration and emigration, and

environmental stochasticity. Typically, model population processes are described as discrete, sequential events using matrix model algebra (Caswell 2001).

In the UK, a wide number of Environmental Statements have used PVAs to assess the impacts of wind farm developments on seabird populations and to inform the consenting process for approval of these developments (Freeman *et al.* 2014; Cook & Robinson 2016). It should be noted that details of PVAs for evaluating the impacts of wind farms are largely available through so called “grey literature” (reports and assessments) rather than ISI published papers. Cook & Robinson (2016) reviewed 27 proposed sites at which the predicted population level impacts of offshore wind farms on seabirds had been considered during assessment: Aberdeen Offshore Wind Farm, Beatrice, Burbo Bank Extension, Docking Shoal, Dogger Bank Creyke Beck A, Dogger Bank Creyke Beck B, Dogger Bank Teesside A, Dogger Bank Teesside B, Dudgeon, East Anglia One, Fife Wind Energy Park, Galloper, Hornsea Project One, Inch Cape, London Array Phase II, MORL (MacColl, Stevenson, Telford), Navitus Bay, Neart na Gaoithe, Race Bank, Rampion, Seagreen Alpha, Seagreen Bravo, Triton Knoll 3, Walney I & Walney Extension (references in Cook & Robinson 2016). In a recent project (Jitlal *et al.* 2017), we synthesised these studies and a further eight reports that used PVA in this context (MacKenzie & Perrow 2009; 2011; Inch Cape Offshore Limited 2011; JNCC & NE 2012; Moray Offshore Renewables Ltd 2013; Freeman *et al.* 2014; Trinder 2014; 2015).

PVAs have aimed to either compare the predicted population trajectory into the future with the wind farm development present to that predicted without the development, or to quantify the risk the development poses by estimating probability of future population declines. Both deterministic and stochastic population models have been used for evaluating the impacts of wind farms, and it has been argued that deterministic models are a more “honest” approach where there is significant uncertainty around demographic parameters because the presented confidence limits from stochastic models may be taken to imply an unjustified level of precision in the underlying data (WWT 2012). However, importantly, deterministic models do not produce a distribution of results and hence cannot employ probabilistic metrics, and so long as the sources of uncertainty included in stochastic PVAs have been clearly articulated, it may be argued they are a better representation of potential future trajectories for population abundance.

A number of different metrics from PVAs, for example the increase in the probability of a population decreasing by a fixed amount over time, have been used to provide assessments of the predicted impact of wind farms on seabird populations. Metrics have been criticised for being sensitive to uncertainties both in the life-history

parameters used to parameterise the models and in the size of the predicted impact of wind farms on the population (Masden *et al.* 2015; Green *et al.* 2016). Uncertainty in the demographic rates used to parameterise models can lead to uncertainty in whether the predicted magnitude of the impact (e.g., increased mortality or reduced productivity) will lead to an adverse effect on the focal population size (Masden *et al.* 2015). Uncertainty in the size of the impact of wind farms on populations arises due to lack of empirical data on collision mortality, displacement or barrier effects on seabird populations. Thus, there is concern that the metrics may not enable sufficiently accurate predictions and good understanding of the predicted impacts of offshore wind farms on seabird populations (Green *et al.* 2016). Recent sensitivity analyses, conducted using simulation approaches (Cook & Robinson 2016; 2017) and real-world data (Jitlal *et al.* 2017), have demonstrated that ratio PVA metrics (e.g. ratio of impacted to unimpacted median population size) are markedly less sensitive to mis-specification in input parameters than probabilistic PVA metrics.

In this project, we performed an initial review of the grey and peer-reviewed literature on the use of PVAs, and identified ten methods that have been used to perform PVAs in seabirds and other species. In the grey literature, we identified ten reports from 2009 onwards (Appendix A). The vast majority used stochastic stage structured matrix models for the PVA, the only exception being an individual based method within this framework for Sandwich terns (MacKenzie & Perrow. 2009; 2011). There was variation in the extent to which methods included density-dependent processes, stochasticity in demography and the environment, and immigration from other colonies. Most methods sampled from beta distributions to set survival and productivity rates with environmental stochasticity (sampling to derive year-specific vital rates to populate the matrix in each projection), and used binomial distributions to deal with demographic stochasticity. All methods required age and stage specific estimates for vital rates (typically survival and reproduction), age at first breeding, and proportion of adults that breed and initial counts. These were most often taken from published reports (mean and SE). Almost all methods assumed closed populations.

Our synthesis of peer-reviewed literature (conducted using two searches in Web Of Science (“population viability analysis and seabirds” – all years; “animal population viability analysis” –2010 inclusive to present) identified 22 relevant peer-reviewed publications covering a wide range of taxa (mostly birds and mammals; Appendix A). Again, the majority of studies used various formulations of stochastic matrix population models (14/22; with one of these being an individual based stochastic simulation model).

Based on these reviews, we selected ten different PVA methodologies to apply and test within this project (see Section 3). We focused on the ability of these methodologies to accurately predict observed counts of abundance for 15 seabird species (northern gannet, fulmar, great cormorant, European shag, Arctic skua, black-legged kittiwake, herring gull, lesser black-backed gull, great black-backed gull, common tern, Sandwich tern, little tern, common guillemot, razorbill and Atlantic puffin), from the set of breeding colonies within the British Isles for which the Seabird Monitoring Programme (SMP) data met a set of minimum data requirements.

We focused upon using population models to predict observed counts (rather than, for example, upon PVA metrics concerning the predicted impact of offshore renewables) because empirical data on abundance are collected directly and are readily available. Therefore, in this sense, the project was more accurately tasked with assessing the performance of alternative baseline population models commonly used to underpin PVAs. A direct empirical evaluation of the performance of PVA methods in accurately predicting PVA metrics is difficult, if not impossible, because PVA metrics typically compare two scenarios (impacted and baseline), of which only one can actually occur.

Caution should be taken in relating the results of our evaluation to the performance of methods in predicting PVA metrics, rather than absolute abundance. It is important to note that the magnitudes of uncertainty in absolute abundance will often be very different to those for PVA metrics - e.g. we would generally expect substantially lower levels of uncertainty in ratio-based PVA metrics than in absolute abundance. We expect, however, that there is likely to be a qualitative link between performance of methods in accurately predicting absolute abundance and performance of methods in accurately predicting PVA metrics. We might reasonably expect that, in most situations, methods that do relatively poorly in predicting absolute abundance are also likely to do relatively poorly in predicting PVA metrics, and that methods that underestimate uncertainty in absolute abundance will also underestimate uncertainty in PVA metrics.

We applied each population modelling method to data for a “training” period and used the method to generate predictions for a “test” period – we then compared the modelled results against the actual counts observed within the test period. We considered different possible splits between the training and test period: specifically, we used four possible definitions of test period (from 2013-2017, 2008-2017, 2003-2017 or 1998-2017), and, in each case, defined the training period to be all years with data prior to this. We assessed the ability of the population models to predict the observed counts using a range of criteria: accuracy and bias, the number of

situations in which it is possible to deploy each method (where minimum data requirements are met), accurate quantification of uncertainty, and the computational time required to implement each method. We applied methods to all species, years and colonies for which the minimum data requirements were met, and assessed performance using all species-colony-year combinations within the test period for which an observed count was available. Note that the minimum data requirements were applied automatically, and that the "colony" definitions we used were based directly on the lowest reporting level in the SMP; this means that some key special protection areas (SPAs) will have been partially or completely excluded from our evaluation. For kittiwakes at the Flamborough and Filey Coast SPA, for example, only one of the smaller SMP count units, at Cayton Bay, has been included, whilst the main Flamborough and Bempton colony has been excluded. The need to use automated rules for data selection was unavoidable, given the large number of species, populations and methods being considered within this project. However, the exclusion of important colonies does mean that the results should be interpreted cautiously. The detailed results of our evaluations are available as a CSV file, and this allows readers to check which colonies have been excluded for each species.

In summary, this project addressed key concerns by testing a range of PVA modelling approaches (population models) across a number of seabird species, data qualities and regional scales to establish the most appropriate method to use and under which circumstances. We provide recommendations to guide end users on how PVAs should (and should not) be produced, enabling PVA model development to be assessed for all relevant species and regions. This addresses the urgent need to determine the feasibility of PVAs that are flexible in application whilst also providing sufficient confidence that they perform appropriately for the population of interest.

2. Data

The UK has some of the best demographic data on seabird populations in the world. The central repository for population count and breeding success data is the Seabird Monitoring Programme (SMP) online database (http://jncc.defra.gov.uk/smp/). Other demographic data, such as adult survival rates and age at first breeding, have been published in the peer-review literature and reviewed in contract reports (e.g., Horswill & Robinson 2015). In this study we considered the 15 UK seabird species for which data coverage is most complete. For these species, and others, the largest amounts of data are typically available for population counts, followed by productivity, with adult survival having the most sparse data. A perennial concern is the paucity of data on adult survival, generally only available at a very small number of colonies (<5) for

a particular species. It should also be noted that the frequency with which count data are collected is often highly variable among colonies. For example, in the Forth/Tay region, some SPA populations are counted annually while others are only counted every few years (Freeman *et al.* 2014). This variation in coverage of count and demographic data underpins the rationale for this project.

2.1. Abundance data

Abundance counts of breeding birds were extracted from the SMP database, using the most up to date dataset available at the time of the request (January 2018). Counts were acquired for 15 seabird species spanning the 1960s to 2017: northern gannet, fulmar, great cormorant, European shag, Arctic skua, black-legged kittiwake, herring gull, lesser black-backed gull, great black-backed gull, common tern, Sandwich tern, little tern, common guillemot, razorbill and Atlantic puffin. Additional colony counts for northern gannets, spanning the period from the early 1900s to 2017, were collated by Prof. S. Wanless and included in the analysis.

Our main evaluation used SMP data at the lowest possible level of recording (e.g. SMP count units). This is because this is the level at which count data are collected, but it is important to note that this does not always correspond to a biologically distinct population or to the spatial level at which PVAs would be run in practice (since a breeding colony may contain multiple SMP count units).

A key objective of this work was to compare the results of running population models using a range of different methods against the results obtained using a previous application of Semi-integrated Population Models in the Forth-Tay (Freeman *et al.* 2014; termed Bayesian State Space Model in that report – see Section 3.4 for rationale for term change). It was not feasible, within the constraints of this project, to re-run that analysis, so, to ensure comparability of our results, we:

a) followed Freeman et al. (2014) in running these analyses at the SPA level, using data that have been aggregated to that level;
b) followed Freeman *et al.* (2014) in manually adding, removing or correcting a number of doubtful counts in the SMP for colonies within the four SPAs in the Forth-Tay region (St. Abb's Head to Fast Castle SPA, Forth Islands SPA, Fowlsheugh SPA, Buchan Ness to Collieston Coast SPA);
c) followed Freeman *et al.* (2014) in using linear regression to convert plot counts into estimated whole colony counts for four combinations of species and SPA (razorbill and common guillemot for St. Abb's Head to Fast Castle

SPA; common guillemot for Fowlsheugh SPA and Buchan Ness to Collieston Coast SPA).

Counts for most species relate to breeding pairs; for guillemot and razorbill, however, the counts relate to individuals, and a conversion factor (the “k-value”), needs to be applied in order to derive the number of pairs from the observed counts (Harris *et al.* 2005a and 2005b updated).

2.2. Survival data

Survival data were only available for a very limited set of colonies. Species level data for adult and immature survival in each of the 15 species (see Table 0-2) were, therefore, collated from the JNCC Report (No. 552) by Horswill & Robinson (2015). A single estimate of mean survival, and an associated standard deviation (representing inter-annual variability) were extracted for each species, and these were, wherever information was available, produced separately for adults and immatures.

It appears possible that the species-level standard deviations presented in Horswill & Robinson (2015) may in some cases be underestimates, because some of the species-level values have been derived from study-level standard errors (SEs), rather than standard deviations (SDs), and SEs (which represent uncertainty) will tend to be systematically lower than SDs (which represent variability). In order to err on the side of caution (e.g. conservatism), we, therefore, re-calculated the species-level SDs. We calculated them using the same basic approach as Horswill & Robinson (2015), but when study-level SEs rather than SDs were used within their calculations, we multiplied these by the square root of the number of years of data available for that study. This is a crude adjustment, derived under some strong and potentially unrealistic simplifying assumptions (i.e., the assumption that the SEs are derived from one data point per year, that years are independent, and that annual survival rates are normally distributed), but should nonetheless help to avoid any systematic under-estimation associated with using SDs in lieu of SEs.

Within the main evaluation of methods, the same survival rates were assumed to hold for all colonies, in the absence of more detailed data. Within our separate evaluation of methods in the Forth-Tay region, however, we do also consider the use of survival rates derived from local colony-specific data from the Isle of May and used in Jitlal *et al.* (2017), in addition to consider the species-level rates from Horswill & Robinson (2015).

2.3. Breeding success data

Breeding success data for each of the 15 species were extracted from the Seabird Monitoring Programme database (1986-2017). Data coverage for each species was generally considerably poorer than for abundance (Table 0-2), but much better than for survival.

2.4. Regional definitions

We selected a range of alternative regional definitions to compare the performance of results obtained using data derived from a range of different spatial scales. We considered ten alternative regional classifications over which abundance or breeding success data could be pooled to inform the demographic rates used with the alternative population models..

The first option is simply to apply population models locally – i.e., to use data from the focal colony only, without any regional pooling (we denote this option **R0**). The remaining nine classifications that we consider are:

R1: SMP regions (based roughly on administrative boundaries used for local government)

R2: ICES regions

R3: Regional Seas (JNCC)

R4: CRA (Cook & Robinson 'ecologically coherent' regions based on trends in abundance)

R5: CRB (Cook & Robinson 'ecologically coherent' regions based on trends in breeding success)

R6: MSFD regions

R7: OSPAR regions

R8: "Global" (i.e. regional pooling over all colonies within the British Isles)

R9: Forth-Tay SPAs (the four Forth-Tay SPAs considered in Freeman at al. 2014)

The number of regions within each classification is given in Table 0-3. Examples of the regional classifications are shown in Figure 1.

“Local” and “Global” pooling were included in order to capture extreme cases (no spatial pooling whatsoever, or pooling data across the entire British Isles). The “Forth-Tay SPAs” were included to allow comparison of our results with those of Freeman *et al.* (2014), because this was the spatial scale at which their analysis was conducted and reported. OSPAR, MSFD, ICES, SMP and Regional Seas classifications were considered because these are widely used for the purposes of reporting.

The remaining two approaches, “CRA” and “CRB”, were designed to provide more ecologically-motivated regional classifications. They are based on the “overall” regions that Cook & Robinson (2010) derived in relation to trends in abundance (Table 4.1 in their report) and breeding success (Table 4.2 in their report). Cook & Robinson (2010) also derived species-specific regions, but we did not consider those because not all colonies were allocated to regions within these, and because species-specific regions were not developed for all of the species included in this study. Examples of the regional classifications are shown in Figure 2.4. Note that the CRA and CRB regions are based on clusters of colonies with similar abundance or breeding success trends, and are not reflective of the at-sea regions used by individuals from different colonies.

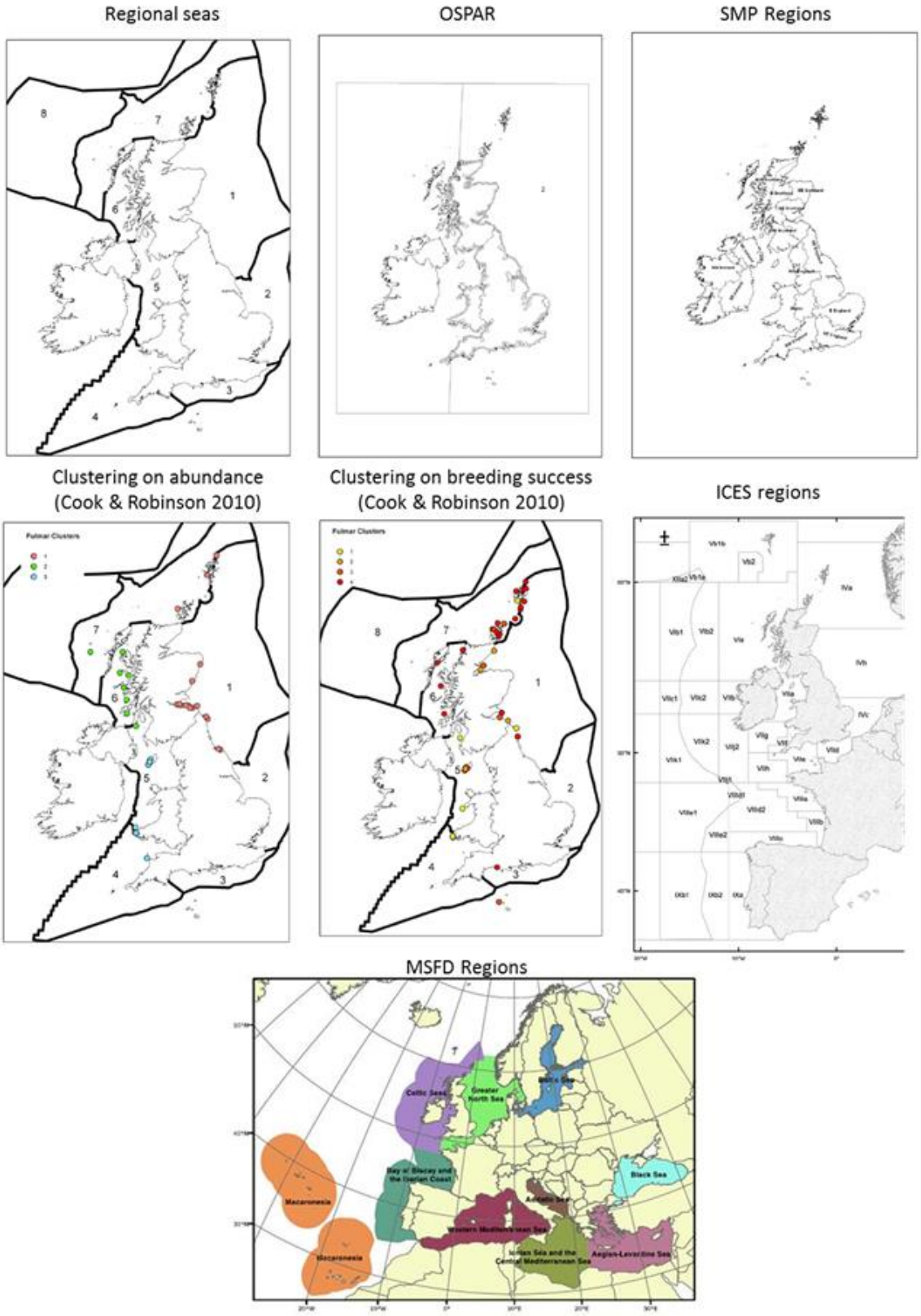


Figure 1: Regions used for data pooling within population modelling methods. Note that examples for CRA (Cook & Robinson 'ecologically coherent' regions based on trends in abundance) and CRB (Cook & Robinson 'ecologically coherent' regions based on trends in breeding are shown only for fulmars. See Cook & Robinson (2010) for full report.

3. Statistical methods for population models used in PVA

We considered a range of different statistical methods for generating population forecasts within PVAs. Each population modelling method is designed to provide predictions for abundance in future years, and each of the methods has a similar basic structure:

1. Use a pooling region classification to decide which input data to use in deriving the parameters of the population model (e.g., abundance , breeding success etc.) for the colony or region of interest;
2. Estimate or derive the values of the parameters of the population model from empirical data for the training period derived from the region in step (1);
3. Find the most recent count for the colony or region of interest;
4. Use (1), (2) and (3) to define the input values for each population model; and
5. Generate predictions - either deterministically, or by simulation, depending on the method.

Population models can either be used to generate predictions for individual colonies, or for wider regions.

3.1. Pooling and reporting regions

It is important to note that we use the term “regions” in two different ways within the context of a PVA:

(a) To specify the level at which outputs from the PVA are produced (“**reporting regions**”); and
(b) In order to determine the input data that will be used in determining the parameter estimates of the PVA model (“**pooling regions**”).

The choice of reporting region classification is related to the objectives of the PVA, i.e. at what spatial level are users interested in running the PVA? In practice, a range of reporting region scales may be of interest when running PVAs, ranging from the SPA level up to the scale of very large spatial regions (e.g., the entire North Sea). For this project, however, our focus is solely upon the empirical assessment of performance, so we focus only on spatial scales at which performance can be assessed empirically in a defensible way. A defensible assessment of performance for a spatial region is really only possible when a count of abundance for the entire region is available for at least one year within the test period, so that there are observed values of abundance in the test period to compare the PVA projections

against. For larger scale region classifications (e.g. OSPAR, ICES, Regional Seas, SMP), exploratory analyses showed that this condition was almost always not met, because in any particular year data are almost always missing for at least some of the colonies within each region. We, therefore, focus in this project solely upon assessing the performance of PVAs that have been run for individual SMP count units (henceforth referred to for convenience as “colonies”, although it is important to note that they do not always correspond to biologically distinct breeding colonies). The only exception is for our separate evaluation of performance in the Forth-Tay region, where we follow Freeman *et al*. (2014) in running PVAs at the SPA level (for four SPAs), using annual counts of abundance that have been summed to the SPA level.

The choice of pooling region classification, by contrast, is a methodological one, i.e., which level of spatial pooling of the PVA inputs produces the most accurate predictions of abundance, at the level of an individual site? In this study, we considered ten different potential regional classifications for use as “pooling regions” (Section 2.4), and compared the performance associated with these different classifications.

3.2. Leslie matrix models

An initial literature review (see Introduction) showed that most PVAs, in both the academic literature and the grey literature, use approaches that are based on Leslie matrices. These used estimated demographic rates – survival and productivity – to generate projections of future population size.

3.2.1. Overview

We focused upon the two main implementations of Leslie matrix approaches to PVAs – a “deterministic” approach and a “stochastic” approach. The stochastic approach is designed to incorporate the effects of demographic and environmental heterogeneity, which induce inter-annual variation in demographic rates. The stochastic approach is arguably more biologically realistic than the deterministic approach, but we considered both because (a) the stochastic approach is more computationally intensive than the deterministic approach and (b) the stochastic approach depends upon having a reliable estimate of the inter-annual standard deviations associated with the demographic rates; and such values are not always available in practice. In both cases, we consider both “local” (i.e., colony level) and “regional” variants of the PVA approach.

The basic inputs to the Leslie matrix PVA approaches that were considered are summarised in

Table *0-4*. These include demographic rates for breeding success (I1) and survival (I2); life history traits (age at first breeding, I3); and population abundance (initial count, I4). A key point is that Leslie matrix approaches only make use of a single count of abundance to initialise the model projection – and subsequent projections are primarily generated using data relating directly to the demographic rates.

The Leslie matrix approaches then involve two main stages – "estimation" and "simulation".

3.2.2. Estimation step

The estimation stage is common to both deterministic and stochastic approaches. It involves three steps (E1, E2 and E3):

	Input	Description	Value determined by
I1	Breeding success	Mean, and, for stochastic version only, SD	Species & pooling region
I2	Survival	Mean, and, for stochastic version only, SD	Species & age class
I3	Age at first breeding	Value	Species
I4	Initial count	Value, year associated with the value	Species & target colony

E1: Extract the mean and SD of breeding success to use within the PVA. For "local" variants of PVA these values are simply taken to be the mean and (inter-annual) standard deviation of breeding success within the target colony (i.e., the colony of interest). For "regional" variants of PVA these values are derived by identifying the mean and SD values for each colony within the pooling region, and then following the recommended procedure in Horswill & Robinson (2015). This procedure averages across breeding successes reported for each constituent colony weighted by the colony size to produce a mean breeding success for the region. Similarly, to generate the SD for breeding success over the region, this method either takes the average over reported breeding success SDs for all constituent colonies, or the SD of the means reported for each colony, and uses whichever of these two values results in the largest SD estimate.

E2: The mean breeding success from stage E1 is then used, in conjunction with the mean survival value (I2) and age at first breeding (I3) to calculate the annual survival

rate s_a and productivity rate r_a for individuals of age a. These rates are, in turn, used to construct the "deterministic" version of the Leslie matrix, $\boldsymbol{L}$, which is of the form:

$$\boldsymbol{L} = \begin{bmatrix} r_1 & r_2 & r_3 & \dots & r_{A-1} & r_A \\ s_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & s_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & s_{A-2} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{A-1} & 0 \end{bmatrix}$$

(where A denotes the maximum age of the species).

E3: The stable age structure is calculated from the Leslie matrix constructed in Step E2 using the function *stable.stage* from the *popbio* package (Stubben & Milligan 2007) within R; this gives the proportion of birds lying in each age class. Technically speaking, the stable age structure is calculated by:

(a) deriving the eigenvectors and eigenvalues of the Leslie matrix $\boldsymbol{L}$;
(b) finding the eigenvector associated with the largest eigenvalue; and
(c) rescaling this eigenvector so that the values sum to one (i.e., dividing each value within the eigenvector by the sum of all values).

3.2.3. Simulation step

S1: The total initial population size was first estimated by converting the most recent count from the target colony (I4) into an estimate of the number of breeding pairs (the approach taken depends on species; see Section 2.1). This was then multiplied by two (to convert from breeding pairs to breeding individuals), and subsequently divided by the proportion of the age structure that corresponds to breeding adults as derived from the stage age structure (E3) (to scale up from breeding adults to the whole population). The initial number of birds within each age class, $\boldsymbol{z}_1$, was then calculated by multiplying the whole population by the proportion of birds in this age class within the stable age structure.

S2: For the deterministic version of the PVA, the counts in subsequent years were calculated by propagating forwards using the deterministic Leslie matrix from Step E2; the counts in each age class in year t were calculated from those in year $t-1$ via the matrix calculation

$$\boldsymbol{z}_t = \boldsymbol{L}\boldsymbol{z}_{t-1}$$

For the stochastic version of the PVA this step is more involved:

S2a: use moment matching to find the parameters of a beta distribution that match the mean and SD of survival from I2 and the mean and SD of breeding success from E1.

S2b: simulate annual breeding success and survival rates using the beta distributions in Step S2a.

S2c: use the simulated demographic rates, from S2b, to generate a Leslie matrix for each year.

S2d: use the Leslie matrices from S2c and the initial counts from S1 to generate projected counts.

3.3. Time series models

The most widely used alternative to Leslie matrix approaches within the literature involves fitting non-linear population growth models to count data on abundance. Such models are very widely used within ecology, and there is a considerable literature regarding the relative advantages and disadvantages of each model.

3.3.1. Model types

If N_t denotes the number of birds at a particular colony in year t , then the most widely used growth models are all of the general form:

$$\log\left(\frac{N_t}{N_{t-1}}\right) = \alpha + \beta f(N_{t-1}) + \varepsilon_t$$

(**Equation 1**)

where the "process-error" term, ε_t, is assumed to be normally distributed:

$$\varepsilon_t \sim N\left(0, \sigma_p^2\right)$$

The parameter α represents the growth (or decay) rate of the species in the absence of any density dependent effects, whilst σ_p^2represents the level of process uncertainty.

The parameter term β represents the level of density dependence, and the function $f(N_{t-1})$ captures the form of the density dependence. We focus here upon three models for f, all of which are widely used in practice:

Simple growth model: $f(x) = 0$

Ricker model: $f(x) = x$

Gompertz model: $f(x) = \log(x)$

The first model (Dennis *et al.* 1991) simply omits the density dependence term, and provides a simple model for growth or decay in the absence of density dependence. The Ricker (1954) and Gompertz (Winsor 1932) models assume that density dependence has a particular parametric form.

3.3.2. Parameter estimation

Two main approaches for estimating the parameters of these models are commonly used:

1. transform the response and explanatory variables in such a way that the resulting model is linear, and then apply a standard simple linear regression; or
2. retain the response and explanatory variables on the original scale, and estimate the parameters via non-linear regression.

Within R (R core team 2019), these approaches will generally utilise either the **lm** or **nlme** functions (respectively).

More specifically, the models all have a log-likelihood function of the form:

$$l(\alpha, \beta, \sigma_p^2, \theta) = \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{\log(N_t) - \log(N_{t-1})\,\alpha - \beta f(N_{t-1}; \theta)}{\sigma_p^2}\right)^2$$

The values of the unknown parameters (α, σ_p^2, and, where relevant, β and θ) can be estimated through maximisation of this log-likelihood function. Numerical optimisation provides one way to do this; an alternative approach, which is faster and less prone to convergence problems, involves noting that the model is effectively equivalent to a simple linear regression model in which $\log\left(\frac{N_t}{N_{t-1}}\right)$ is the response variable and $f(N_{t-1})$ is the explanatory variable. In situations where $f(N_{t-1})$ is a function of the data alone – i.e., does not depend on any unknown parameters – the parameters α, β and σ_p^2 can be calculated using standard software for fitting a linear regression model (e.g., within R, the **lm** function in the **stats** package). We use **lm** in this project, because initial results suggested that the results were more stable, and less prone to issues of non-convergence, than when using **nlme.** Note that for a real-world analysis of smaller scope than this project, **nlme** would still be an attractive option, but non-convergence issues mean it is not straightforward to automate the running of **nlme** so that it can successfully fit a large number of models without manual intervention.

3.3.3. Dealing with missing data

The parameters of the population growth models can be estimated using **lm** whenever a count is available for both the current and previous year. We required at

least ten years of data on abundance to be available for which this criterion was fulfilled in order to be able to apply these methods. Note the choice of ten years as a threshold was somewhat arbitrary, but was designed to be extremely minimal - it could certainly be argued that ten years of data is still insufficient to reliably fit time series models of this complexity.

Within a draft version of this report we attempted to deal with situations in which missing data meant minimum data requirements were not met by using an alternative form of parameter estimation - data cloning, a modern computational statistical approach to maximum likelihood estimation in complex models (Lele *et al.* 2007) – to fit the models. The results, however, were extremely poor, and this appeared to be due to issues of non-convergence. The R package for applying data cloning to PVAs (PVAClone (Nadeem & Lele 2012)) makes use of the JAGS package (Plummer 2003), which is a widely-used piece of software for fitting complex models in a Bayesian framework via Markov chain Monte Carlo, but which is very difficult to apply successfully in an automated way.

Within the context of this project it was not possible to check and solve model convergence issues within PVAClone (given the large number of analyses being considered). Therefore, at the suggestion of the Project Steering Group, we considered an alternative approach, in which parameter estimation was performed in the standard way (using **lm**), but imputation was used to infill missing data prior to fitting the models.

Imputation is a widely used statistical approach to deal with missing data by replacing missing data values with “imputed” values, which are then analysed as if they were real data values. There are two broad approaches to imputation – “single imputation”, in which a single value is selected to replace each missing value, and “multiple imputation”, in which multiple possible values are selected. Multiple imputation is more defensible, because it accounts for the uncertainty associated with the imputation of missing data, but also more computer intensive, because it means the analysis needs to be re-run for each imputed dataset. The large numbers of models being fitted for this project meant that it was only feasible to use single imputation.

Various modelling approaches to imputation are possible. We followed the non-parametric approach of JNCC[1], which, in turn, follows that of Thomas (1993). This approach is a generalisation of the “simple chaining method” that has traditionally

[1]http://archive.jncc.gov.uk/pdf/Methods%20of%20analysis%20for%20production%20of%20indices%20of%20abundance%20and%20estimation%20of%20productivity1.pdf

been widely used in calculating annual indices of population size within ecology; the generalized version makes more efficient use of the available data than simple chaining. For a specific colony, and pooling region classification, the approach involves two stages:

Stage 1: Determine the pooling region containing the colony of interest, and within this region for each pair of years, j and k, calculate the ratio r_{jk} of total abundance in year k to total abundance in year j, where these totals are calculated by summing observed counts for the set of colonies within the pooling region that are observed in both years j and k.

Stage 2: For the colony of interest i and a particular year j, in which no count is available, calculate the imputed count to be:

$$z_{ij} = \frac{1}{|T_i|} \sum_{k \in T_i} y_{ik} r_{jk}$$

where y_{ik} denote the observed counts at this colony for the set of years T_i in which observed counts are available. The imputed value is, therefore, a weighted sum of the observed counts for the colony of interest, where the weights are calculated based on year-to-year variation at the level of the pooling region.

The time series for the colony interest is then taken to consist of observed counts, for years where these are available, and imputed counts, for years where no observed count is available.

We applied this method to all colonies that have at least five observed counts; imputed values were generated using each of the possible pooling region classifications.

Diagnostic checks on values generated by the imputation procedure suggested that it will sometimes produce implausible values. We, therefore, imposed a set of filters, in order to remove imputed values that lacked defensibility. Specifically, we removed values:

a. that were for years outside the set of years with observed counts for the colony of interest – i.e. years that were being extrapolated rather than interpolated;

b. that lay outside the range
minimum observed count$/2$, maximum observed count$*2$) and so were inconsistent with the observed counts for this colony;

c. that were associated with colonies at which either the maximum observed count was very low (less than ten) or the minimum observed count was zero, on the grounds that the results of ratio-based calculations become highly unstable when applied to very small counts, and become meaningless when applied to zero values.

3.4. Semi-integrated Bayesian population models

The key limitation of the approaches outlined in Sections 3.2 and 3.3 is that they make only partial use of the available data. Leslie matrix approaches primarily only use data on survival and breeding success, and ignore data on abundance (asides from the initial population size). Time series approaches use data on abundance but ignore data on survival and breeding success. Both forms of data – abundance and demography – should, in principle, be able to provide information that is relevant to the generation of predictions, so it is problematic that each of these approaches makes only partial use of the available data.

Integrated population models (IPMs) attempt to overcome this by making use of all available data. The models considered here were those used in Freeman *et al.* (2014). They are not, strictly speaking, Integrated Population Models, because they do not estimate the model parameters simultaneously from multiple data sources. However, they do have the key practical feature of Integrated Population Models – that they use both demographic data and data on abundance to infer the underlying size of the population in each year, and in generating predictions for future years. We therefore refer to these as “Semi-integrated” population models (SIPMs; we could also refer to them as Bayesian state space models, as they were termed in Freeman *et al.* 2014 and Jitlal *et al.* 2017, but that terminology is ambiguous in this context because time series models can also be fitted as Bayesian state space models).

The basic model structure of Freeman *et al.* (2014) is similar to that of a Leslie matrix model; it is the fact that parameter estimation is based upon multiple sources of data, rather than the underlying structure of the model, that separates this approach from the simpler Leslie matrix approaches outlined in Section 3.2.

The model considered by Freeman *et al.* (2014), was, more specifically, of the following form for each year t:

Number of adults: $A_t \sim \text{Poisson}(N_{t-1}s_{t-1})$

Number of individuals recruited: $R_t \sim \text{Poisson}\left(N_{t-a}\frac{f_{t-a}}{2}v^a\right)$

Total population size (unobserved)**:** $N_t = A_t + R_t$

Observed abundance: $y_t \sim \text{N}(N_t, \sigma_E^2)$

where s_t and f_t refer to year-specific adult survival and fecundity rates, v represents the immature survival rate, and a represents the age at first breeding. The annual survival and fecundity rates are assumed to follow log-normal distributions, of which mean and standard deviation are determined *a priori* from empirical data on survival and breeding success (respectively). The remaining model parameters - the immature survival rate v, the level of observed error σ_E^2, and the initial true population sizes (at the start of the time series) - are then estimated by fitting the model to observed abundance data via MCMC using the JAGS software (Plummer 2003). Further information on priors, inference and data pre-processing is given in Freeman *et al.* (2014).

3.5. Summary of methods

The set of PVA methods that we considered is summarised in Table 0-5.

Table *0-5*.

Method	Model type	Specific model	Type of data required	Minimum data requirements	Survival rates
ATG	Abundance time series models	Simple growth model	abundance	10 years+ in TP for which abundance data are available in both current and previous year	Not relevant
ATR		Ricker	Abundance		
ATZ		Gompertz	Abundance		
LDN	Leslie matrix models	Deterministic	Demographic rates	1+ years breeding success data in TP, and 1+ years abundance data in TP	National
LDF					Forth-Tay
LMN		Stochastic – constrained productivity	Demographic rates	2+ years breeding success data in TP, and 1+ years abundance data in TP	National
LMF			Demographic rates		Forth-Tay
LUN		Stochastic – unconstrained productivity	Demographic rates		National
LUF			Demographic rates		Forth-Tay
IPM	Semi-integrated population model	Freeman et al. (2014)	Abundance and demographic rates	See Freeman et al. (2014)	Forth-Tay

We considered all possible combinations of PVA methods and pooling region classifications, except that for Semi-integrated population models we only considered the Forth-Tay SPAs, for which outputs have already been generated. We therefore considered a total of (9 * 10) + 1 = 91 PVA “methods” (combinations of statistical methodology and pooling region classification).

4. Additional considerations in PVA modelling

In this section, we briefly review the impact of two key issues upon PVAs: density dependence and the existence of meta-populations. These processes were not able to be represented within this project, however, we briefly summarise current understanding regarding the influence of these processes upon population dynamics in seabirds, of relevance to PVAs.

4.1. Density dependence

Should an Offshore Renewables Development (ORD) act to reduce abundance at a colony (through lethal or cumulative sub-lethal effects), density-dependent demographic responses may partially compensate for these losses through increased productivity or survival of remaining individuals (Horswill & Robinson 2015; Horswill *et al.* 2016; Cook & Robinson 2016). There is considerable evidence for density-dependent regulation of population processes in UK seabirds (Horswill & Robinson 2015); however, the precise form and strength of these relationships remain uncertain (Cook & Robinson 2016), in part due to a lack of broad-scale studies and to site-specific variation in environmental effects.

Life history theory suggests that for long-lived species such as seabirds with low productivity, adults will buffer against the effects of adverse environmental variation by sacrificing reproductive success over their own survival (Williams 1966). Therefore, the most likely impact of density dependence will be to act on reproduction, leading to negative density-dependent effects on per capita population growth rates. The incorporation of density-dependent processes into population models will tend to lead to a reduction in the rate of projected population declines, for species that have a negative population trajectory (Weimerskirch 2001). The inclusion of such density dependent demographic processes in PVA models will, therefore, tend to lead to results that are less precautionary, but hopefully also more biologically realistic and, therefore, accurate, than the results obtained using density-independent models lacking any compensatory feedback mechanisms (Cook & Robinson 2016).

A review and simulation study by Cook & Robinson (2016) demonstrated that whilst most PVA output metrics were sensitive to inclusion of density dependence in models, they were relatively insensitive to the assumed form of density dependence. The authors, therefore, recommended that where there is good evidence for the presence and direction of density dependent relationships in seabirds they should be incorporated into population models and used to generate PVA metrics. However, with the important caveat that these relationships appear to be highly site-specific and therefore cannot be assumed to be present at all sites (Cook & Robinson 2016), thereby affecting the efficacy of incorporating such processes into regional PVAs.

Density dependent effects on survival of immatures in the context of PVAs have already been assessed within state-of-the-art SIPMs for one species in the Forth-Tay region, the common guillemot (Freeman *et al.* 2014). This model can only be applied

to data for relatively data-rich colonies, however, and so an assessment of its performance within this project is only made for a single species in this region.

4.2. Meta-populations

PVA models lacking a meta-population structure assume populations are closed, which is an unrealistic assumption for many seabird species. Seabird colonies, including those designated as SPAs, are typically sub-populations within larger meta-populations, which are collections of spatially distinct sub-populations linked by dispersal and migration (Hanski 1999). When PVA models define populations in terms of SMP count units (as in this study) this is likely to be a particular issue, because these units often do not refer to biologically distinct populations, but rather to subdivisions of a breeding colony. Meta-population models aim to quantify profoundly more complex dynamics than single population models, which typically assume closed populations, by factoring in movements among sub-populations (dispersal, immigration, emigration) as well as colony-specific survival and productivity rates. Such models are highly relevant to conservation policy, particularly for mobile species such as seabirds where interchange among colonies arises through dispersal and emigration. However, the impact of meta-population dynamics has rarely been considered in conservation policy (but see Sanz-Aguilar *et al.* 2016), including marine renewable assessments.

Whilst meta-population models have been applied to seabird populations (e.g. Spendelow *et al.* 1995; Inchausti & Weimerskirch 2002; Dugger *et al.* 2010), they require the estimation of between-colony rates of movement involving the simultaneous study of marked individuals in several colonies or areas containing local populations. Some data exist for movement and dispersal of individuals between colonies in seabirds (e.g. for European shag and kittiwake, Danchin & Cam 2002; Coulson 2011; Barlow *et al.* 2013; Grist *et al.* 2014; 2017), and it is broadly recognised that this process should be included within population and PVA models used to assess impacts of disturbance (Furness & Trinder 2016). However, even when such data do exist, the statistical estimation of rates of probability of movement between colonies is non-trivial (Dugger *et al.* 2010), and often highly context dependent in terms of population densities and environmental conditions. Therefore, few reliable estimates of inter-colony movement processes exist for seabirds, hindering the use of meta-population dynamics within PVA assessments. The development and fitting of meta-population models is beyond the scope of this project, and so is not considered further.

5. Methodology for comparing PVA approaches

The key objective of this project is to evaluate and compare the performance of different population models underlying PVA methods in practice, within the context of specific seabird populations (15 species, in the waters of the British Isles).

PVA is concerned with the use of population models to predict future population sizes, potentially under different impact scenarios. In this project, we compare how well different population modelling methods, typically employed within PVAs, perform in terms of accurately predicting future population sizes. We note that predictions of absolute future population sizes are often not the key output from PVAs – we focus on them here primarily because they are an observable quantity, for which extensive data already exist, and so provide a basis for assessing the empirical performance of the population models that underpin PVAs. Care should be taken in generalising the results of the evaluations to the performance of methods in producing PVA metrics, although we think the ability of methods to produce accurate predictions of absolute abundance provides a good overall assessment of their defensibility.

The best way of evaluating the performance of these models in predicting abundance is to split the available population abundance data into a “training period” that will be used as a basis for generating PVAs and a “test period” that will be used as a basis for evaluation. Such an approach is standard in assessing the performance of predictive models. The evaluation will involve comparing the actual counts within the test period to the predicted values for this period generated by the population models; we will also account for the uncertainty measures produced by the population models within this evaluation.

5.1. Length of test period

We define the split into training and test data in terms of the length of the test period: the final T years (leading up to 2017, the final year for which data are available) are assumed to form the test period, and all years prior to that are assumed to form the training period.

The choice of the length of the test period is a challenging issue: using a short test period reduces the amount of data available for assessing performance, but using a short training period will necessarily reduce the performance of the methods being evaluated. In practice, PVA is often used to predict relatively far into the future (e.g. 20 years ahead); using a long period might therefore seem to be of most practical relevance, but this is not necessarily the case. PVAs generated now, in practice, will

be based on datasets that have been collected for over 30 years, but if we consider a test period of 20 years then we will be comparing PVAs that have been generated using a dataset that spans only around 10 years. Focusing on a short test period (of, say, five years) is, however, not necessarily indicative of the ability of a method to generate accurate predictions over longer periods.

We therefore consider four potential test periods lengths within our main (national) evaluation:

- 5 years (2013 – 2017)
- 10 years (2008 – 2017)
- 15 years (2003 – 2017)
- 20 years (1998 – 2017)

These choices are designed to cover a range of timespans, and levels of coverage. Test periods that are longer than 20 years were not considered, due to the sparsity of data within the training period when considering longer test periods.

For the Forth-Tay evaluation we only consider a fove year training period (2013-2017), because SIPM outputs are only available for a single time period (the period ending in 2012).

5.2. Choice of test data to use for evaluation

Within the test period, we focused upon using count data for evaluation. We used *all* available count data within this period for evaluation – i.e., for each species an evaluation is performed for every colony-year combination that has data within the test period. Let y_{ij} denote the observed count for year j at colony i. Where relevant (e.g. for guillemot and razorbill) counts are converted into pairs prior to modelling and testing, to ensure comparability between species and methods (the SIPM models of Freeman *et al.* 2014 worked with pairs).

We evaluated the performance of different population modelling methods in predicting counts in the test period using a range of different criteria (Table 0-6):

- Ability to run model – i.e. whether it is possible to generate a projected population size
- Occurrence of highly implausible results
- Lack of systematic bias
- Low error

- Accurate quantification of uncertainty
- Low level of uncertainty
- Ease of computation – i.e., time for computation

We included all sites that have at least one count within the test period *and* one count within the training period in the evaluation. The requirement to have at least one count within the training period is necessary because all population modelling methods require at least one observed count from the colony of interest (in order to initialize the PVA).

5.3. PVA outputs

Not all PVA methods will produce predictions for all of these colony-year combinations: for some colonies it will be impossible to apply some methods because of data sparsity.

Where a PVA method, k, can be applied, the method will produce:

1) a "best" prediction (predicted mean) μ_{ijk} for the colony-year combination; and,
2) for most methods, an estimated distribution of predictions, capturing the uncertainty associated with the "best" prediction.

The estimated distribution may either be represented as a theoretical distribution (e.g., the cumulative distribution function, CDF, F_{ij} associated with a parametric probability distribution) or else consist of a set of simulated values that have been drawn from this distribution. In either case, we can summarise this distribution in the following ways:

a) as the standard deviation of the distribution ("prediction standard error");
b) as a 95% prediction interval; or
c) as the probability of being below or above the observed count. For theoretical predictive distributions the probability of being below the observed count can be calculated as $F_{ijk}(y_{ij})$, and for simulated values it can be calculated as the proportion of values that are less than the observed count y_{ij}.

Regardless of how it is calculated, we let p_{ijk} denote the estimated probability of the prediction generated by method k being below the observed count y_{ij}. The probability of being above the observed count will be equal to one minus this quantity.

5.4. Evaluations

5.4.1. Main evaluation

Our main evaluation involves applying six different population models (three time series models [ATG, ATR, ATZ] and three Leslie matrix models [LDN, LMN, LUN]) to all species-colony-test period combinations for which minimum data requirements are met. We then compare these against observed numbers of breeding pairs for all years within the test period for which counts at the colony are available. Each of the PVA methods is applied using nine different possible pooling region classifications – the regional definitions given in Table 3 with the exception of R9. These classifications are used to pool productivity data (for the Leslie matrix methods: LDN, LMN and LUN, see Table 5 for model definitions) or in imputing abundance data (for the abundance time series models: ATG, ATR and ATZ, see Table 5 for model definitions). Survival rates are, in all cases, based on those in Horswill & Robinson (2015). A total of 6 * 9 = 54 PVA methods (combinations of population model and pooling region classification) were, therefore, potentially applied for each species-colony-test period combination.

5.4.2. Forth-Tay evaluation

However, a key interest in this project lies in comparing the SIPM against the other methods. The SIPM can be regarded as the current “gold standard” approach because it is the approach that makes use of the widest range of data and is the most methodologically defensible approach. The approach is more time consuming than other approaches to implements, however, in terms of both computational and human resources, which meant that it was not possible to generate additional PVAs using this approach within the remit of this project. We therefore focused upon using PVAs that were generated using this method in previous projects (Freeman *et al.* 2014, Jitlal *et al.* 2017). The comparison of the SIPM against other methods was therefore restricted to a single five year period (2013-2017), a single region (the Forth-Tay region, which contains four SPAs) and five species (herring gull, kittiwake, guillemot, razorbill and puffin). As for the main evaluation, it was also not possible to evaluate all species for all populations. Given the restricted temporal and spatial coverage, the results of this evaluation should therefore be interpreted very cautiously.

The methodology for this evaluation was, insofar as possible, identical to that for the main evaluation, but there were necessarily a number of differences:

1. Four additional “population models” (IPM, LDF, LMF, LUF; see Table 5 for model definitions) and an additional “pooling region classification” (R9) were considered within the Forth-Tay evaluation, but not within the main (national) evaluation. One of the additional population models was the SIPM (the inclusion of this being the main rationale for running this additional evaluation). The other three additional “populations models” (LDF, LMF, LUF; see Table 5 for model definitions) were not really distinct new models, but were, rather, variants of the Leslie matrix models in which local survival rates from the Forth-Tay region were used in place of the species-level rates from Horswill & Robinson (2015). These were included in order to ascertain whether differences between the SIPM outputs and other methods were due to differences in input data (the SIPM uses the local survival rates) or differences in model structure.

2. When the abundance time series models were applied using the SPA-level pooling regions (R9) the imputation method used by Freeman *et al.* (2014) was used in place of the JNCC imputation method, to ensure consistency in the comparison against the SIPM results.

3. For the Forth Islands SPA, the PVA methods were used to produce SPA-level predictions (summed across all colonies within the SPA), and these were compared against SPA-level counts within the test period (also summed across colonies), following the approach used in the SIPM modelling. For Fowlsheugh SPA and Buchan Ness to Collieston Coast SPAs, the SPAs correspond to a single colony in the SMP data, so no summation across colonies is needed in order to produce SPA-level population sizes. For St. Abb’s Head to Fast Castle SPA, we follow Jitlal *et al.* (2017) in only modelling one of the colonies within the SPA, St Abbs Head NNR, because of the sparse data for the remaining colonies.

5.5. Assessing performance

5.5.1. Comparing predicted and observed counts

We compared predicted to observed counts using the log-ratio between them, where logs were calculated to base 10. A value of one was added to both predicted and observed counts prior to the calculation, in order to deal with the possibility that the observed count is zero. In mathematical terms, we considered the following ratio:

$$r_{ij} = \log_{10}\left(\frac{y_{ij}+1}{\mu_{ij}+1}\right)$$

for each combination of colony and year.

Focusing on log ratios ensures that increases and decreases are dealt with in a symmetric way, and ensures that large increases in absolute numbers do not dominate the calculations. The use of base ten for the logarithms is to aid interpretation; it means that:

- A value of $r_{ij} = 2$ means that the predicted count is $10^2 = 100$ times higher than the observed count;
- A value of $r_{ij} = -2$ means that the predicted count is $10^{-2} = 1/100^{\text{th}}$ the value of the observed count;
- A value of $r_{ij} = 0$ means that the predicted and observed counts are identical.

5.5.2. Measures of performance

For each species-colony-year-method combination for which an evaluation was possible, we assessed the ability of the method to predict the observed count using a range of criteria, and then averaged these criteria across colonies, year and species to look at the overall performance of each method. It should be noted that the set of species, colonies and years for which an evaluation is possible is neither a random sample of the set of all species, colonies and years, nor likely to be representative of that set, so the results of these evaluations should be interpreted cautiously.

We focused upon seven different criteria for assessing for performance, which we represented as a hierarchy – the criteria that are lower in the hierarchy are only relevant if a method performs well at the higher levels.

C1: Ability to run. The first criterion assesses whether the method is able to produce a prediction at all – i.e., whether the minimum data requirement for

producing a population model is met, and whether the method can be used to successfully generate a prediction (e.g., no fatal problems with non-convergence or fitting are encountered). For any particular combination this criterion can simply be represented by a binary variable, which is either one (prediction fails to be produced) or zero (prediction can be produced); by averaging across years, colonies and species we can estimate the percentage of situations in which each method fails. Note that we did not assess SIPMs against this criteria, because for the Forth-Tay evaluation we only focused on the species-SPA combinations for which SIPMs had already been run.

C2: Occurrence of highly implausible results. Methods sometimes run, but produce highly implausible predictions – e.g., predicted colony sizes that are many thousands of times higher than the current colony size. We classed a result as being "highly implausible" if $|r_{ij}| > 2$ – i.e., if the predicted count was more than 100 times the observed count or less than 1/100th of the observed count. This definition is clearly very conservative – we only classed a result as being "highly implausible" if the method produced predictions that differ from the observed count by an extremely wide margin. In practice, we anticipate that the PVA outputs that we class as "highly implausible" are so different to observed abundance values that they would always, or almost always, be flagged as being implausible based on expert judgement, and so would not actually be used. We calculated the percentage of results that are not highly implausible (NHI): ideally, this should be close to 100%.

C3: Lack of systematic bias. The next criterion assessed whether the method produced predictions that systematically over-estimated or under-estimated the true (logged) counts. We calculated this by looking at the mean or median value of the log-ratio r_{ij}, averaged across colonies, years, and, where relevant, species. Note that "highly implausible" results were excluded when calculating this (and subsequent) criterion. This value should ideally be close to zero.

C4: Low error in specific situations. A method that is systematically unbiased may nonetheless show considerable error – i.e., the method can over-estimate the true count in some situations, and under-estimate it in others, even if it shows no overall tendency to underestimation or overestimation.

We quantified this by focusing solely on the size, or magnitude, of the difference, $|r_{ij}|$, and averaging this magnitude across years, colonies and, where relevant, species (using either the mean or median).

C5: Reliable quantification of uncertainty. It is important that methods not only provide reliable predictions, but that they provide a reliable quantification of uncertainty. A simple way to evaluate this is by testing whether the observed count lies within the 95% prediction interval (1) or not (0). We calculated the percentage of situations in which this occurred. This would ideally be close to 95%.

C6: Level of uncertainty. All else being equal, it is desirable for the level of uncertainty associated with the prediction to be as low as possible. We therefore report the width of the 95% prediction interval. This should ideally be as small as possible. Note, however, that this criterion is only meaningful if the previous criterion is met – i.e. a low level of uncertainty is only desirable if the confidence intervals also have adequate coverage (i.e. contain the true value close to 95% of the time).

C7: Ease of computation. The final criterion refers to the practicality of implementing each method. We quantified this by the total computer time, in seconds, required to generate PVA predictions using each method.

These criteria are summarised in Table 0-6.

5.5.3. Summarising performance

We averaged each of these criteria of performance across colonies and years, to report overall performance of each method, species and test-training split at a range of different levels of spatial aggregation. As well as summarising overall performance for each test-training split, we also showed how performance varied in relation to the length of the gap between the final count in the training period and the year being used for method evaluation.

We also looked at the percentage of evaluations for which each method was the "best" method – either the method with the minimum absolute difference between predicted and observed values, or the minimum value of $|r_{ij}|$. By considering differences in absolute values, as well as ratios, we can check whether our focus on ratios, in the other criterion, is missing some aspect of performance.

One possible issue with these evaluations derives from the fact that not all methods could be applied in all situations. In order to account for the potential effects of this, we also run a separate additional set of evaluations solely using the species-colony-test period combinations for which it is possible to run all methods.

6. Results

6.1. National evaluation

We compared the performance of 54 different methods for performing population modelling forecasts – these were all possible combinations of six statistical methods for population models (ATG, ATR, ATZ, LDN, LMN, LUN; see Table 5 for model definitions) and nine methods for regional pooling (R0-R9; Table 3).

6.1.1. Raw results

The raw outputs from the population models recorded the observed count, predicted mean, predicted SE, and the value of the CDF at the observed count (p_{ij}) for each combination of test-training split, reporting region, species, colony, year and modelling method for which an assessment was possible.

The raw outputs are included, as a CSV file, in the Supplementary Information (SI1).

6.1.2. Criterion 1 – ability to run

We calculated the percentage of situations in which each particular population modelling method could be used. For each combination (of test-training split, reporting region, species, colony, year and method) a binary variable was used to indicate whether the method could be applied (1) or not (0): these were then averaged across year, colony and species to get an overall assessment of the range of situations in which each method could be run.

The overall results, averaged across species, are shown below (Figure 2); a breakdown by species is given in the electronic supplementary material.

The percentage of situations in which assessments could be run was highest for the regional versions of the Leslie matrix approaches (LDN, LMN, LUM) where regional classifications that include a fairly small number of regions were used (R3, R4, R5, R7, R8), with percentages of over 80% in most cases. The percentage of assessments able to be run dropped off in regional classifications with larger numbers of regions (R2 & R1) to around 50-60% for the Leslie Matrix approaches. The most local versions of the Leslie matrix approaches run at the site level (R0) were possible in around 20% of cases.

The time series approaches (ATG, ATR, ATZ) were possible in many fewer cases, with between 10% and 20% of potential assessments possible, regardless of the regional pooling method.

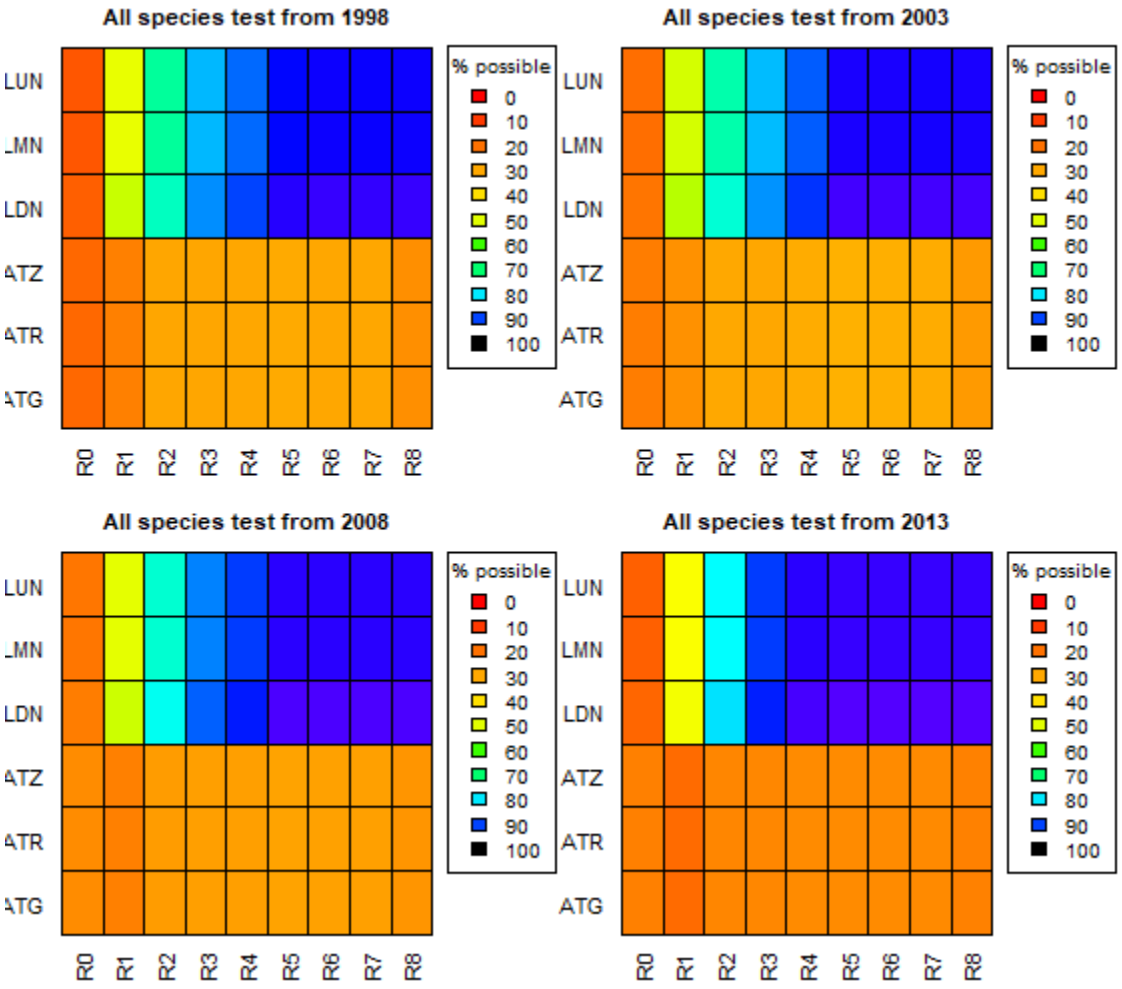


Figure 2: Summary of Assessment Criterion 1 – ability to run. The percentage of all potential model runs that were possible (i.e., that achieved minimum data requirements), calculated across all combinations of species and colony. Each cell represents the percentage of runs that were possible for each statistical modelling method (y-axis) and pooling region classification (x-axis). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

6.1.3. Exploratory graphs

We produce scatterplots (Figures Figure 3-Figure 7) of observed against predicted counts for each combination of statistical method and pooling method. The results show that, for all statistical and pooling methods, there is considerable variability in performance between species, sites and years - there are many combinations for which observed and predicted counts are quite similar (i.e., close to the 1:1 line), but also a large proportion of combinations for which there are substantial differences between them. Within each statistical method the results are qualitatively similar for all pooling region classifications (R0-R8).

There are however, some consistent differences between the different statistical methods. The Leslie matrix models (Figures Figure 3-Figure 5) frequently show large differences between observed and predicted counts, whereas the time series models (Figure 6-Figure 7) tend, overall, to show rather smaller differences. The time series models, however, do sometimes predict extremely high abundance values (in excess of 1 million), whereas this does not happen with the Leslie matrix models. Surprisingly, the Leslie matrix models never predict abundances of zero, but the time series models predict this relatively frequently – future work fully investigating why this is the case would be worthwhile.

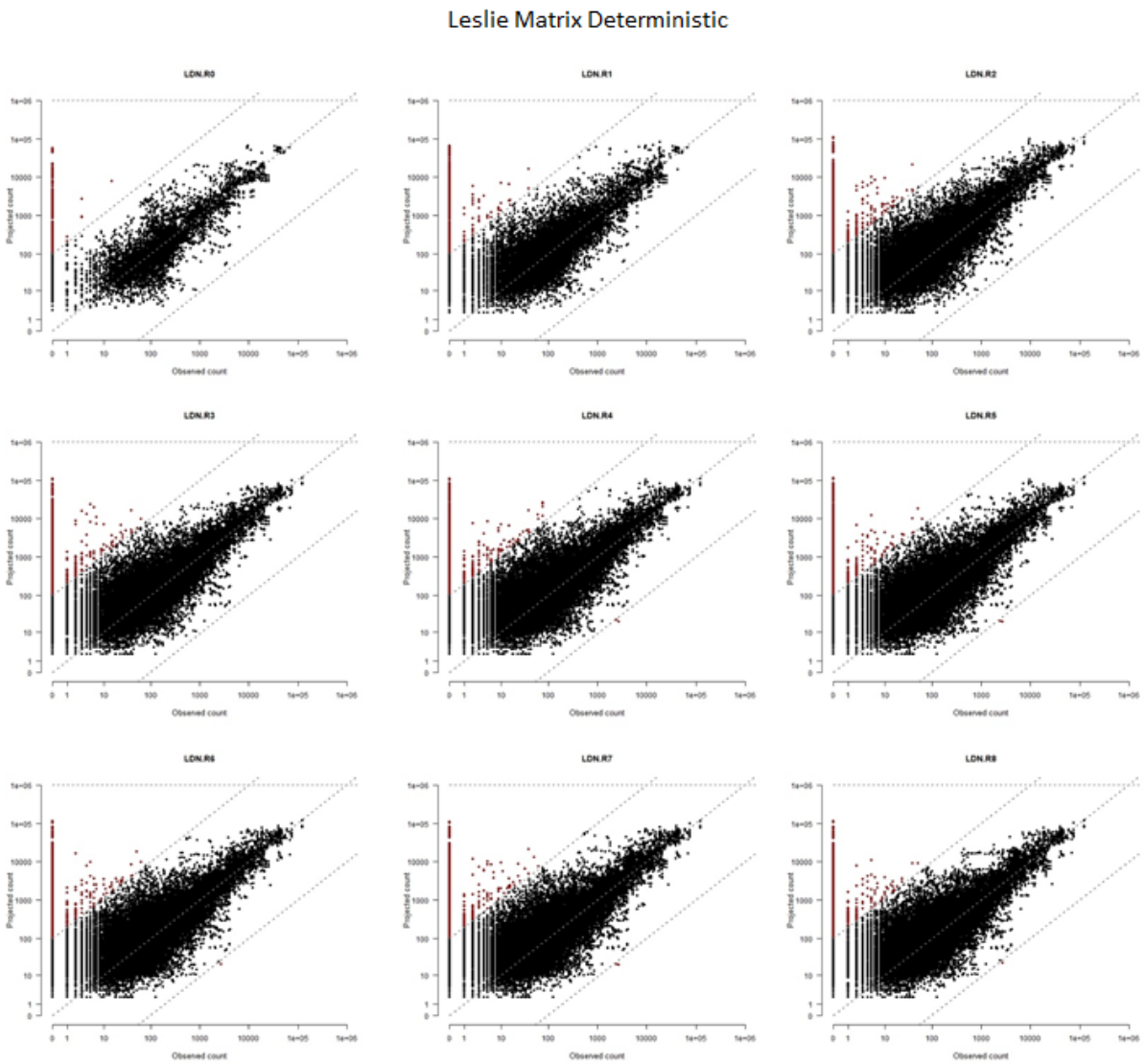


Figure 3: Scatter plots for observed versus predicted abundance over all species using the deterministic Leslie Matrix model (LDN). Each panel represents models using demographic data (productivity) from different pooling regions. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Top/bottom lines are where $|r_{ij}|<2$.

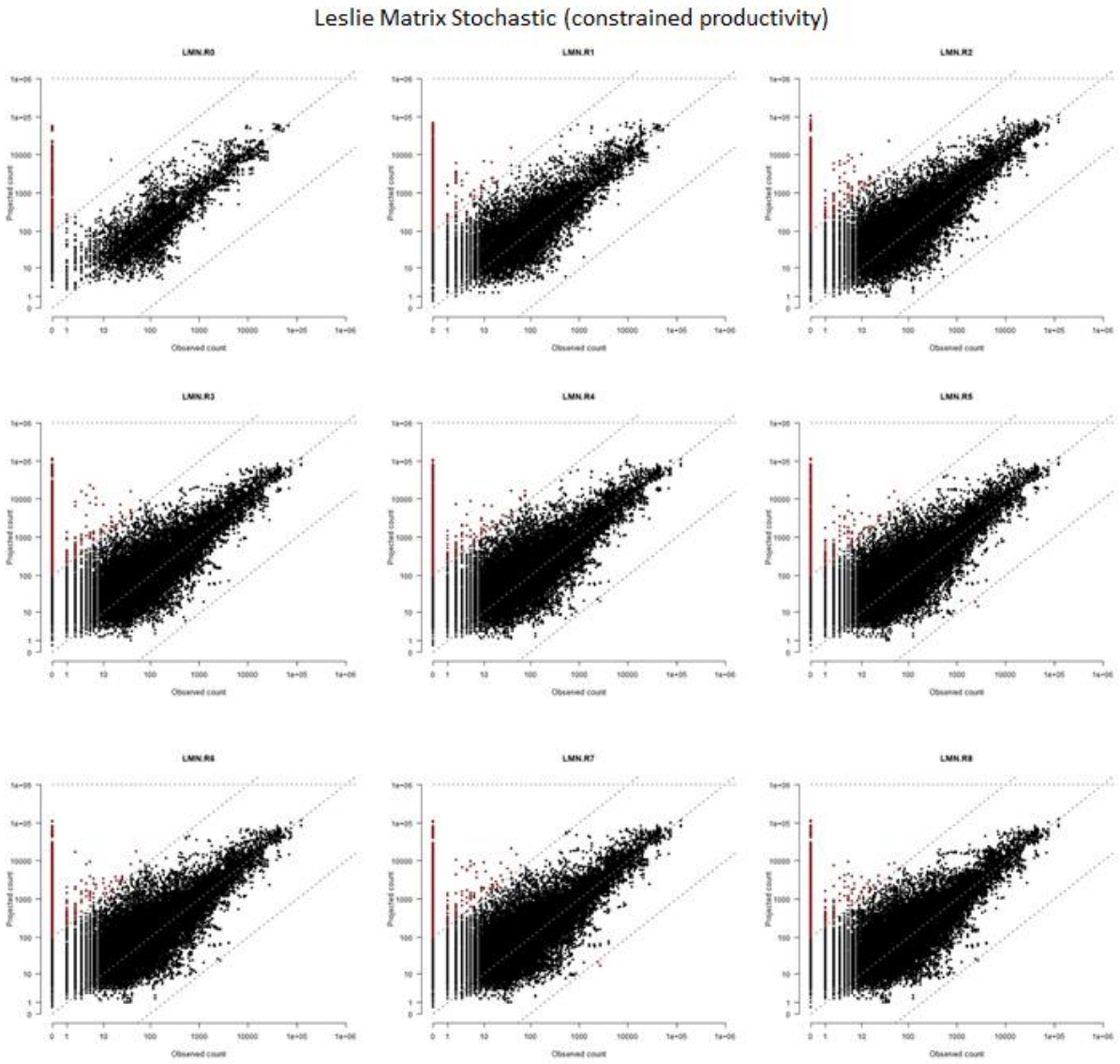


Figure 4: Scatter plots for observed versus predicted abundance over all species using the stochastic Leslie Matrix model with constrained productivity (LMN). Each panel represents models using demographic data (productivity) from different pooling regions. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Top/bottom lines are where $|r_{ij}|<2$.

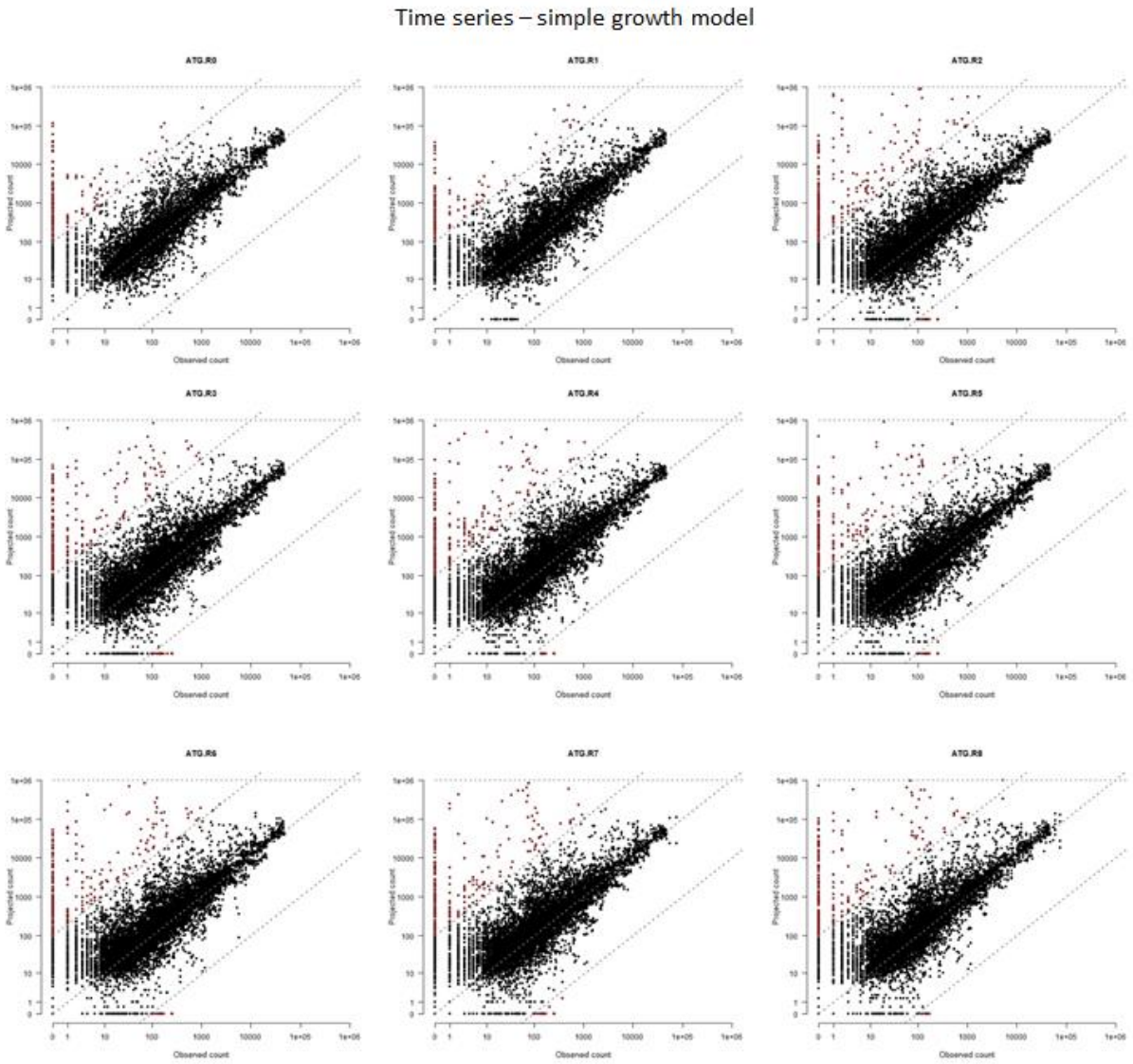


Figure 5: Scatter plots for observed versus predicted abundance over all species using the time series simple growth model (ATG). Each panel represents models using demographic data (productivity) from different pooling regions. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Top/bottom lines are where $|r_{ij}|<2$.

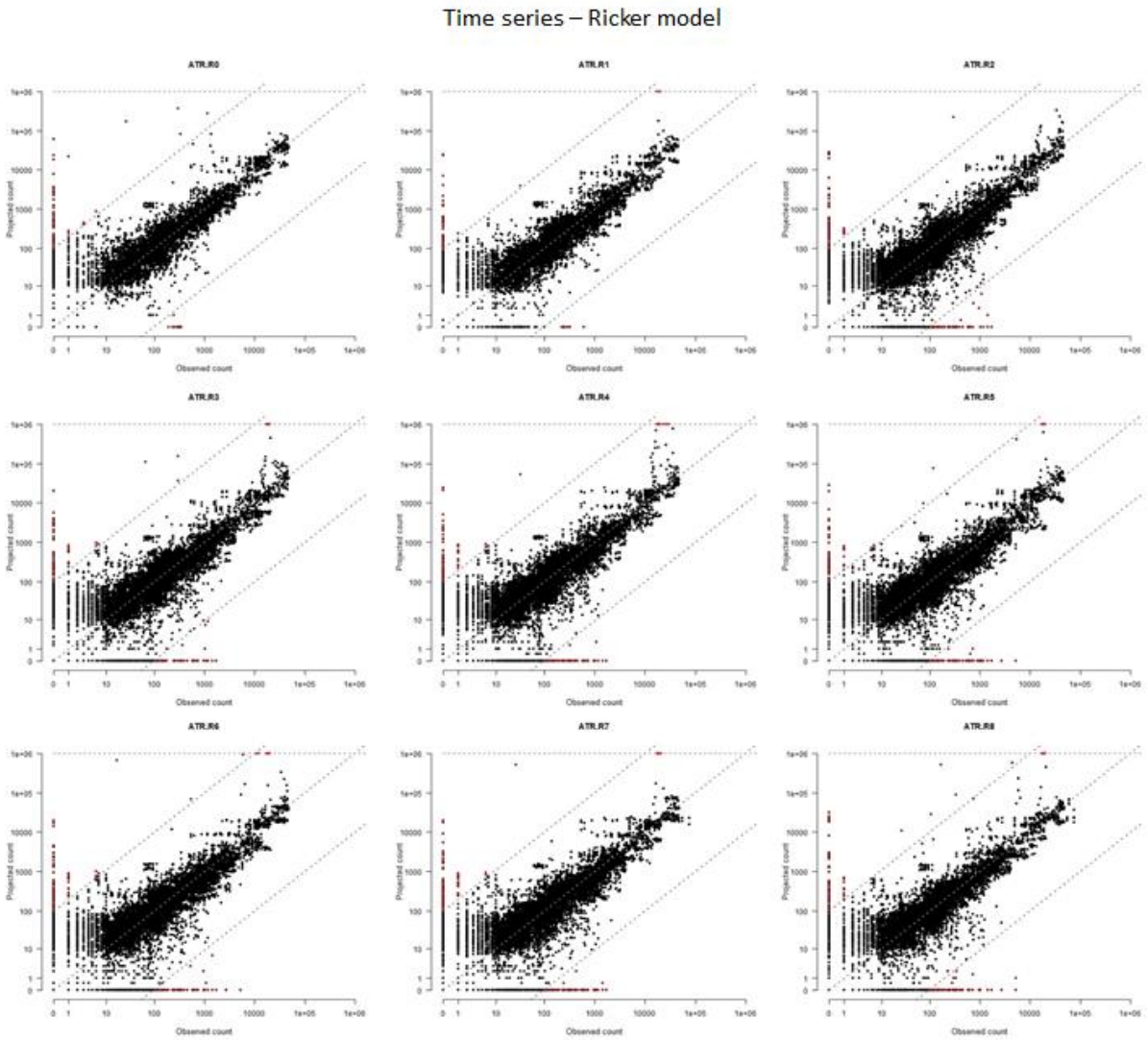


Figure 6: Scatter plots for observed versus predicted abundance over all species using the time series Ricker model (ATR). Each panel represents models using demographic data (productivity) from different pooling regions. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Top/bottom lines are where $|r_{ij}|<2$.

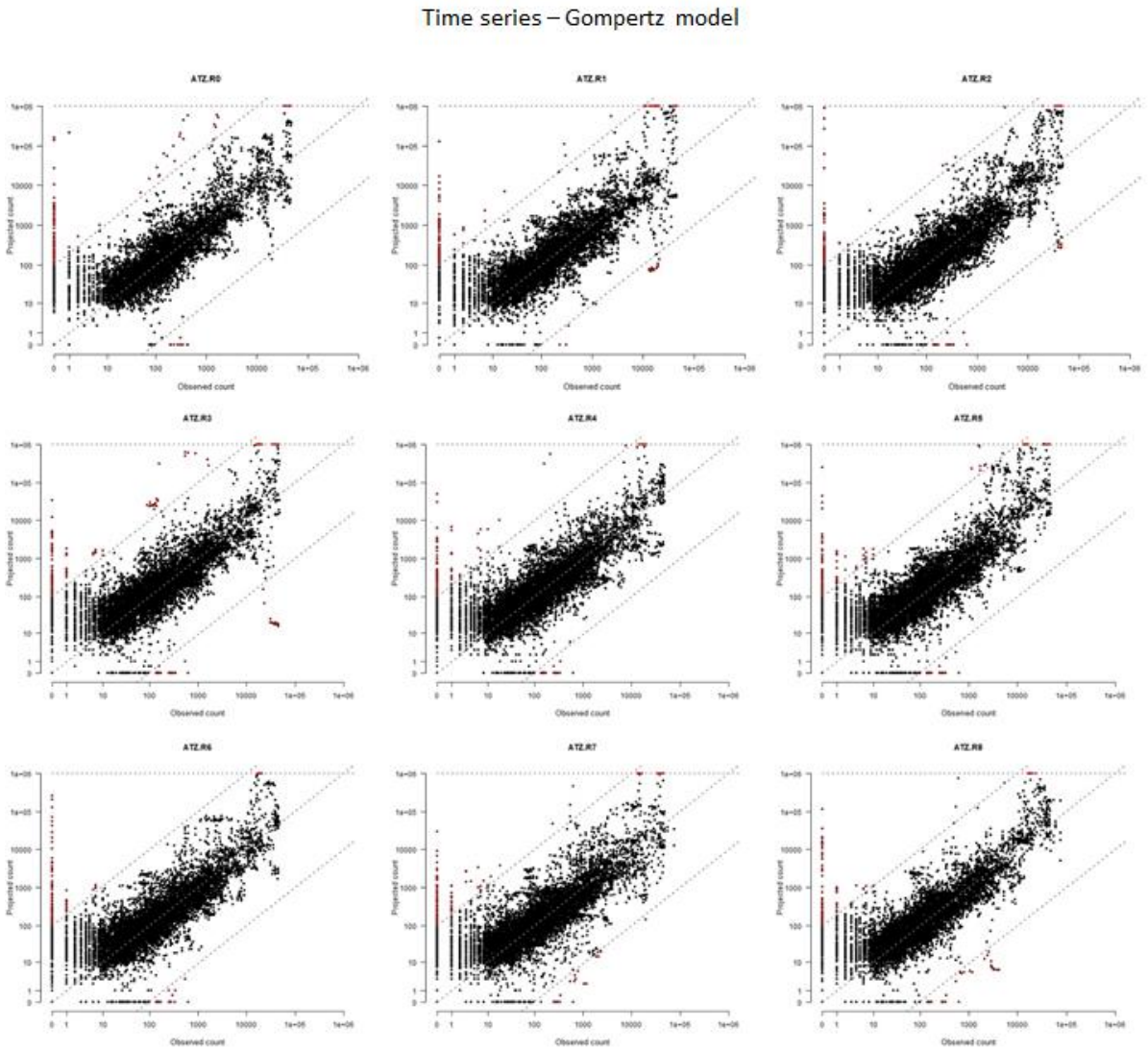


Figure 7: Scatter plots for observed versus predicted abundance over all species using the time series Gompertz model (ATZ). Each panel represents models using demographic data (productivity) from different pooling regions. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Top/bottom lines are where $|r_{ij}|<2$.

6.1.4. Criterion 2 – occurrence of highly implausible results

We used an extreme criterion to classify results as “highly implausible” (predicted abundances that were more than 100 times larger or smaller than the observed abundance), and therefore expected that methods should ideally have 100% of results classed as “Not Highly Implausible” (NHI).

When we looked across all combinations of species, colonies and years (Figure 8), it appears that the Leslie Matrix methods (LDM, LMN, LUN) result in substantially more highly implausible results than the time series methods. However, this is due to the much greater number of instances in which the Leslie Matrix methods can be applied (Figure 2), so these results include species-colony combinations for which data availability is poor, conditions in which any population modelling method is likely to perform poorly.

Therefore, assessing the different modelling methods across only those instances where all modelling methods could be assessed results in a fairer comparison. When we focus only upon the instances where all modelling methods could be applied (Figure 9), we see that the three Leslie Matrix approaches (LDN, LMN, LUN) performed consistently well, with only a very small proportion of results being flagged as highly implausible. The time series methods (ATG, ATR, ATZ) performed less well, with more than 5% of results being flagged as highly implausible in some situations. The simple growth model (ATG) resulted in the highest number of highly implausible results.

Using the mean rather than median predicted abundance (Figure 10), resulted in many more highly implausible results, particularly for the time series methods (ATG, ATR, ATZ). This is because some individual simulations of the time series models produced extremely large or small future population sizes, and these simulations impact the mean far more than the median. This highlights the importance of using different assessment criteria in the assessment of population models; in particular, it highlights the importance of interpreting the mean population size, averaged across simulations, carefully, given that the mean can be highly sensitive to the existence of extreme values.

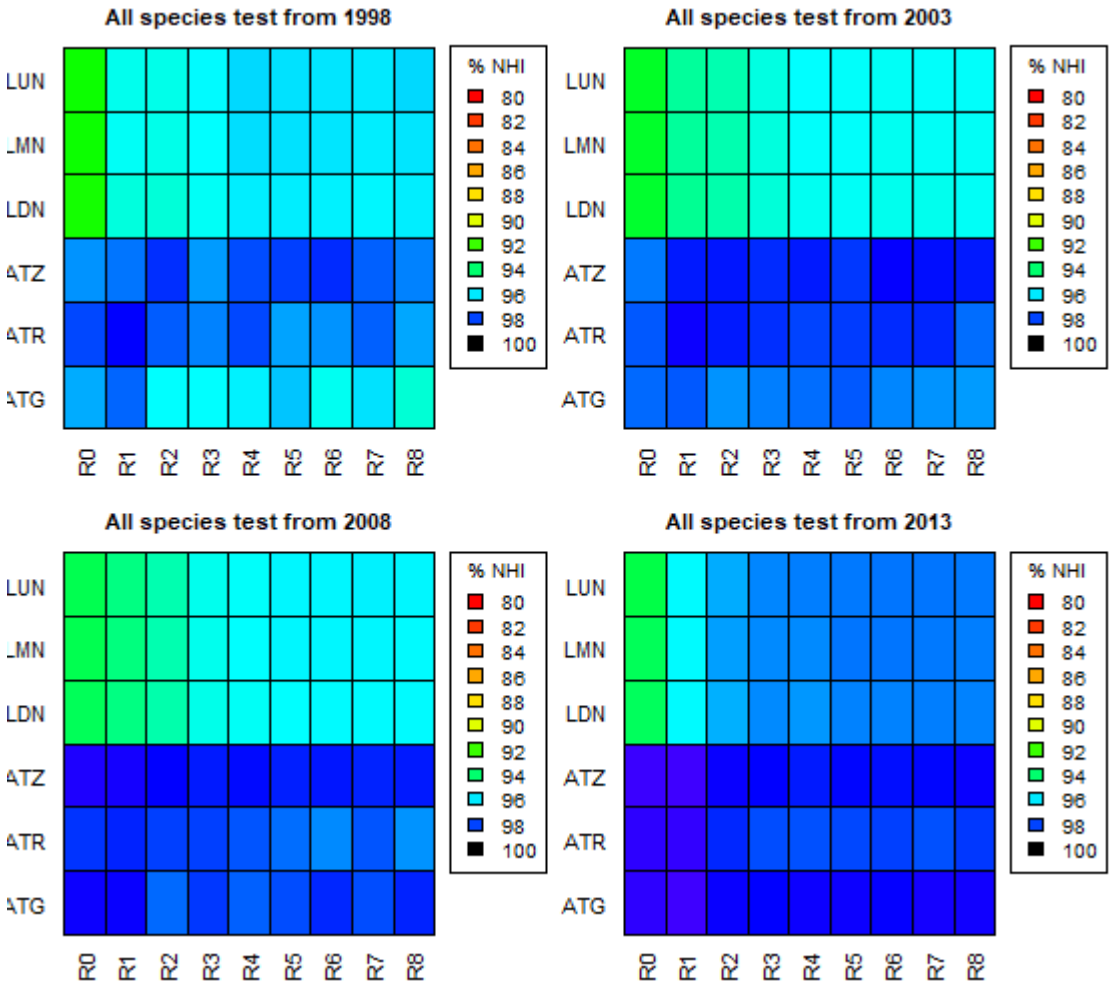


Figure 8: Assessment Criterion 2 – occurrence of highly implausible results. Shown as the percentage of **median** predicted population sizes that are 'not highly implausible', calculated across of all species, colonies and years. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

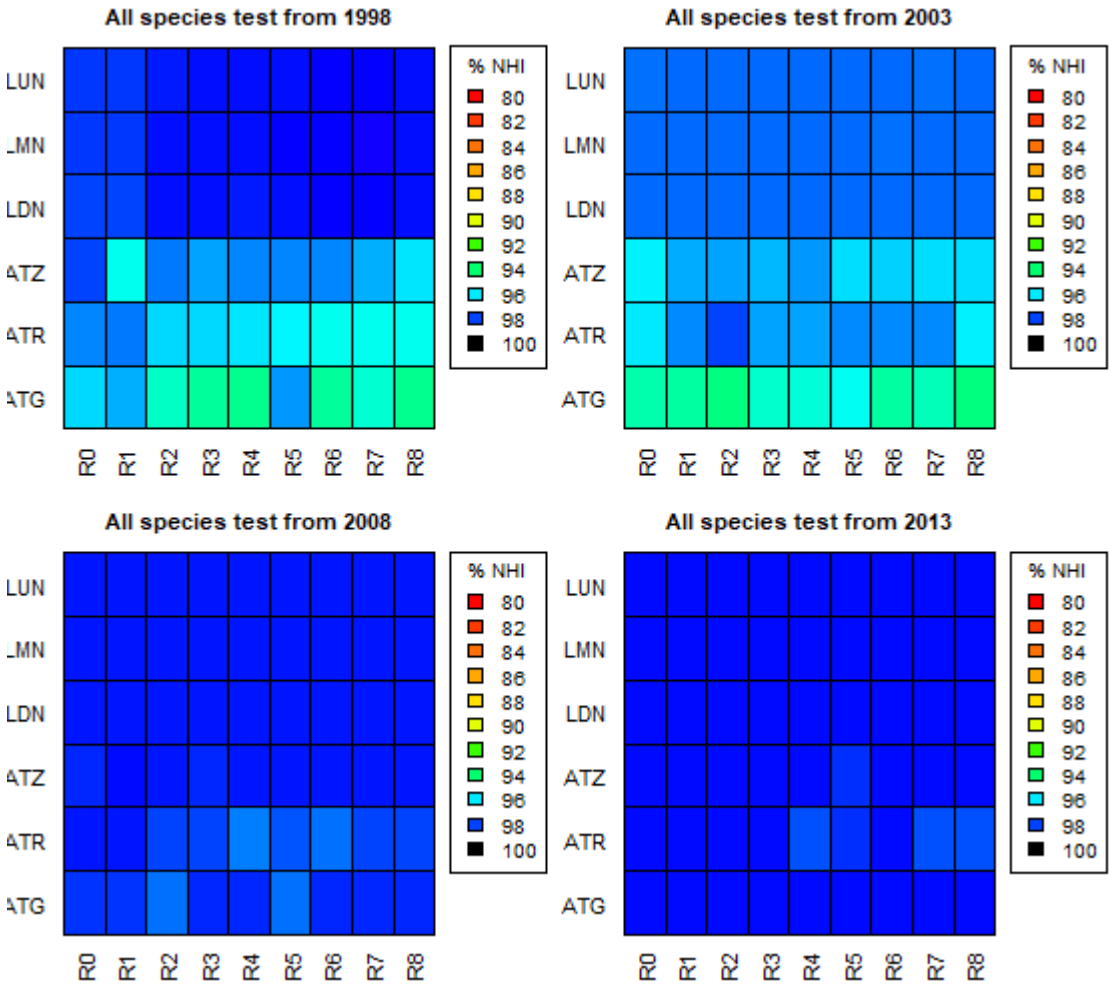


Figure 9: Assessment Criterion 2 – occurrence of highly implausible results. Shown as the percentage of **median** predicted population sizes that are 'not highly implausible', calculated across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

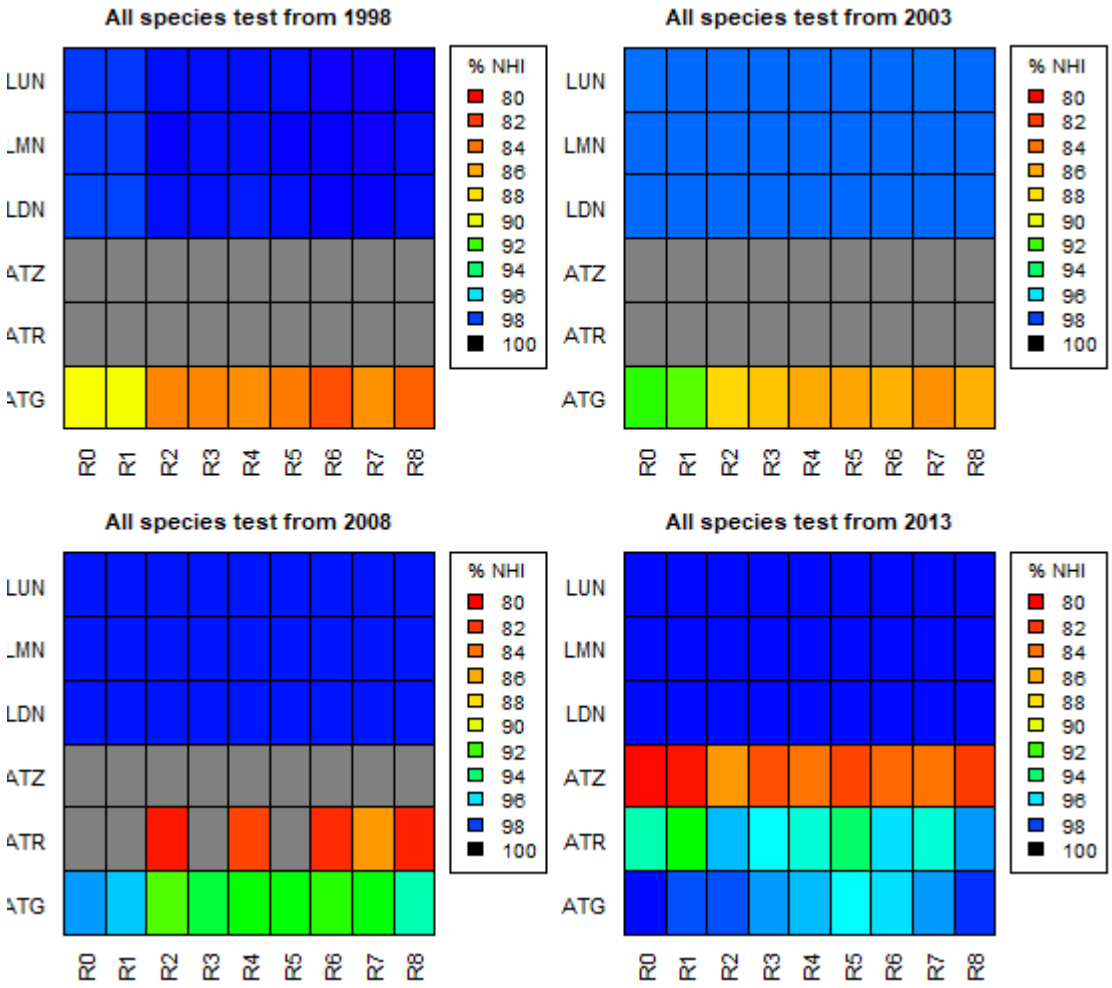


Figure 10: Assessment Criterion 2 – occurrence of highly implausible results. Shown as the percentage of **mean** predicted population sizes that are 'not highly implausible', calculated across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

6.1.5. Criterion 3 – Systematic bias

For this and later criteria, we focus solely on the subset of species-colony-year combinations for which all methods can be applied, because this gives the fairest basis for comparison. The results obtained by averaging across all combinations of species, colonies and years are shown in the Electronic Supplementary Information.

Levels of bias for median predicted abundance are generally lower for the time series methods than the Leslie matrix models, with the Leslie matrix methods tending to produce median abundance estimates that consistently underestimated the observed abundances in the test periods (Figure 11). There was no obvious change in bias across the different pooling regions for any of the modelling methods. However, for all methods, bias tended to become increasingly positive as the length of the training period increased (Figure 11).

Calculating bias using mean rather than median predicted abundance (Electronic Supplementary Information) led to different results: in this case the systematic bias for the Leslie matrix methods (LDN, LMN and LUN) was consistently lower than that for the time series models, and the simple growth (ATG) and Gompertz (ATZ) models had particularly high levels of bias. The biases for the time series methods tended to be positive (greater than zero in the legend), implying that these methods tended to systematically predict higher abundances than those observed.

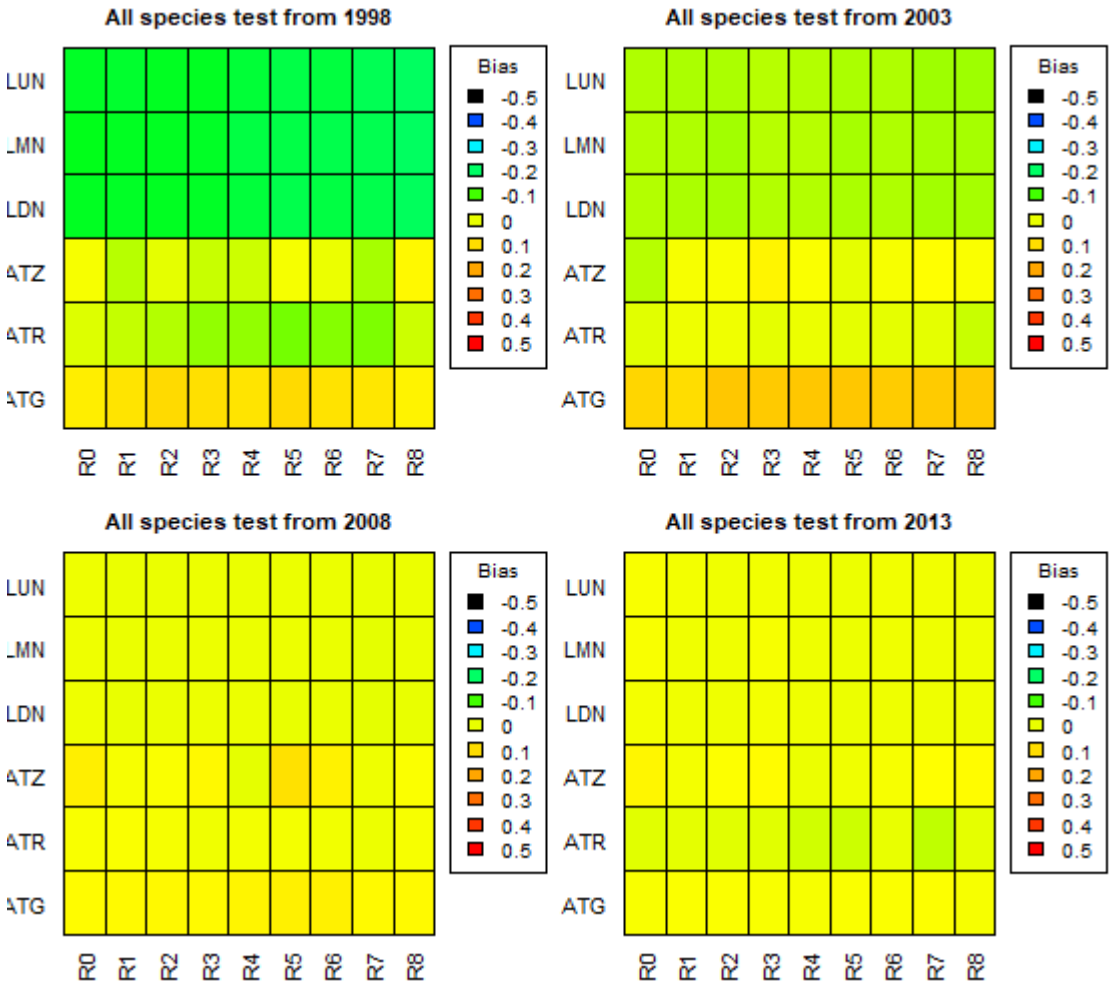


Figure 11: Assessment Criterion 3 – Systematic bias. Shown as the bias averaged across all species, colonies and years. Bias is calculated using the **median** predicted abundance from each model, and is calculated across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

6.1.6. Criterion 4 – Error

The results for error were qualitatively similar to those for bias. The errors associated with median predicted abundance were generally similar for different methods, but tended to be lowest for the simple growth and Ricker models, intermediate for Leslie matrix models, and highest for the Gompertz model (Figure 12). The level of regional pooling had very little impact on the magnitude of error associated with any of the modelling methods. Error decreased as the length of the training period increased for all methods (Figure 12).

Errors associated with mean predicted abundance (Electronic Supplementary Information) were, by contrast, consistently much higher for the time series models than the Leslie matrix models (with the Gompertz model, in particular, having very high levels of error in the mean predicted abundance).

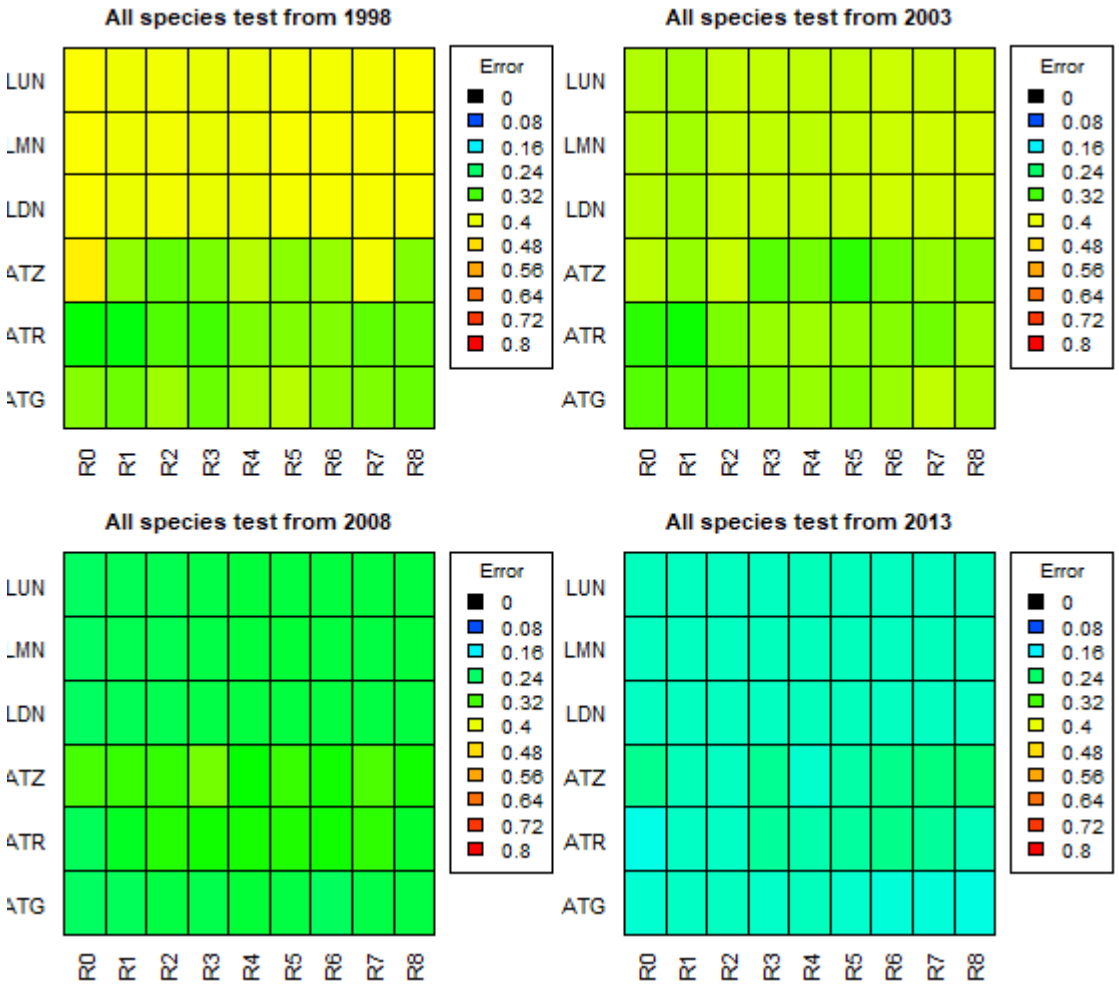


Figure 12: Magnitude of error for **median** predicted abundances, averaged across all species, colonies and years. Error is calculated using the **median** predicted abundance from each model, and is calculated across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

6.1.7. Criterion 5 – Quantification of uncertainty

Coverage (percent of observed abundances that were within the predicted 95% confidence interval) was good for all three of the time-series approaches, generally around 90% (Figure 13). There was little variation in coverage across the different pooling regions for any of the methods, suggesting that regional pooling did not strongly affect whether observed abundances were captured by the confidence intervals around time series model predictions.

The two versions of the stochastic Leslie matrix performed poorly in terms of quantification of uncertainty. Both models considerably underestimated uncertainty, with only around 20% of observed abundances being captured by the predicted 95% confidence intervals (LMN and LUN; Figure 13).

The deterministic Leslie matrix approach does not allow for any quantification of uncertainty, so automatically performed extremely poorly on this criterion.

6.1.8. Criterion 6 – Magnitude of uncertainty

For methods that provide an accurate representation of uncertainty, it is also important to consider how wide the predicted confidence intervals were for each method, because very wide confidence intervals do not necessarily provide useful information within PVA assessments. Of the three time-series methods, the Gompertz model in particular generated extremely wide confidence intervals (*Figure 14*), with the simple growth and Ricker model methods also generating confidence intervals that were very large, detracting from the usefulness of these methods. The width of the confidence intervals for both stochastic Leslie matrix approaches was narrow (*Figure 14*), but since the confidence intervals for these methods had poor coverage this just seems to represent the fact that these methods are failing to accurately capture uncertainty.

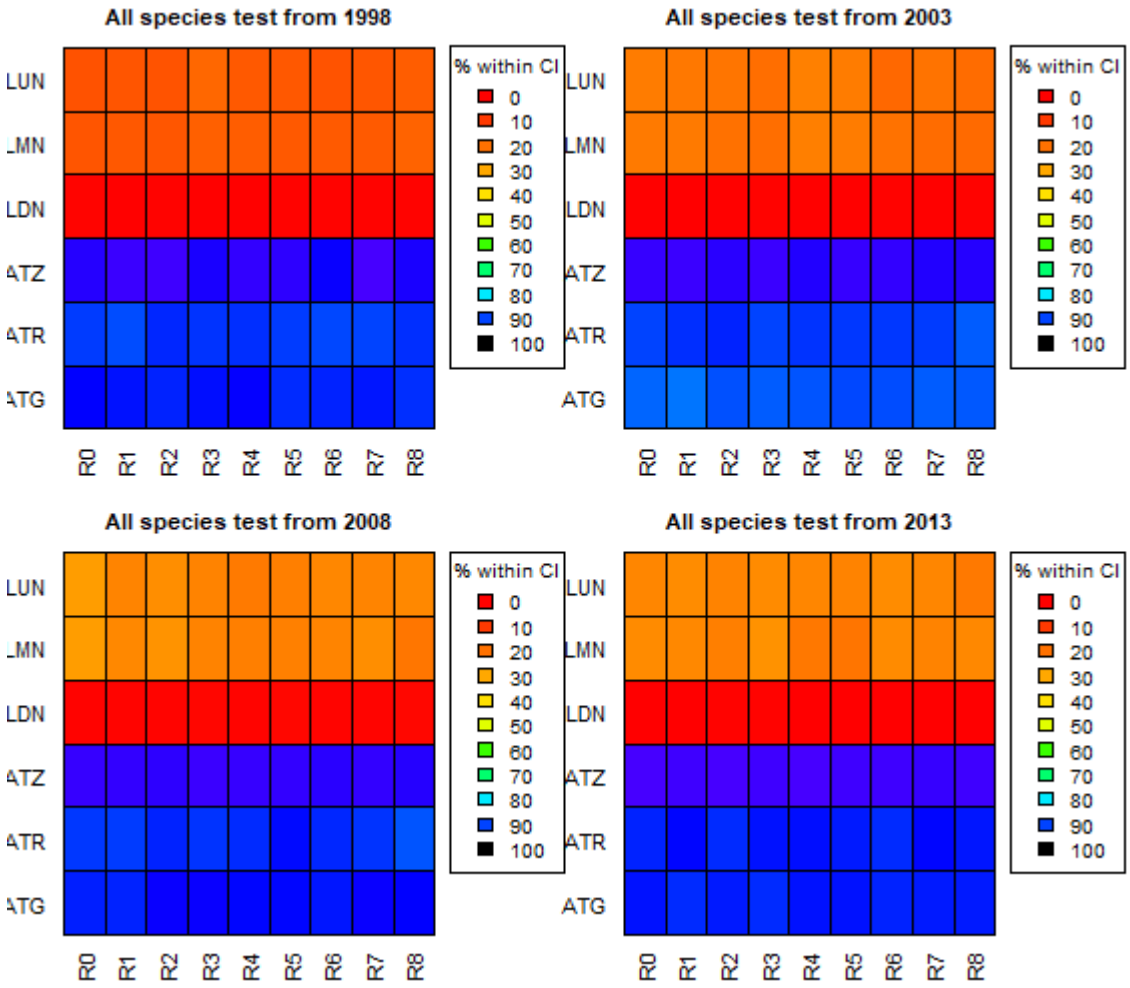


Figure 13: Accuracy of uncertainty quantification, assessed as the percentage of observed abundances that fell within the predicted 95% confidence interval for each model, calculated across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

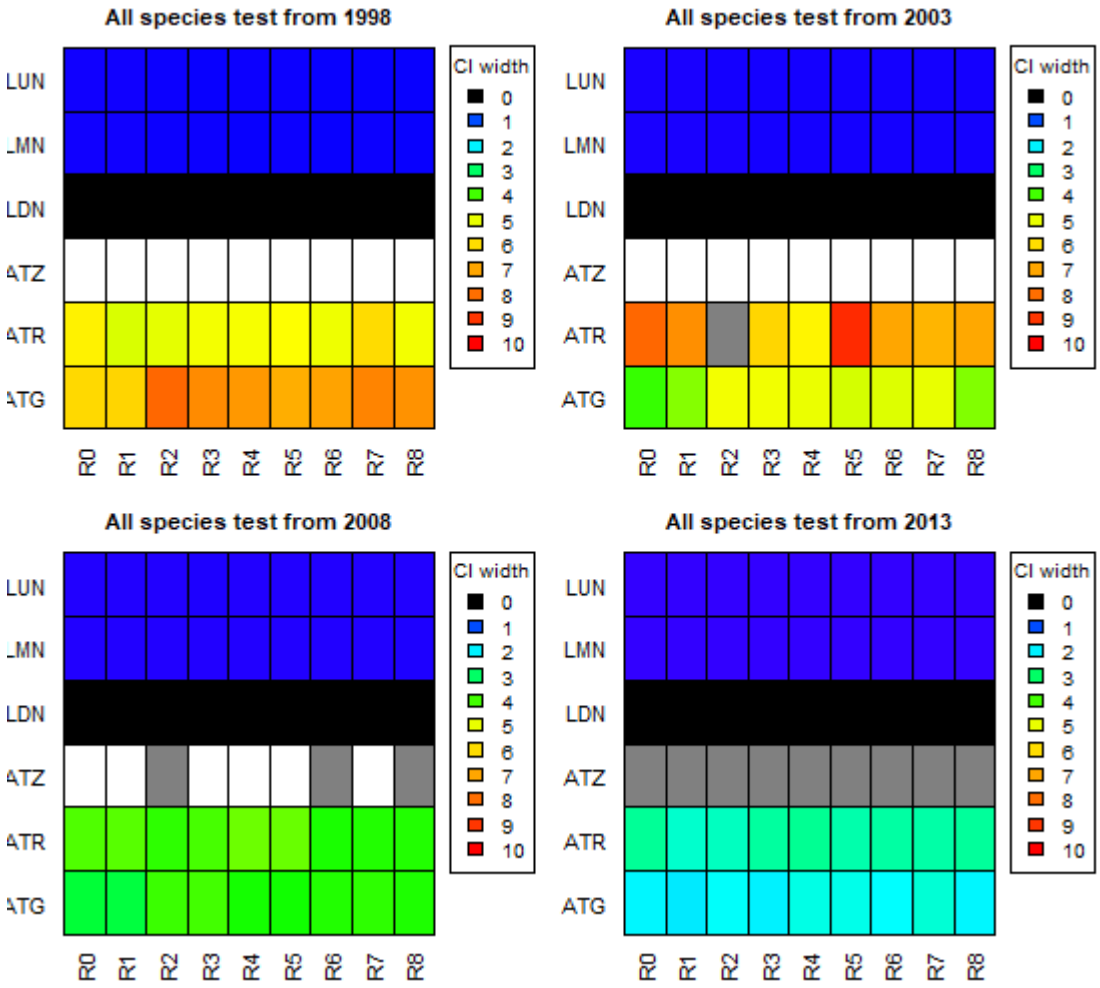


Figure 14: Magnitude of uncertainty, represented as $\log_{10}$(width of 95% confidence interval) for each model, averaged across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Note that white and grey indicate the 95% confidence intervals were so wide as to be outside of the range used in generating the colour legend.

6.1.9. Criterion 7 – Time for computation

The differences in speed between methods were substantial (Figure 15). The slowest methods were the stochastic Leslie Matrix methods, followed by the deterministic Leslie Matrix, and then the three time series methods. However, the computational times associated with the deterministic Leslie Matrix and the time series methods (a few seconds) are likely to be negligible in most practical situations. Note that this comparison does not include the SIPM models from the Forth-Tay region (see next section), which takes a considerable amount of time and expertise to successfully fit in many cases.

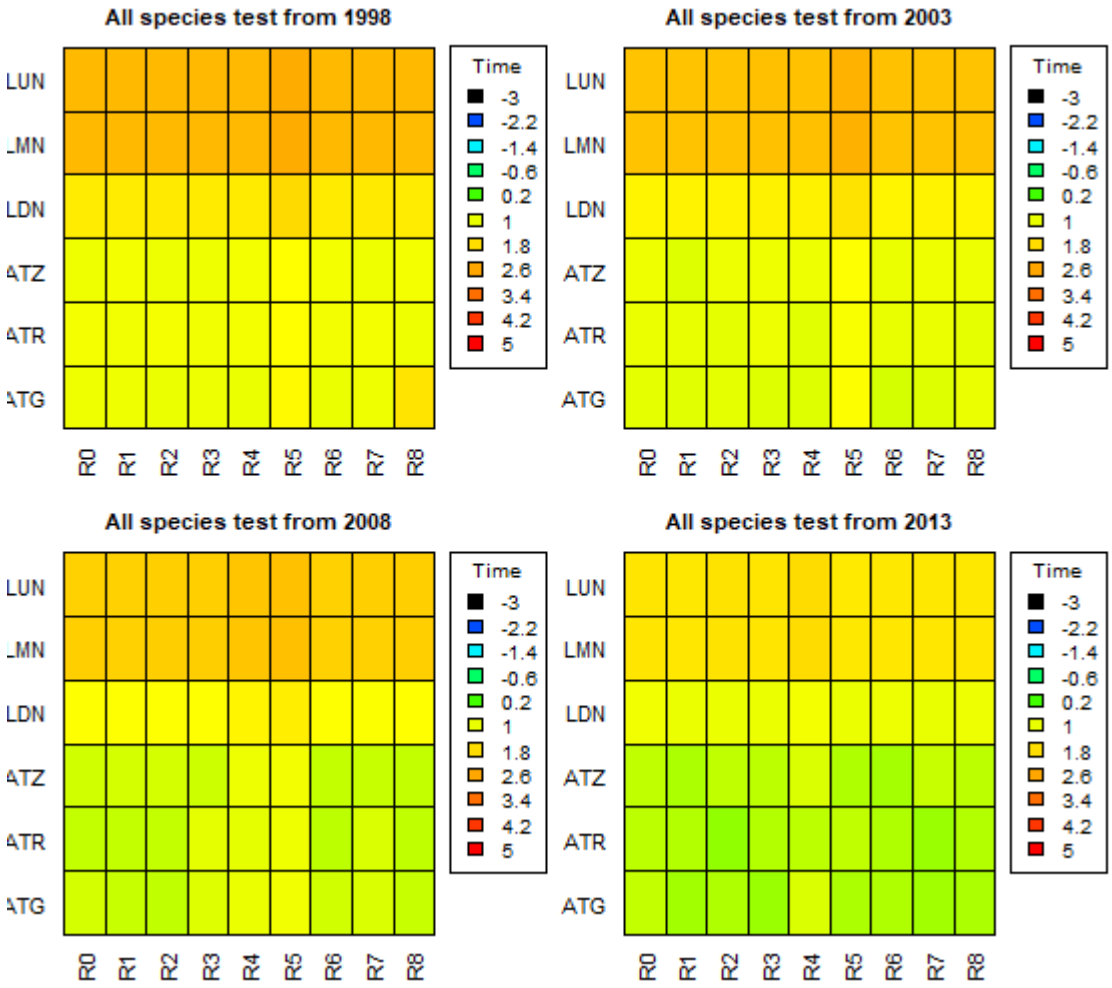


Figure 15: Assessment of time required for computation for each method (in $\log_{10}$ seconds), averaged across the **subset** of species, colonies and years for which all possible methods could be assessed. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

6.1.10. Percentage of occasions where each model performed the best

We consider the percentage of situations (species-colony-year combinations) in which each pooling method and statistical method provided the best performance. We assessed the "best" method as that which minimised the absolute difference between the observed and predicted count. It is possible that these comparisons will be heavily influenced by species-colony combinations with high mean abundance, so we also considered an alternative definition of the "best" method, by determining which method minimised the absolute value of log(predicted count/observed count). However, this assessment led to qualitatively similar results, and so is only included in the Electronic Supplementary Information.

We begin by considering the percentage of times that each combination of statistical method and pooling method provided the best performance. If we consider all possible combinations of species, colony and year then the Leslie matrix models have the highest percentages, because these models can be applied in all situations (Figure 16). However, if we focus only on the situations where all methods can be applied, then the time series models, particularly the Gompertz model, have the highest percentages (Figure 17). Whichever way the evaluation is performed, however, no specific combination of statistical method and pooling method has a high percentage of being the best method. The percentages are typically lower than 5% for all combinations of methods, and almost never exceed 10%. The results of these evaluations should therefore, be interpreted with great caution, as they suggest that the "best" method is highly variable between species, colonies and years.

In order to tease apart the differences in performance between methods, we therefore focus on assessing the percentage of times that each statistical method is best, for each choice of pooling method, and vice versa.

We begin by focusing upon the choice of statistical method. In Figure 18 we focus upon the best choice of method, calculated across all species-colony-year combinations, for each choice of pooling method. The Leslie matrix methods are most often the "best" method, with Leslie matrix methods collectively being the best methods in more than 50% of colony-site-year combinations in almost all circumstances – the one exception is when pooling method R0 was used, and the test period was 2013-2017. In general, the Leslie matrix methods were much more often the best method when regional pooling of any form was used (R1-R8), but were only slightly more often the "best" model in comparison to the time series models when there was no regional pooling (R0).

However, if we focus only on situations in which all methods can be compared, the results look very different. Time series models now collectively perform best in all situations (i.e. for all levels of regional pooling, and all test period definitions), and their advantage is greatest when regional pooling is used (R1-R8). These two sets of results can be interpreted as follows:

- Leslie matrix models can be applied in many more situations than time series models, but;
- In situations where time series models *can* be used, they will often provide the best performance.

Within the time series models, the best performing models tend to be either the Gompertz or simple growth models.

We compared model performance (predicted versus observed abundance) in relation to the definition of the regions used for spatial pooling of demographic information. This is of most relevance to the Leslie matrix approaches, where regional pooling of productivity rates is widely used in practice. This is primarily because Leslie matrix methods with regional pooling of demographic rates can be used in very data-sparse situations where other modelling methods cannot be used. The regional pooling regions we considered (for pooling productivity values and, for the time series models, for imputing abundance values) (Table 0-3) were:

- R0: site level;
- R1: SMP regions;
- R2: ICES regions;
- R3: JNCC regional seas;
- R4: Cook & Robinson Abundance;
- R5: Cook & Robinson Breeding Success;
- R6: MSFD;
- R7: OSPAR;
- R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

The key comparison to consider here is between site-specific methods without any regional pooling (R0), regional pooling methods that use a large number of fairly small spatial regions (R1, which has 114 regions), and regional pooling methods based on a small number of fairly large spatial regions (methods R2-R8, which contain between 1 and 11 regions).

When we consider all combinations of species, colonies and species, there is high variability in the “best” pooling method, with no indication that any method consistently outperforms any other (Figure 20).

When we focus only on combinations for which all methods could be applied, local methods, which avoid much regional pooling (R0 and R1) tend to have the best performance in a higher proportion of situations than the methods that allow for regional pooling. Methods that allow for a higher level of regional pooling (R2 to R8) still, however, collectively account for the “best” method in over 60% of all situation, indicating that there is substantial variability in whether pooling improves performance or not.

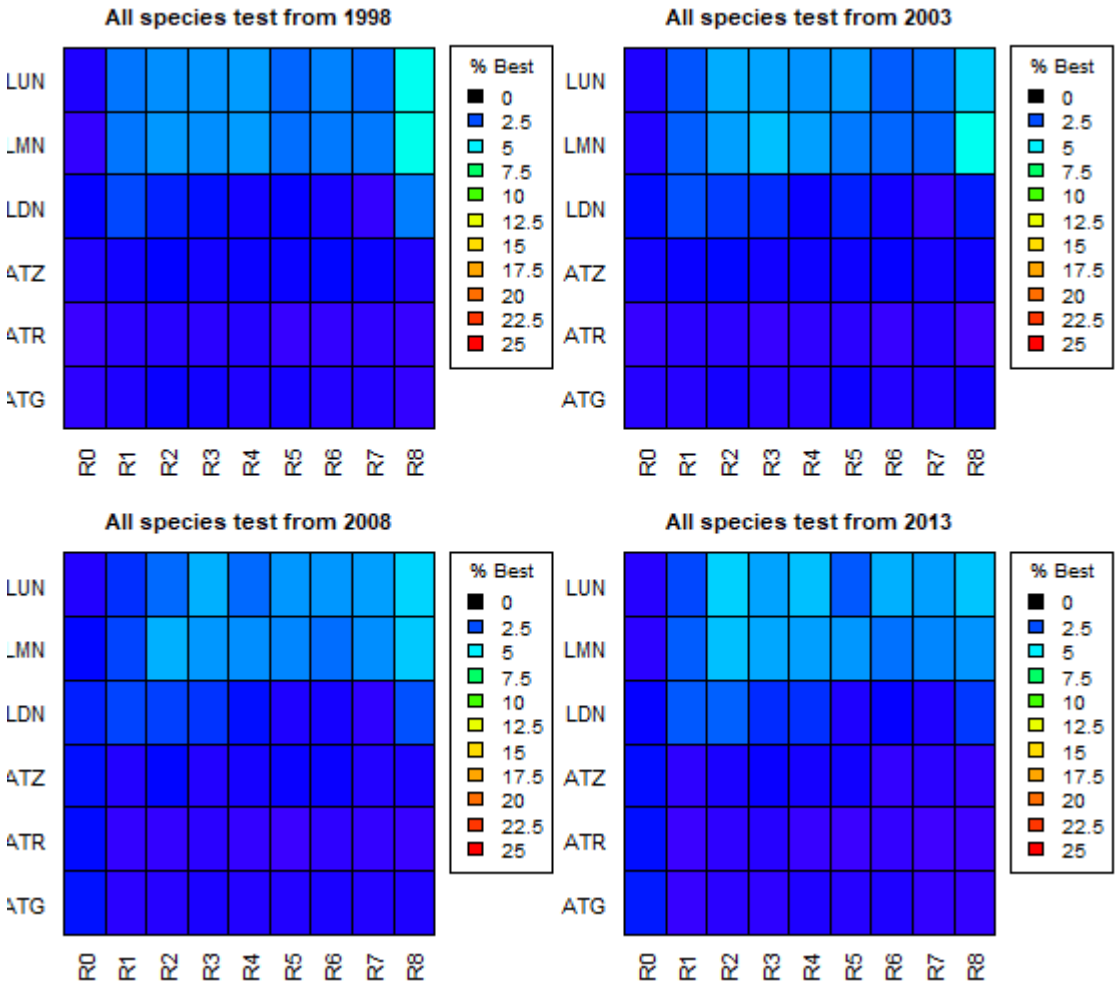


Figure 16: Assessment of which method (combination of statistical modelling method and pooling region classification), performed the best averaged across the full set of all species, colonies and years. Performance was assessed by selecting the percentage of instances in which each method led to a median prediction that was closer (in absolute value) to the observed count than any other method. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

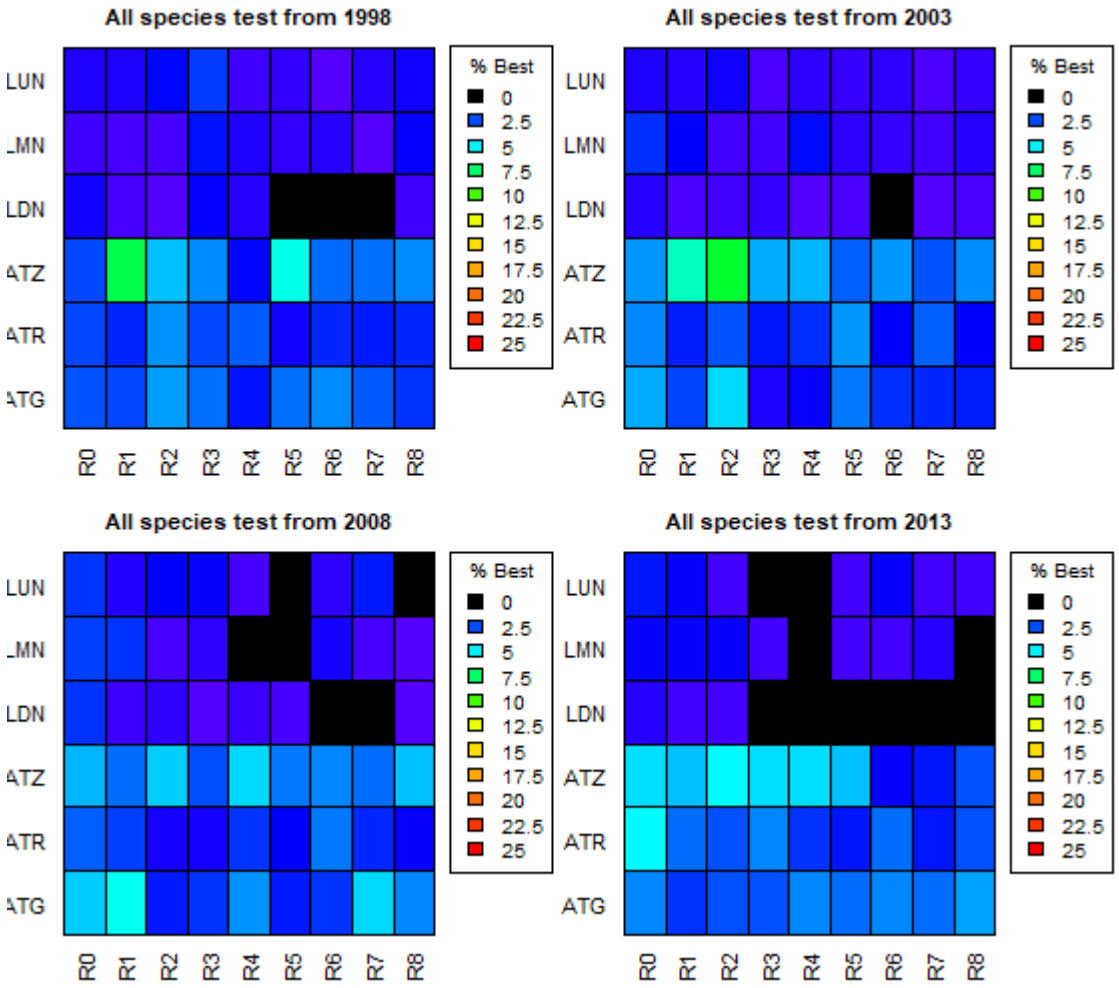


Figure 17: Assessment of which method (combination of statistical modelling method and pooling region classification) performed the best, averaged across only those combinations of species, colony and year for which all possible methods could be applied. Performance was assessed by selecting the percentage of instances in which each method led to a median prediction that was closer (in absolute value) to the observed count than any other method. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Reporting regions were: R0: site level; R1: SMP regions; R2: IC ES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man).

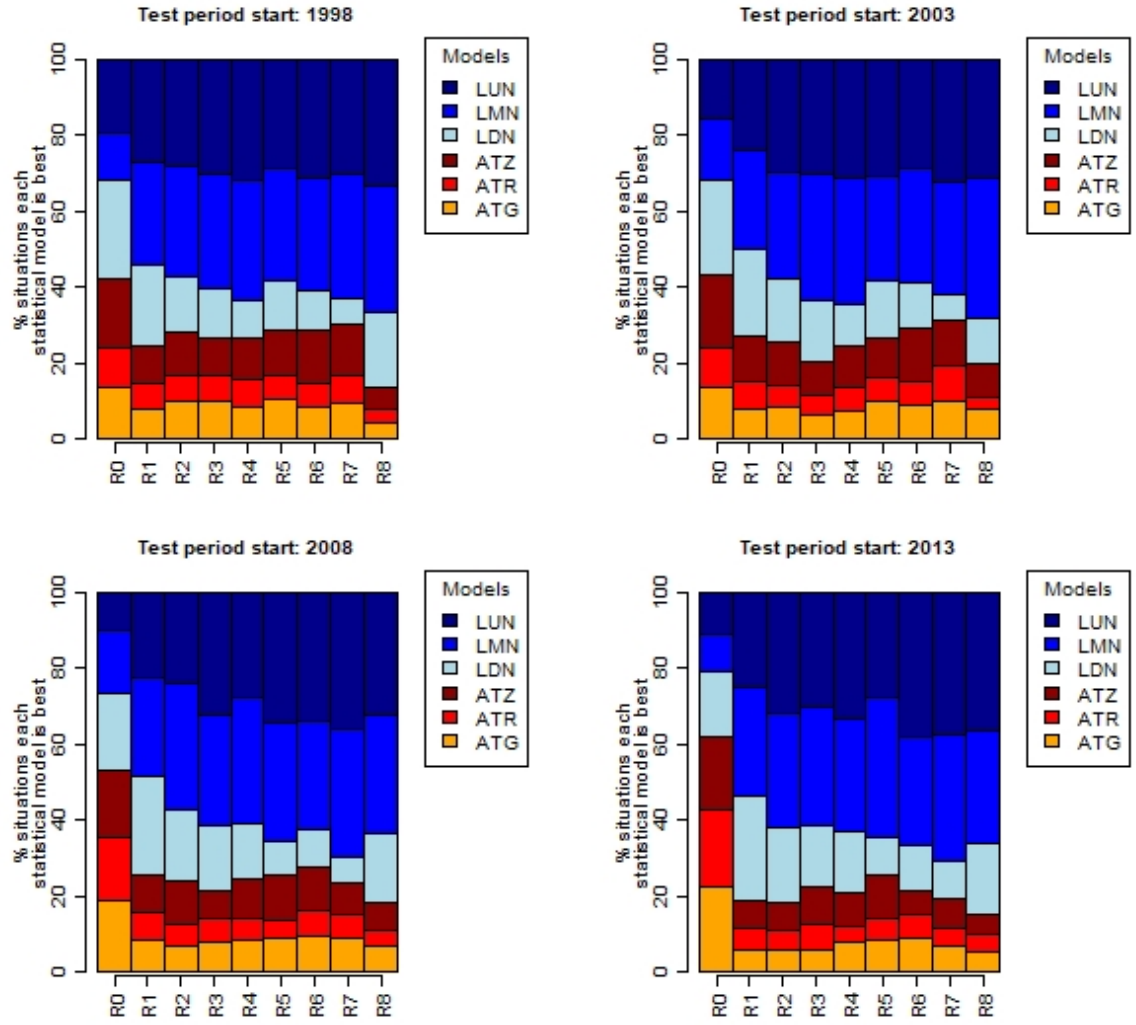


Figure 18: Percentage of species-colony-year combinations for which each statistical method has the best performance (in terms of having the lowest absolute difference between the median predicted value and the observed count), calculated separately for each pooling method and each test period definition. Percentages were calculated using all species, colony, year combinations.

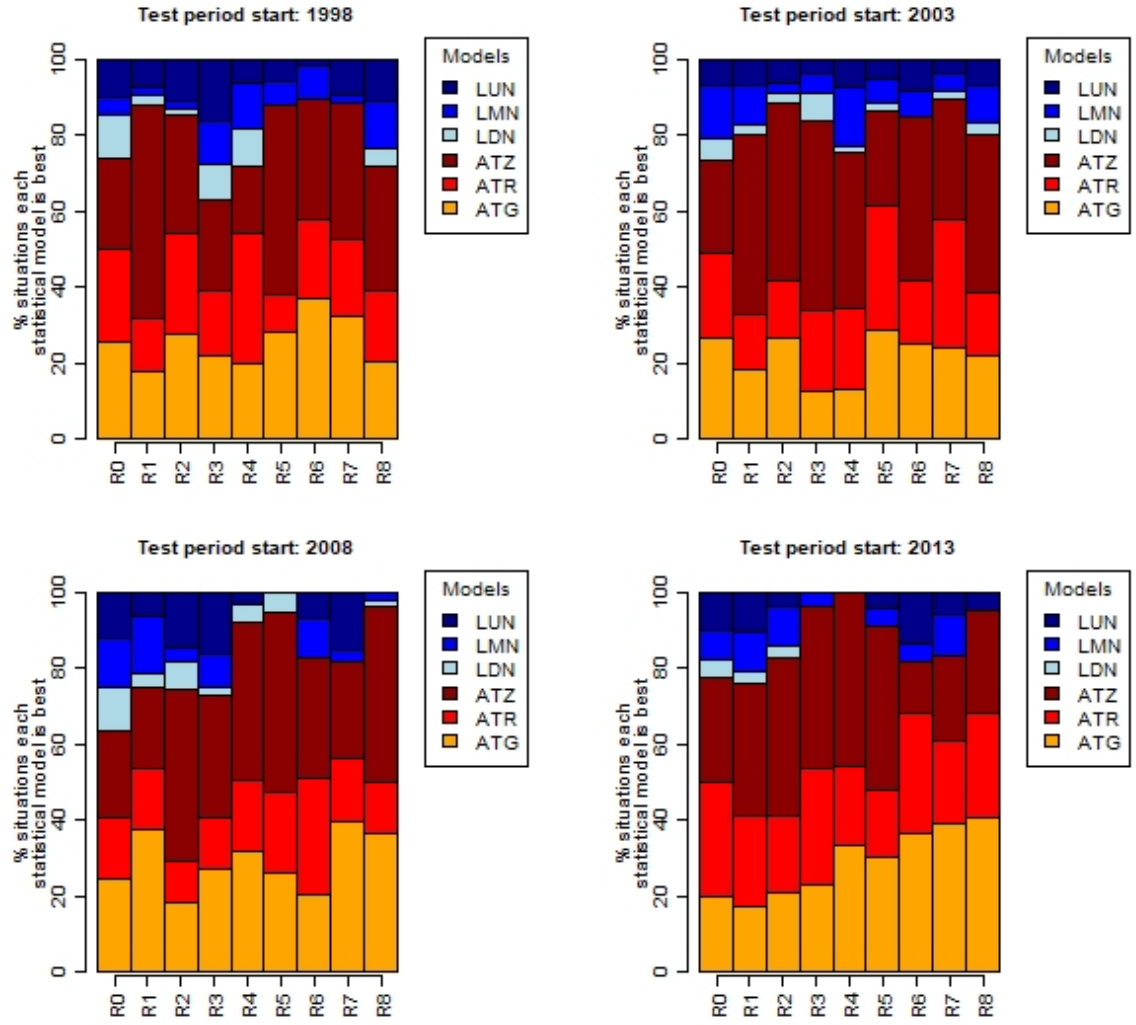


Figure 19: Percentage of species-colony-year combinations for which each statistical method has the best performance (in terms of having the lowest absolute difference between the median predicted value and the observed count), calculated separately for each pooling method and each test period definition. Percentages were either calculated across the **subset** of species, colonies and years for which all possible methods could be assessed.

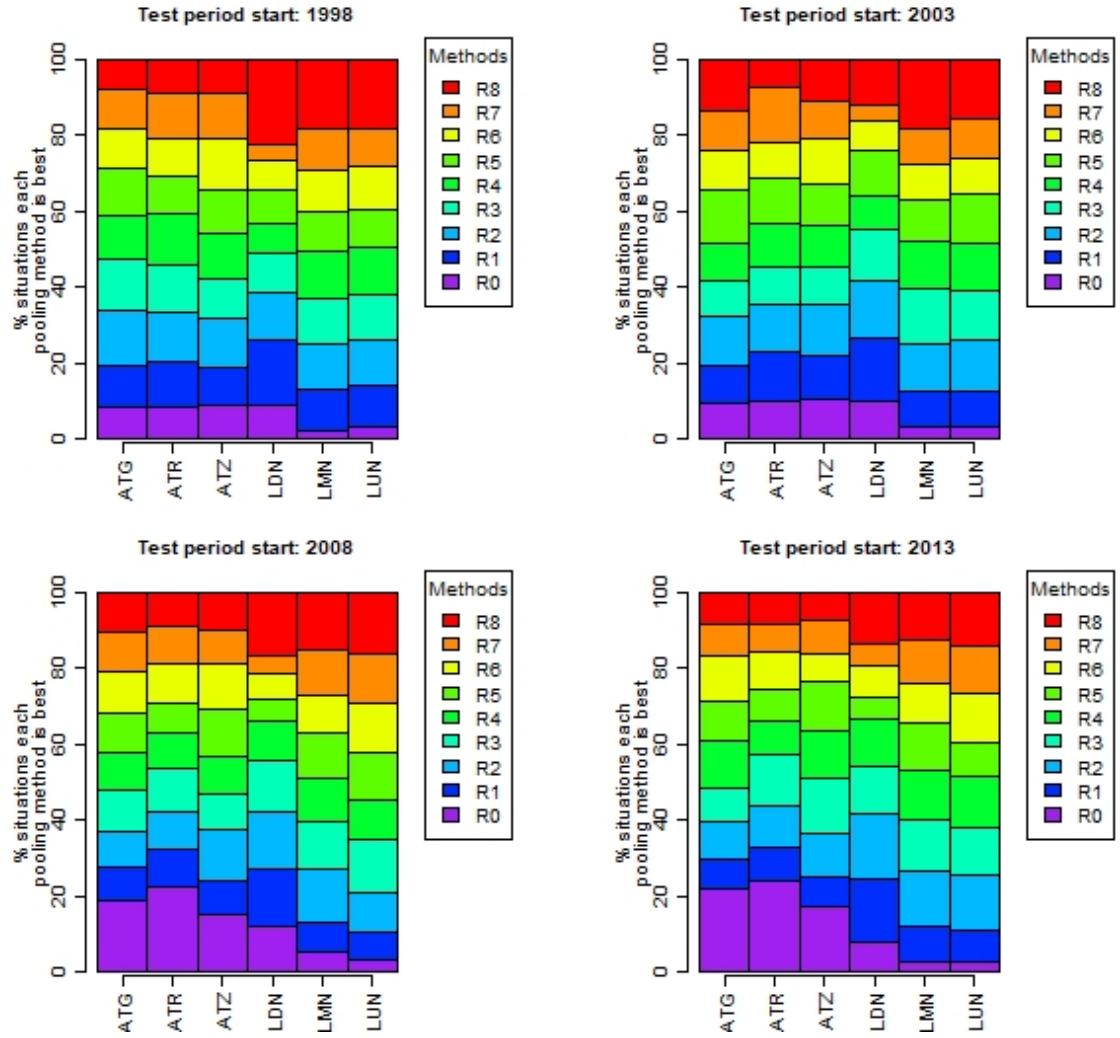


Figure 20: Percentage of species-colony-year combinations for which each pooling method has the best performance (in terms of having the lowest absolute difference between the median predicted value and the observed count), calculated separately for each statistical method and each test period definition. Percentages were calculated using all species, colony, year combinations.

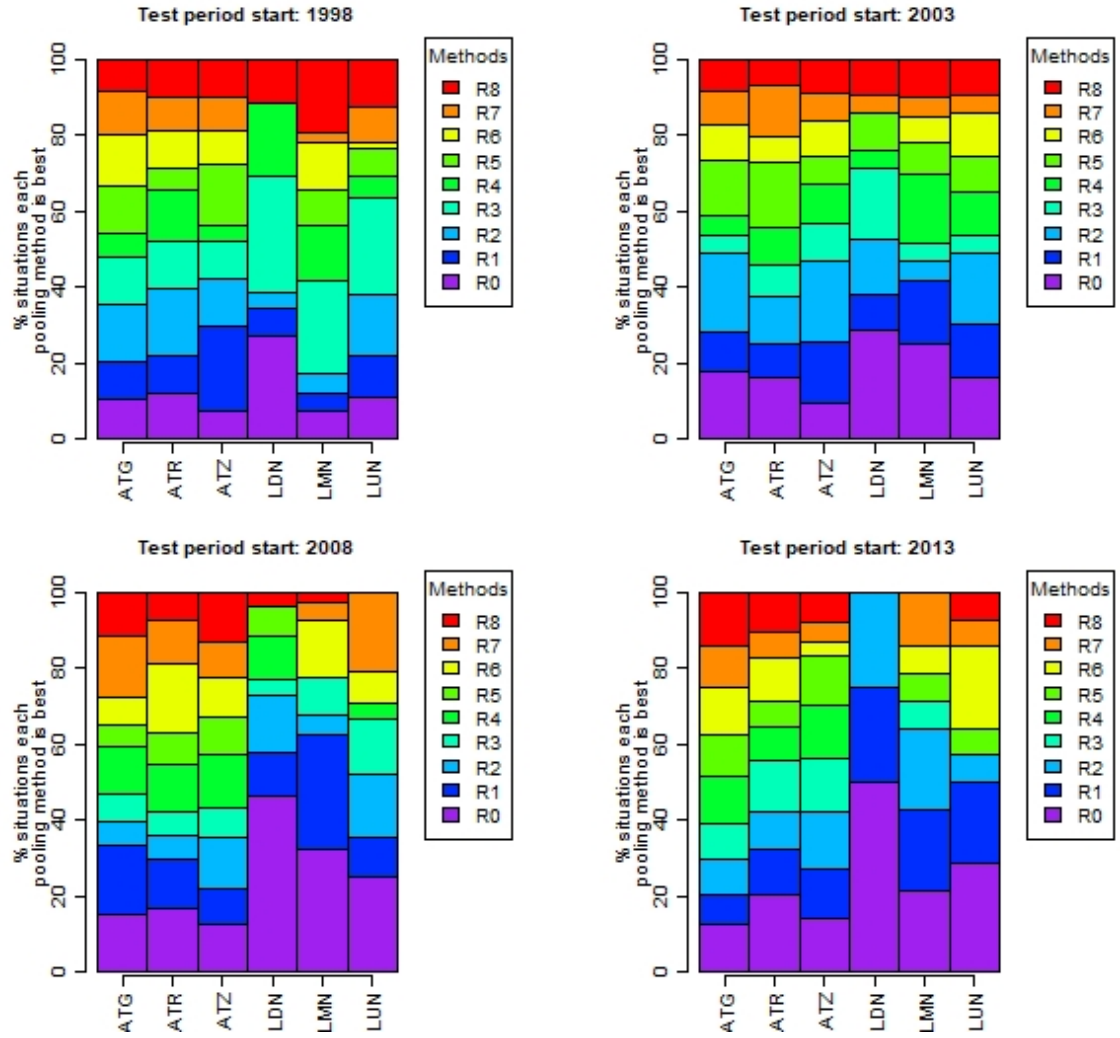


Figure 21: Percentage of species-colony-year combinations for which each pooling method has the best performance (in terms of having the lowest absolute difference between the median predicted value and the observed count), calculated separately for each statistical method and each test period definition. Percentages were calculated across the **subset** of species, colonies and years for which all possible methods could be assessed.

6.1.11. Relation to time since last training period

We examined if performance of the different modelling methods was related to the gap between the year in which evaluation occurred (i.e., the year with the observed count within the test period against which the prediction was compared) and the year used to initialise the model (i.e., the year with the latest observed count within the training period at the colony of interest). This was because we expected levels of error and bias to increase as the length of this gap increased, and also to investigate if there was an evidence that some of the modelling methods were more robust to longer gaps between the year of initialisation of the model and the subsequent year in which the comparison for observed versus predicted was made.

The percentage of highly implausible results (more than or less than 100 times the observed abundance) increased as the length of this gap increased (Figure 22**:**).

The level of bias in predictions also increased as the length of the gap between the last year of data used in model fitting or initialisation and the year in which comparisons of predicted versus observed abundance were made (Figure 23). Bias was generally low across all modelling methods when gaps were very short (1-3 years), however as gaps increased, bias became much more pronounced, with most methods tending to overestimate the observed abundance (Figure 23). This increase in bias with increasing gap width was most pronounced for the time series simple growth model, which increasingly overestimated observed abundance as the length of the gap exceeded around five years (Figure 23). Of all the modelling methods, the time series Ricker model appeared to be the most robust to maintaining low levels of bias as the length of the gap increased, followed by the stochastic Leslie matrix methods (Figure 23).

In terms of the overall magnitude of error, all methods showed an increase in error as the length of the gap increased. For most methods, error started to increase significantly as the gap reached over five years, with the three Leslie matrix methods reaching the highest levels of absolute error across all modelling methods once the gap had exceeded more than ten years. The notable exception was the time series Ricker model, which maintained relatively lower levels of error with increasing gap length up to around ten years, when compared to the other modelling methods.

There was relatively little impact of the length of the gap on the percentage of observed abundances that were within the predicted 95% confidence interval for each method (Figure 25). For the time series methods this was likely because the width of the confidence intervals increased greatly as the length of the gap increased

(Figure 26), meaning that the percentage of times the observed abundance fell within the predicted confidence interval remained high at around 80-90% for all three time series methods (Figure 25). For the Leslie matrix methods, this was because the length of the gap had very little impact on the width of the confidence intervals (Figure 26), meaning that the percentage of times the observed abundance was within the predicted confidence interval remained low (around 30%) regardless of the length of the gap (Figure 25).

The percentage of times each model was the best performing model showed no obvious relationship with the length of period (Figure 27).

The relationships between these assessment criteria and the length of the gap appeared to be fairly similar, regardless of the definition of the test period. This suggests that it was the data gap between initialisation and testing at the colony of interest, rather than the overall breakdown of the data into training and test periods, that was of key relevance in determining levels of accuracy, bias and coverage.

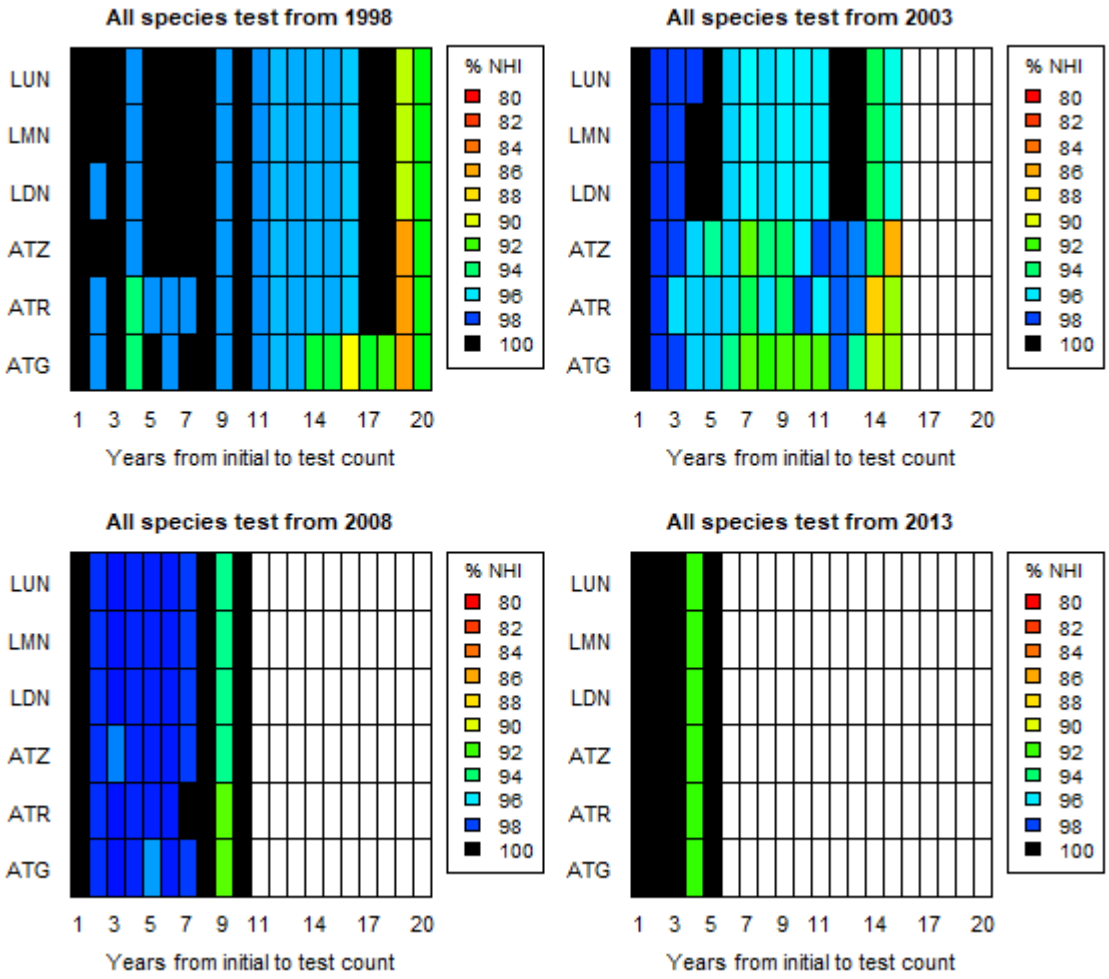


Figure 22: Percentage of median predicted population sizes that are "not highly implausible", averaged across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained.

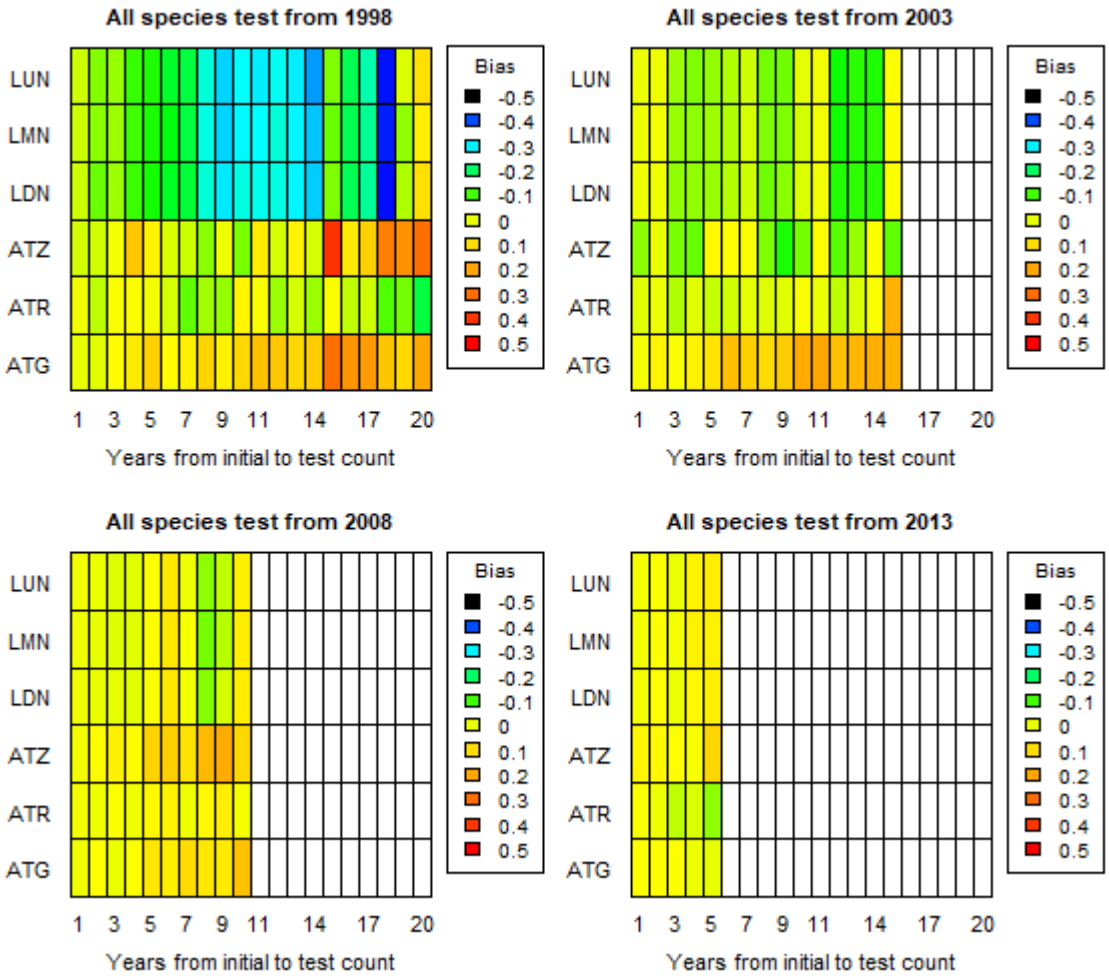


Figure 23: Median bias, averaged across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained.

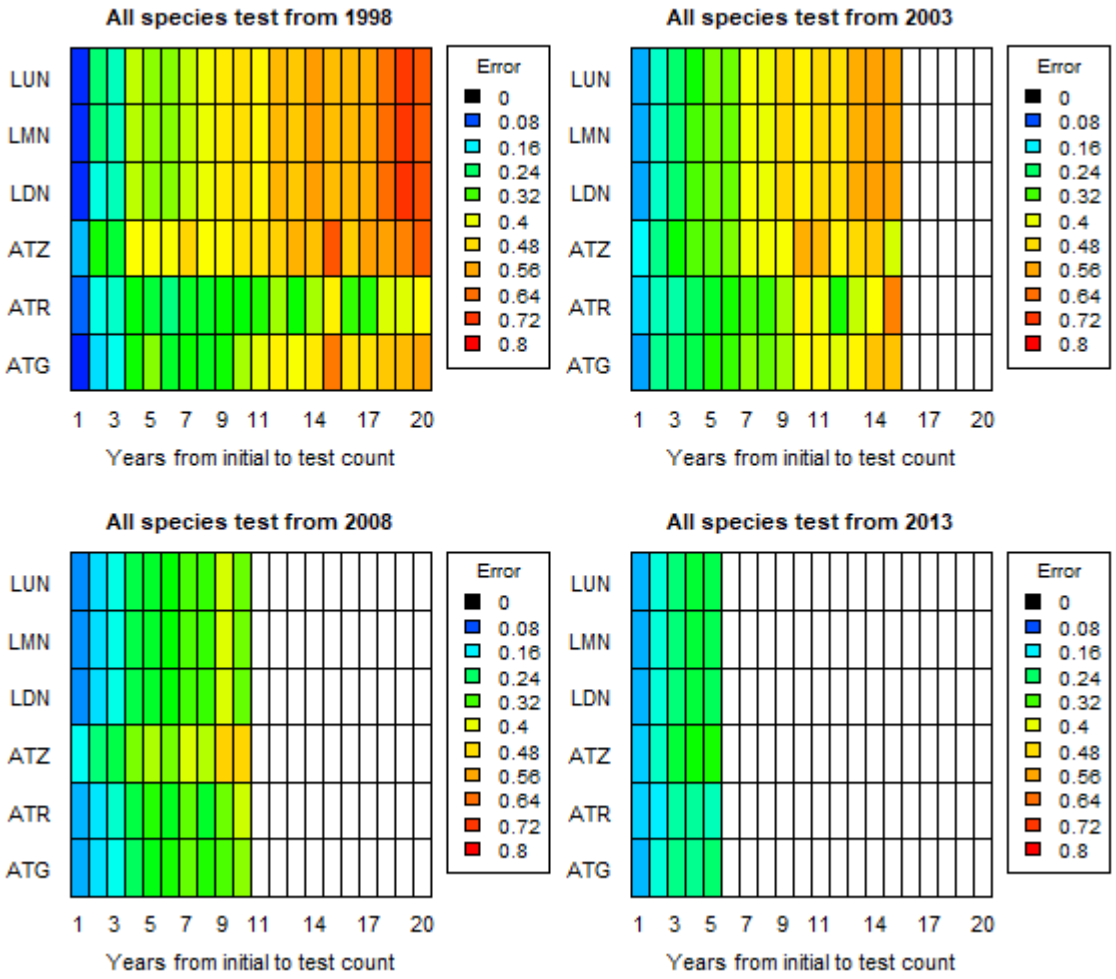


Figure 24: Median error, averaged across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained.

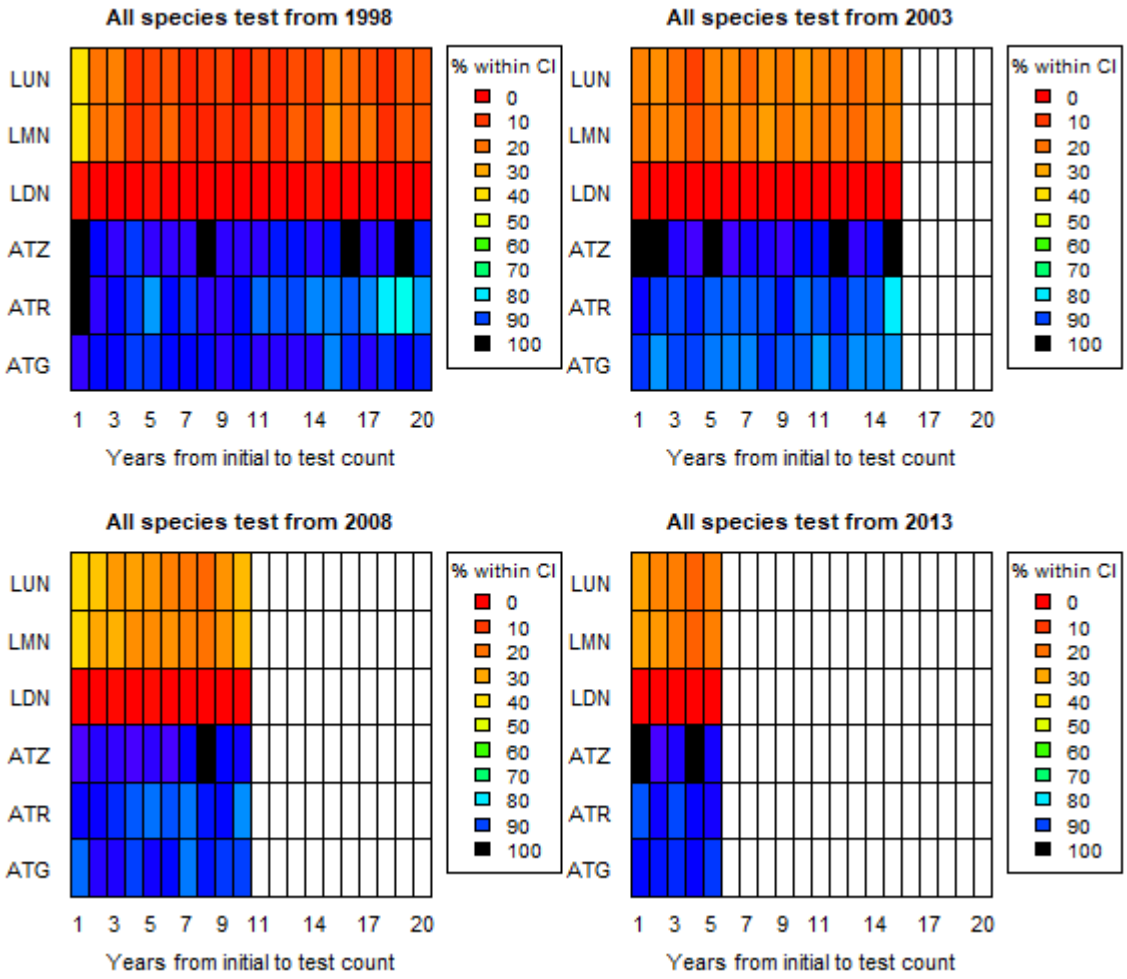


Figure 25: Percentage of times the observed abundance was within the 95% confidence interval, calculated across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained.

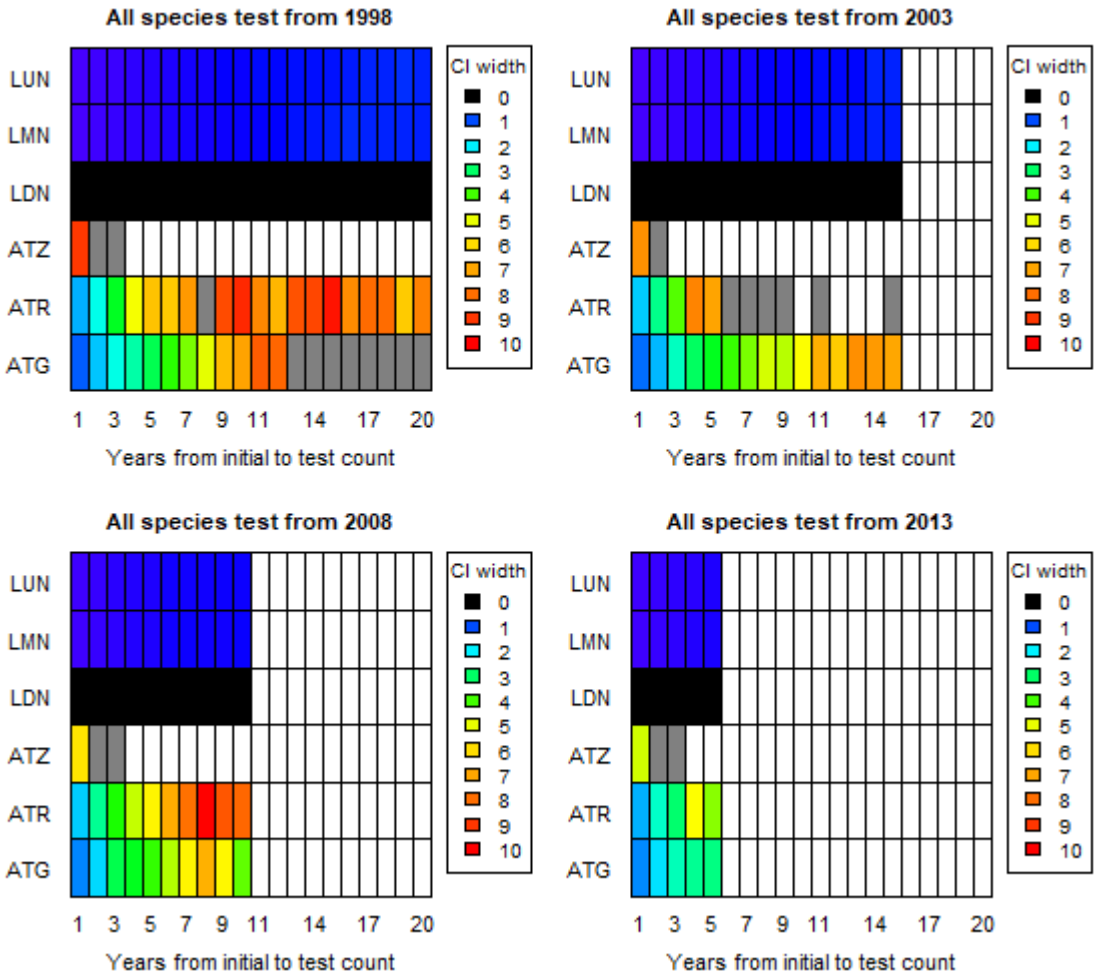
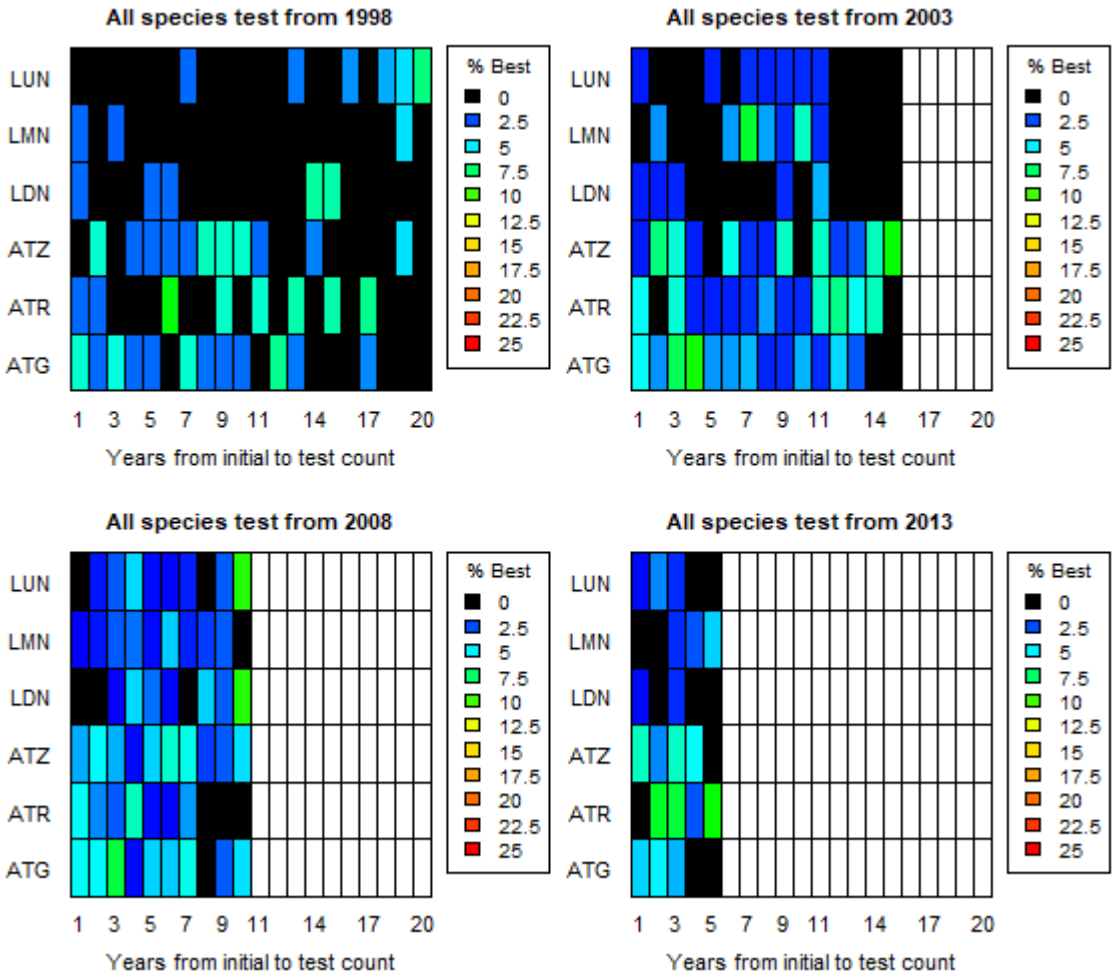


Figure 26: Median $\log_{10}$(width of confidence interval), calculated across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained. Note that white and grey indicate the 95% confidence intervals were so wide as to be outside of the range used in generating the colour legend. Note that white and grey indicate the 95% confidence intervals were so wide as to be outside of the range used in generating the colour legend.



Figure

Figure 27: Assessment of which method (combination of statistical modelling method and pooling region classification) performed the best, calculated across the **subset** of species, colonies and years for which all possible methods could be assessed, expressed in relation the number of years between the evaluation year and initial count for one particular regional pooling classification (site-level, R0). Performance was assessed by selecting the percentage of instances in which each method led to a median prediction that was closer (in absolute value) to the observed count than any other method. Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; LDN: Leslie Matrix deterministic model; LMN: Leslie Matrix Stochastic model with productivity constrained; LUN: Leslie Matrix Stochastic model with productivity unconstrained.

6.2. Comparison for Forth-Tay SPAs

SIPMs were produced by Freeman et al. (2014) for 14 populations (species-by-SPA combinations) in the Forth-Tay region – these populations are listed in

Table ***6-1***, and the number of years for which an evaluation of performance was possible within the test period (2013-2017) is shown for each population. In total, evaluation was possible for 35 of the 70 (i.e. 50%) of the possible species-SPA-year combinations, but there was considerable variation between populations. For five populations an evaluation was possible in all five years, for two populations it was possible in two of the five years, for six populations it was only possible in a single year, and for one population (Herring Gull – Forth Islands) it was not possible in any year.

Table 6-1

Populations used in the Forth-Tay evaluation, and the set of years within the test period for which an evaluation of performance was possible (i.e. for which an SPA-wide count of abundance was available).

Species	SPA	Evaluation possible?					# years of evaluation
		2013	2014	2015	2016	2017	
Black-legged kittiwake	Forth Islands	Yes	Yes	Yes	Yes	Yes	5
	St Abbs to Fast Castle	Yes	Yes	Yes	Yes	Yes	5
	Fowlsheugh			Yes			1
	Buchan Ness to Collieston				Yes		1
Common guillemot	Forth Islands	Yes	Yes	Yes	Yes	Yes	5
	St Abbs to Fast Castle	Yes			Yes		2
	Fowlsheugh			Yes			1
	Buchan Ness to Collieston				Yes		1
Razorbill	Forth Islands	Yes	Yes	Yes	Yes	Yes	5
	St Abbs to Fast Castle	Yes			Yes		2
	Fowlsheugh			Yes			1
Herring Gull	Forth Islands						0
	St Abbs to Fast Castle	Yes	Yes	Yes	Yes	Yes	5
Atlantic Puffin	Forth Islands	Yes					1

We summarised performance for all modelling methods in the Forth-Tay comparison, across all combinations of statistical method and regional pooling method, by averaging across all species-colony-year combinations using the testing period starting in 2013 (Figure 28 and

Figure *29*).

6.2.1. Criterion 1 – Ability to run

Within the Forth-Tay region (R9), the SIPM could be run in all circumstances, as could the various Leslie Matrix methods, however the three time-series methods could only be run in around 80-90% of instances (Figure 28). The fact that the SIPM can be run in all situations is true by definition, as we only focused here upon the set of SPAs for which an SIPM had already been run, and upon a region (the Forth-Tay) where there was sufficient data to be able to fit these models for a range of species and SPAs. Therefore, this result would not generalise to other regions or species. In general, we might expect that SIPMs can be applied in a similar set of situations to those in which time series models can be applied, since the key restriction to their use, in terms of data availability, is the fact that they can only be applied to populations for which data on abundance are collected relatively frequently.

6.2.2. Criterion 2 - Occurrence of highly implausible results

The occurrence of highly implausible results (more than 100 times above or below the observed abundance) was very low for all methods in the Forth-Tay region (R9). When using the mean predicted abundance, the SIPM, all Leslie Matrix methods and the simple growth time series method resulted in 100% of predictions that were not highly implausible, with the Ricker model resulting in around 95% of predictions that were not highly implausible (Figure 28). When using the median predicted abundance, all methods resulted in 100% of results that were not highly implausible (not shown).

6.2.3. Criterion 3 - Systematic bias

Systematic bias was very low for all modelling methods in the Forth-Tay region, although marginally higher for the Gompertz time series method compared to the other modelling approaches (Figure 28).

6.2.4. Criterion 4 – Error

Similarly, all modelling methods resulted in low error in the Forth-Tay region (R9), again with the exception of the Gompertz time-series method which resulted in considerably more error than any of the other methods (Figure 28).

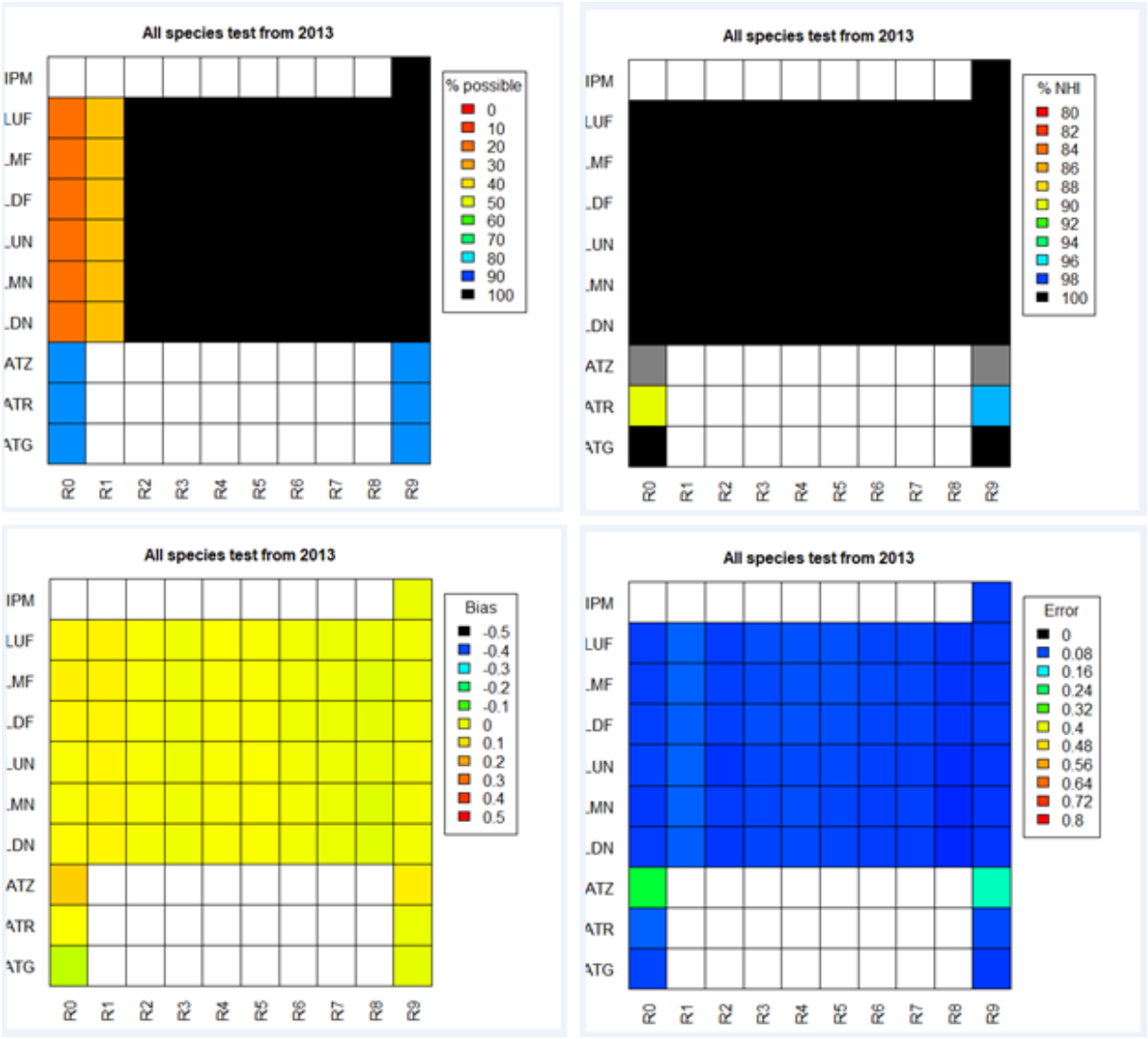


Figure 28: Summary of results for the Forth-Tay comparison for all modelling methods for Assessment Criterion 1 to 4. Criterion 1 is the percent of occasions where it was possible to use the modelling method to generate a prediction (top left panel). Criterion 2 is the occurrence of highly implausible results estimated using the mean predicted abundance from each modelling method (top right panel). Criterion 3 is the level of bias in model predictions using the median predicted abundance from each modelling method (lower left panel). Criterion 4 is the magnitude of error in model predictions using the median predicted abundance from each modelling method (lower right panel). Each criterion is produced by averaging over all species by SPA comparisons that were possible within the Forth-Tay region, using a model testing period starting in 2013. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix deterministic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

6.2.5. Criterion 5 - Quantification of uncertainty

The SIPM method and time series simple growth and Ricker models all performed well in terms of the percentage of observed abundances that were within the 95% predicted confidence intervals (around 90%; Figure 29), whilst the Gompertz time series method achieved 100% of observed abundances within the predicted confidence intervals (Figure 29).

The Leslie matrix methods performed more poorly, with around 60% of observed abundances falling within the predicted 95% confidence intervals, indicating that these methods may severely underestimate uncertainty (Figure 29).

6.2.6. Criterion 6 – Magnitude of uncertainty

Of the methods that produced accurate representations of uncertainty, the SIPM produced substantially narrower 95% credible intervals than the 95% confidence intervals produced by the time series methods, although the credible/confidence intervals associated with these methods were all very wide. The Gompertz model produced the widest intervals of all.

6.2.7. Criterion 7 - Computational time

Computation time was similar over all modelling methods apart from the SIPM, being lowest for the time series simple growth model, followed by the other time series methods and deterministic Leslie Matrix methods, with the stochastic Leslie Matrix methods taking the longest computationally. The SIPM models take considerably longer to fit than any of the other methods. We did not quantify this explicitly in this project, because we did not refit the SIPMs, but, broadly speaking, the SIPMs take a few hours to run for a single population, whereas the other methods can all be run for a single population within a few seconds. The length of computation time required to fit SIPMs to any particular population will vary considerably depending upon the complexity and amount of data used in model fitting. Time will also depend upon the specification of the computer being used (and, clearly the computational time required to fit the models today will be lower than that required when the models were originally fit in 2014, due to advances in processing power).

6.2.8. Percentage of situations in which each model had “best” predictions

In the Forth-Tay comparison (R9), the SIPM method was the “best” model, in terms of most accurate median prediction for abundance, around 25% of the time (Figure 29). This method was followed by the time series simple growth and Ricker modelling methods, which both performed best in around 5% of comparisons (Figure 29). The stochastic Leslie Matrix with unconstrained productivity (LUF) was the best method in just 2.5% of all comparisons in the Forth-Tay, whilst the other Leslie Matrix methods were never the best performing model

(Figure 29). The time series Gompertz method also failed to be the best model in any comparison (Figure 29).

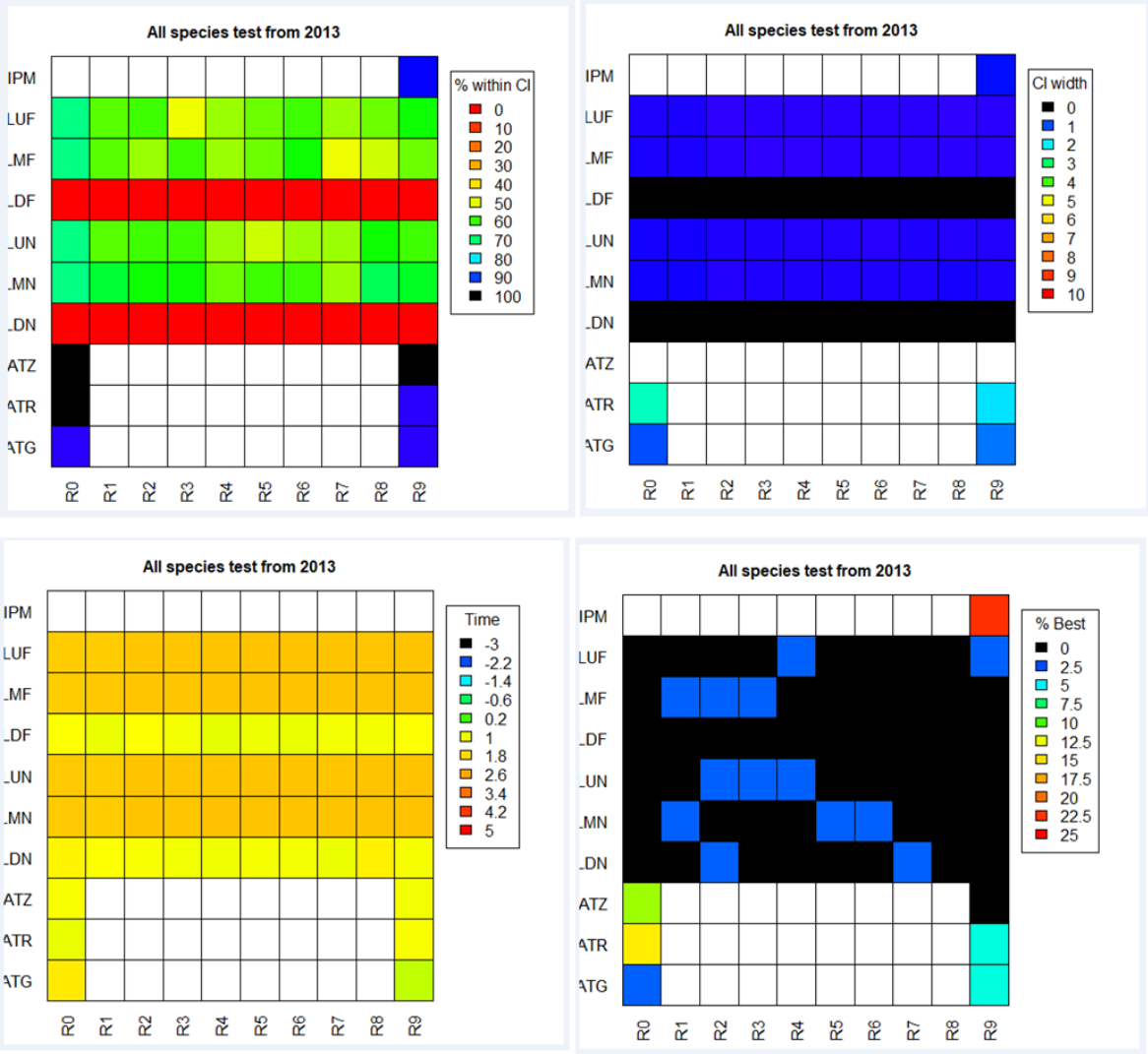


Figure 29. Summary of results for the Forth-Tay comparison for all modelling methods for Assessment Criterion 5 to 6. Criterion 5 is the assessment of uncertainty for each modelling method (top left panel: percentage of observations that were within the 95% confidence interval from the model; top right panel: width of 95% confidence interval from the method). Criterion 6 is the time required for computation for each of the modelling methods (bottom left panel). Overall performance is also shown by calculating the percentage of times in which each modelling method resulted in the best abundance prediction, comparing the median prediction to the observed abundance (lower right panel). Each criterion is produced by averaging over all species by SPA comparisons that were possible within the Forth-Tay region, using a model testing period starting in 2013. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix determinisitic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

6.2.9. Population-specific results

Detailed results of performance for specific populations are given in Appendix B. Overall, the predicted abundance values produced by SIPM were substantially closer to observed counts of abundance for two populations – Herring Gull at Forth Islands, and kittiwakes at Buchan Ness and Collieston Coast – although in the latter case Leslie matrix approaches performed similarly to the SIPMs when high levels of regional spatial pooling were used. The poor performance of the Leslie matrix models for these populations probably arises from limited availability of relevant demographic data. SIPMs performed worse than Leslie matrix methods, in terms of the accuracy of predictions, for three populations – razorbill and kittiwake at Fowlsheugh, and puffin at Forth Islands. These three populations all had small amounts of abundance data in the training period – too little to meet the minimum data requirements we imposed for the use of time series models – which probably explains the poor performance of the SIPMs in these situations. For the remaining populations the different methods showed relatively similar performance, in terms of the accuracy of the median/mean predictions produced.

For all populations the levels of uncertainty produced by different methods showed a consistent pattern: uncertainties associated with SIPMs were substantially higher than those produced by Leslie matrix models, and uncertainties associated with time series models were substantially higher than those produced by SIPMs.

7. Discussion

7.1. Summary of key empirical findings

The overall performance of each modelling method, was extremely variable, and overall differences in performance between methods were relatively subtle (with one important exception: the underestimation of uncertainty by stochastic Leslie matrix methods, which was a strong and consistent finding). All methods performed well in some situations, in terms of the discrepancy between the predicted and observed abundance within the test period, and poorly in others, and it is not clear from the results which factors are driving these differences in performance. All methods have some situations in which they perform extremely poorly. For stochastic methods (Leslie Matrix and time-series methods), it was often the case that a handful of simulations produced extremely implausible results. Using the median across simulations, rather than mean, to estimate abundance helps to eliminate some of these extremely implausible results.

Despite the considerable variations been populations, and the various caveats associated with our assessments (Section 7.2), our comparisons did reveal a number of consistent findings, and we focus upon these here. We focus primarily upon the findings of the national evaluation, as the small number of species, populations and test years considered within the Forth-Tay evaluation mean that the results of this comparison should be interpreted with substantial caution.

The two most consistent, and unexpected, findings, of our comparisons (both nationally and in the Forth-Tay region) were that:

1. Deterministic and stochastic Leslie matrix models often produced less accurate predictions than time series models, even the very simple, and biologically implausible, simple growth model.

2. Stochastic Leslie matrix models seemed to systematically underestimate uncertainty, which led to low percentages for the observed abundance falling within the predicted 95% confidence interval. By contrast, the time series models (simple growth, Ricker and Gompertz), and, where applicable, SIPMs, provided an assessment of uncertainty that avoided the underestimation seen with the stochastic Leslie matrix models. Therefore, for these methods, the confidence or credible intervals did tend to include the observed count in approximately 95% or more or the time.

In addition, there were a number of other consistent findings, which reinforced existing knowledge:

a. In terms of overall ability to apply different modelling methods, the Leslie matrix methods that used pooled demographic rates could be applied to the highest proportion of populations. Leslie matrix methods that used local empirical rates, or time series models, could only be applied for the minority of colony by species combinations where there was extensive local data for either demographic rates or abundance counts.
b. The percentage of highly implausible results (i.e. the percentage of results for which predicted abundance was more than or less than 100 times the observed abundance), the level of bias and the level of error all tended to systematically increase as the length of the gap between the last year used for model fitting or for model initialisation and the year in which the comparison between observed and predicted abundance was made increased.

More specifically, in the context of Leslie matrix models:

a. Pooling of information for productivity rates seemed to have relatively little overall impact on model performance, although methods that avoided regional pooling (R0) generally had slightly higher performance than methods that used it, in situations where local data on breeding success existed for the colony of interest.
b. Methods that used a large number of fairly small regions for regional pooling could be used in fewer situations than methods that used a small number of fairly large regions, but had higher performance in situations where they could be used.

In terms of time required for computation, it was possible to run all of the modelling methods that we considered here, except the SIPMs, relatively quickly – within a few seconds for a single population. Note that this relied on efficient programming of the algorithms, and a simpler implementation of the stochastic Leslie matrix models in R using loops would have been much slower. The SIPMs are much slower to set up and run than any of the other modelling methods, in terms of both computer time and staff time.

7.2. Caveats, limitations and further work

Within this study, we compared a large number of population modelling methods (91), commonly used in PVAs, for a relatively large number of species (15), across four different test-training period splits, across a wide range of colonies and years. It has necessarily only been possible within this report to focus upon key overall summaries of the results, and it has not been feasible to investigate in detail all of the specific situations (species, geographical areas and years) in which particular methods performed well or badly. The raw outputs of the comparisons for individual species-colony-year-method combination are made available as part of the Electronic Supplementary Information, in order to allow stakeholders to investigate the performance of the methods in specific situations in more detail if this is of interest.

In summarising the results, we averaged across colonies, years and, where relevant, species. The set of species-colony-year combinations for which evaluations of method performance were possible was a fairly small, and not necessarily representative, proportion of the set of all species-colony-year combinations. Therefore, the results may be influenced by patterns of data availability, and the comparisons between methods will partly be confounded by differences between the species, colonies and years in data availability. It would be worthwhile to explore this issue in more detail, and to attempt to adjust for this where possible (e.g., through the use of weighting when combining results, to try and improve the generalisability of the results). However, fundamentally, this limitation arises from the availability of data on abundance, productivity and, particularly, survival, and so will always be an important caveat associated with evaluations of this kind.

A key comparison in this study has been between SIPMs, which effectively form the current “gold standard” for running PVAs, and other, simpler but potentially less defensible, methods. Our ability to make this comparison has been limited, however, by the fact that SIPMs have to date only been run for a very limited set of species-by-colony combinations within Scotland, and the fact that the timescale and resourcing of this project did not allow additional SIPM runs at other colonies to be produced. The results of this particular comparison should, therefore, be treated with caution, as it is based on a relatively small, and not necessarily representative, set of species and colonies, and because this comparison could only be performed over a relatively short test period (2013-2017).

We have focused here upon PVA methods that are widely used in practice. With the exception of SIPMs, we have focused upon relatively simple methods that can be

run in an automated way, with little or no manual intervention required to set up and run the population model. A wide range of other statistical methods could potentially be used for generating PVAs, and it would be interesting to consider comparisons against these in future work. It would be possible, in particular, to consider statistical methods – e.g., extensions of the Ricker or Gompertz models – that allow for regional pooling within the model. It would also be interesting to investigate the performance of PVA methods that account for metapopulations, where this is feasible.

The focus of this project has been upon using population models to generate predictions of actual counts. PVAs are often used, in practice, to compare the predicted values that are obtained when an intervention is introduced (e.g., an offshore renewable energy development) against the predicted values that are obtained in the absence of the intervention (a “baseline”), in order to quantify the impact of the intervention upon seabird demographics. The levels of error associated with using PVAs to predict actual abundance will, in general, not be the same as the levels of error associated with using PVAs to compare relative levels of abundance under two scenarios in this way. We would expect, in general, that the levels of error associated with making relative comparisons will be lower than the levels of error associated with predicting absolute abundance, so long as the relative comparisons are performed in a sensible way (e.g., with appropriate matching of stochastic simulations). However, we hypothesise that the levels of error associated with predicting absolute error are likely to be, nonetheless, strongly associated (e.g., correlated) with the levels of error associated with making relative comparisons, but it is beyond the scope of this project to examine whether this is indeed the case.

7.3. Implications of the results

Our evaluation has shown that deterministic and stochastic Leslie matrix models can often have poor performance in predicting observed abundance levels, and in a substantial number of cases perform poorly even relative to a time series model (the simple growth model) that has an extremely simple and biologically unrealistic structure. Our interpretation of this result is that the poor performance of the Leslie matrix models relative to time series models largely relates to the fact that there is relatively little direct data available on adult and immature survival rates, with the result that the rates being used in running the Leslie matrix models are likely to provide inaccurate estimates of the demographic rates for the population of interest. Our proposed solutions either involve collecting more data relating to these rates, or using models that leverage both demographic and abundance data, which is much more widely available, to try and quantify and adjust for these inaccuracies. We also

recommend the use of sensitivity analyses, to evaluate how important these issues are for PVA metrics.

The results of our comparisons also consistently suggest that the types of stochastic Leslie matrix models that are typically used in practice for running PVAs of seabird species tend to systematically underestimate uncertainty. Our evaluation demonstrated that the proportion of situations in which the 95% confidence intervals of projected abundance contained observed abundance was substantially lower than the target level of 95%. This occurred even though the inter-annual standard deviations in demographic rates used in the models were, in many cases, quite large, suggesting that the issue is not solely with the input values themselves, but that the structure of the models must also be leading to under-estimation of uncertainty. Statistical arguments suggest that a key underlying cause of this under-estimation is likely to be the assumption within current stochastic Leslie matrix models that stochasticity is (a) independent between demographic rates and (b) independent between years. Both assumptions are likely to be biologically implausible, and any failure of the independence assumption would be likely to lead to a systematic underestimation of uncertainty. More specifically, we note that:

a. the models implemented here assumed that inter-annual variations in stochastic demographic rates (survival and productivity) are independent of each other - if there is positive correlation between rates and/or years (as seems likely in practice) then the independence assumption is likely to lead to uncertainty being underestimated, potentially substantially;
b. these models accounted for variability in annual rates but not for uncertainty in the means and SDs of these rates.

As a consequence, the confidence intervals associated with these models should be interpreted extremely carefully.

IPMs and SIPMs provide a potential solution to this issue because these methods effectively adjust for this effect by directly quantifying the resulting variability in abundance. We anticipate that an extension of stochastic Leslie matrix models to incorporate empirical estimates of correlations between rates and years may also resolve the under-estimation. However, the key challenge in achieving this is not the technical ability to include correlations between rates and years into the Leslie matrix projections (which is relatively straightforward, and has already been done, for

example, in the Seabird_PVA_Tool[2], but rather the lack of empirical data on the magnitude and directions of correlations.

We have only focused here on the under-estimation of uncertainty in predicting abundance, but we strongly suspect that this will also carry over into systematic bias in the values of uncertainty-related PVA metrics, such as the probability of quasi-extinction, and to underestimation of the uncertainty associated with ratio-based PVA metrics.

The results of our evaluations also suggest that simple time series models tend to produce extremely high estimates of uncertainty, particularly in the case of the Gompertz model. This is because, in their standard usage, these models are purely empirical, imposing no biological constraints upon the rates of growth or decline that are possible. These models allow behaviour that is biologically implausible, for certain parameter values (i.e., allowing the populations to increase to extremely high levels), and the abundance data are often insufficient to rule out such parameter values. This tendency could be constrained by specification of prior information on key model parameters, if available for the species in question. However, it is likely that these methods will only perform well when there are relatively long time series of abundance data available for model fitting. In addition, these methods can only be used for PVAs in situations where impacts can reasonably be assumed to operate solely upon a single demographic rate (e.g. adult survival), because they do not allow a partitioning of impacts into different demographic processes.

7.4. Specific recommendations

7.4.1 Recommendations arising from empirical findings in the project

On the basis of the key empirical findings of the comparisons within this project we make a number of specific recommendations regarding the use of Leslie matrix models for running PVAs of seabirds, given that these are by far the most widely used approach for this purpose in practice:

Recommendation 1. Empirical validation of Leslie Matrix Models.

In situations where deterministic or stochastic Leslie matrix models are being used to produce PVAs, and relevant abundance data exist for multiple years, we recommend that the performance of the Leslie Matrix models should be validated empirically.

[2] https://github.com/naturalengland/Seabird_PVA_Tool

This can be achieved by using the Leslie matrix model to produce projections of baseline abundance for the period for which abundance data are available, and then comparing the projected and observed abundance values against each other. Where substantive differences arise, we recommend using methods that adjust the rates within the Leslie matrix model to better match the abundance data (e.g. using a SIPM, or “tuning”).

Recommendation 2. Caution in interpreting uncertainty ranges from Stochastic Leslie Matrix Models.

We recommend that the uncertainty ranges derived from stochastic Leslie matrix models, and any metrics that involve these uncertainty ranges (such as quasi-extinction probabilities) should be interpreted with great caution, given that the results of this evaluation suggest that these ranges may often provide substantial underestimates of actual uncertainty.

Recommendation 3. Sensitivity testing of PVA outputs.

We recommend that the sensitivity of the PVA outputs of interest to the values of input parameters within the Leslie matrix models should be assessed wherever possible, and that, where relevant, the sensitivity of the outputs to the choice of model structure (e.g. inclusion or exclusion of density dependence) should also be assessed.

7.4.2 Recommendations on future research areas

We also make a number of specific recommendations regarding future research in this area:

Recommendation 4. Need for quantification of relative performance of model abundance predictions and its effect on a range of PVA metrics.

We recommend that a simulation-based study be undertaken to quantify the likely relationship between the performance of models in predicting absolute abundance and their performance in predicting a range of PVA metrics. Note that such a study would need to go beyond a standard sensitivity analysis, because it would need to account for the potential for structural uncertainties in the models being used to generate the PVAs, and not only for uncertainties in the values of the inputs to the PVA models.

Recommendation 5. Additional data collection on demography.

We recommend additional data collection in order to improve the defensibility of demographic rates.

Recommendation 6. Extension of Leslie Matrix methods to include correlation in demographic rates.

We recommend that stochastic Leslie matrix methods should be extended to include correlations between demographic rates – both between different rates, and between the values of rates in different years – because we believe that this is a key reason that these methods consistently underestimated uncertainty within our comparisons. The extension of the model structure to achieve this is relatively straightforward – the key issue lies in the need to quantify levels of correlation empirically, as there is currently little or no empirical evidence relating to this. Data collection to address this issue would involve annual population level estimates of adult survival, immature survival and breeding success, from the same population, and ideally from the same sample of individuals.

8. PVA guidelines

We conclude by briefly outlining a set of guidelines for best practice when running PVAs. These guidelines are partially based upon the results of the current evaluation, but are also based upon existing knowledge and expert judgement. As such, not all of the guidelines presented here are based directly on the results of the evaluation undertaken within this project.

8.1. Recommendation 1.

We recommend the use of Leslie matrices as providing the best general framework for running PVAs - they provide a flexible and biologically meaningful approach for accounting for the impacts of anthropogenic pressures upon demographic rates, and allow impacts on productivity, as well as survival, to be considered.

Our remaining recommendations concern the methods that are used for calculating the inputs to the Leslie matrix approach, and for quantifying and propagating the uncertainties associated with these inputs. The demographic inputs can either be calculated solely using demographic data ("deterministic Leslie matrix approach", "stochastic Leslie matrix approach"), or using a combination of abundance data and demographic data ("IPMs", "SIPMs", "non-Bayesian tuning methods").

8.2. Recommendation 2.

In situations where a reasonable amount of abundance data are available, we recommend the use of methods that adjust the demographic rates within a Leslie matrix to fit the available abundance data, and also provide a proper quantification of uncertainty. Specifically:

a. In general, we recommend the use of IPMs, as these methods make the most effective use of all available data, on both demography and abundance, and provide a defensible quantification of uncertainty.

b. In situations where there is very poor evidence for one demographic rate (e.g. juvenile survival) and a reasonable amount of empirical data on abundance and on the remaining demographic rates then we recommend that SIPMs can be used as a defensible (and simpler) alternative to IPMs.

c. Non-Bayesian approaches to tuning also allow the rates in the Leslie matrix models to be adjusted to fit the abundance data. However, these approaches do not currently allow the uncertainty associated with tuning to be defensibly quantified and accounted for, so we do not currently recommend the use of these methods (because of the importance of properly quantifying uncertainty). If these methods can be refined so that they can defensibly quantify uncertainty (e.g. through the development of an appropriate bootstrap procedure), then it may become appropriate to use these approach as an alternative to SIPMs.

8.3. Recommendation 3.

In situations where there is minimal abundance data, there is little choice but to just use standard (untuned) Leslie matrix approaches. In this context we think it is also important to quantify uncertainty, and hence to use stochastic rather than deterministic Leslie matrix approaches, but the results of the evaluations within this project suggest that current stochastic Leslie matrix approaches tend to systematically underestimate uncertainty. Until this issue is resolved, these estimates of uncertainty should be interpreted with considerable caution. Research to empirically quantify correlations between demographic rates, and to incorporate these into the stochastic Leslie matrix calculations, is likely to help to make this approach more defensible.

9. Acknowledgements

We thank the Project Steering Group for guidance throughout the project. We thank Aonghais Cook for providing additional background information to the Cook & Robinson (2010) report. Data have been provided to the SMP by the generous contributions of its partners, other organisations and volunteers throughout Britain and Ireland. Partners to the SMP are: BirdWatch Ireland; The British Trust for Ornithology; Centre for Ecology & Hydrology; Countryside Council for Wales; Department of Agriculture, Fisheries and Forestry (Isle of Man); Department of Environment, Heritage and Local Government (Republic of Ireland); States of Guernsey Government; JNCC; Manx Birdlife; Manx National Heritage; The National Trust; National Trust for Scotland; Natural England; Northern Ireland Environment Agency; The Royal Society for the Protection of Birds; Scottish Natural Heritage; Seabird Group; Shetland Oil Terminal Environmental Advisory Group; Scottish Wildlife Trust. We also acknowledge funding provided by Scottish Government Rural and Environmental Science and Analytical Services Division (RESAS) through their Contract Research Fund.

10. References

Addison, P.F.E., Rumpff, L., Bau, S.S., Carey, J.M., Chee, Y.E., Jarrad, F.C., McBride, M.F. & Burgman, M.A. (2013) Practical solutions for making models indispensable in conservation decision-making. Diversity and Distributions, 19, 490-502

Barlow, E.J., Daunt, F., Wanless, S. & Reid, J.M. (2013) Estimating dispersal distributions at multiple scales: within-colony and among-colony dispersal rates, distances and directions in European shags Phalacrocorax aristotelis. Ibis 155: 762-778

Caswell, H. (2001) Matrix population models - construction, analysis, and interpretation. Sinauer, Sunderland, MA. 722p

Cook, A.S.C.P. & R.A. Robinson (2010) How Representative is the Current Monitoring of Breeding Seabirds in the UK? BTO Research Report No. 573.

Cook, A.S.C.P. & Robinson, R.A. (2016). Testing sensitivity of metrics of seabird population response to offshore wind farm effects. JNCC Report No. 553. JNCC, Peterborough.

Cook, A.S.C.P. & Robinson, R.A. (2017) Towards a framework for quantifying the population-level consequences of anthropogenic pressures on the environment: The case of seabirds and windfarms. Journal of Enviromental Management 190:113-121

Coulson, J.C. (2011) The Kittiwake. T & AD Poyser, London.

Daunt, F., Benvenuti, S., Harris, M.P., Dall'Antonia, L., Elston, D.A. & Wanless, S. (2002) Foraging strategies of the black-legged kittiwake Rissa trydactyla at a North sea colony: evidence for a maximum foraging range. Marine Ecology Progress Series 245: 239-247

Danchin, E. & Cam, E. (2002) Can non-breeding be a cost of breeding dispersal? Behavioral Ecology & Sociobiology 51:153-163

Dennis, B., Munholland, P.L. & Scott, J.M. (1991) Estimation of Growth and Extinction Parameters for Endangered Species. *Ecological Monographs* **61**(2), 115-143.

Drewitt, A.L., & Langston, R.H.W. (2006) Assessing the impacts of wind projects on birds. Ibis 148: S29–42

Dugger, A.M., DG. Ainley, P. O'B., Lyver, Barton, K. & Ballard, G. (2010) Survival differences and the effect of environmental instability on breeding dispersal in an

Adélie penguin meta-population. PNAS 107(27): 12375–12380.

Enstipp, M.R., Daunt, F., Wanless, S., Humphreys, E., Hamer, K.C., Benvenuti, S. & Gremillet, D. (2006) Foraging energetics of North Sea birds confronted with fluctuating prey availability. In: Top predators in marine ecosystems: their role in monitoring and management. (Eds I.L. Boyd, S. Wanless & K. Camphuysen). Cambridge University Press, Cambridge, pp191-210

Freeman, S., Searle, K., Bogdanova, M., Wanless, S. & Daunt, F. (2014) Population dynamics of Forth & Tay breeding seabirds: review of available models and modelling of key breeding populations (MSQ – 0006). Contract report to Marine Scotland Science.

Furness, R. & M. Trinder (2016) Qualifying impact assessments for selected seabird populations: A review of recent literature and understanding. MacArthur Green Report: Report commissioned by Vattenfall, Statkraft and ScottishPower

Renewables.
http://www.macarthurgreen.com/files/Seabird_Knowledge_Gap_Literature_Review.pdf

Green, R E., Langston, R H. W., McCluskie, A, Sutherland, R & Wilson, J D. (2016) Lack of sound science in assessing wind farm impacts on seabirds. Journal of Applied Ecology doi: 10.1111/1365-2664.12731.

Grecian, W.J., Inger, R., Attrill, M.J., Bearhop, S., Godley, B.J., Witt, M.J. & Votier, S.C. (2010) Potential impacts of wave-powered marine renewable energy installations on marine birds. Ibis 152: 683-97

Grist, H., Daunt, F., Wanless, S., Burthe, S., Newell, M.A., Harris, M.P. & Reid, J.M. (2017) Relationships between male and female migratory strategy and reproductive performance in partially migratory European shags (Phalacrocorax aristotelis). Journal of Animal Ecology 86: 1010-1021

Grist, H., Daunt, F., Wanless, S., Nelson, E.J., Harris, M.P., Newell, M. Burthe, S. & Reid, J.M. (2014) Site fidelity and individual variation in winter location in partially migratory European shags. PLOS One 9: e98562

Harris, M. P., Heubeck, M., Newell, M. A. & Wanless, S. (2015a) The need for year-specific correction factors (k values) when converting counts of individual Common Guillemots Uria aalge to breeding pairs. Bird Study 62: 276–279.

Harris, M.P., Newell, M.A. & Wanless, S. (2015b) The use of k values to convert counts of individual Razorbills Alca torda to breeding pairs. Seabird 28: 30–36

Herrando-Pérez, S., Delean, S., Brook, B. W., Cassey, P. & Bradshaw, C. J. A. (2014). Spatial climate patterns explain negligible variation in strength of compensatory density feedback mechanisms in birds and mammals. Plos One 9(3):e91536.

Hanski, I. (1999). Habitat Connectivity, Habitat Continuity, and Metapopulations in Dynamic Landscapes. Oikos 87(2): 209-219.

Horswill, C. & Robinson, R.A. (2015) Review of Seabird Demographic Rates and Density Dependence. JNCC Report No. 552. Joint Nature Conservation Committee, Peterborough.

Horswill, C., O'Brien, S.H. & Robinson, R.A. (2016) Density dependence and marine bird populations: are wind farm assessments precautionary. Journal of Applied Ecology doi: 10.1111/1365-2664.12841
http://jncc.defra.gov.uk/pdf/WKSEQUIN2008.pdf

Inch Cape Offshore Limited (2011) Inch Cape Offshore Wind Farm Environmental Statement: Appendix 15B Population Viability Analysis

Inchausti, P. & Weimerskirch, H. (2002) Dispersal and metapopulation dynamics of an oceanic seabird, the wandering albatross, and its consequences for its response to long-line fisheries. Journal of Animal Ecology 71(5): 765-770.

Jitlal, M., Burthe, S. Freeman, S. & Daunt, F. (2017) Testing and validating metrics of change produced by Population Viability Analysis (PVA) (Ref CR/2014/16). Draft report to Scottish Government

JNCC & NE (2012) Defining the level of additional mortality that the North Norfolk Coast SPA Sandwich tern population can sustain. JNCC & NE

JNCC (imputation methodology)
http://jncc.defra.gov.uk/pdf/Methods%20of%20analysis%20for%20production%20of%20indices%20of%20abundance%20and%20estimation%20of%20productivity1.pdf

Knight, A.T., Cowling, R.M., Rouget, M., Balmford, A., Lombard, A.T. & Campbell, B.M. (2008) Knowing but not doing: Selecting priority conservation areas and the research-implementation gap. Conservation Biology, 22, 610-617

Langston, R., Davies, I.M. & Scott, B.E. (2011) Seabird conservation and tidal stream and wave power generation: information needs for predicting and managing potential impacts. Marine Policy 35: 623-30

Larsen, J.K. & Guillemette, M. (2007) Effects of wind turbines on flight behaviour of wintering common eiders: implications for habitat use and collision risk. Journal of Applied Ecology 44: 516-522

Lele, S.R., Dennis, B. & Lutscher, F. (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* **10**, 551–563.

MacKenzie, A. & Perrow, M.R. (2009) Population viability analysis of the north Norfolk Sandwich tern Sterna sandvicensis population. Report for Centrica Renewable Energy ltd and AMEC Power & Process.

MacKenzie, A. & Perrow, M.R. (2011) Population viability analysis of the north Norfolk Sandwich tern Sterna sandvicensis population. Report for Centrica Renewable Energy ltd and AMEC Power & Process

Maclean, I.M.D., Frederikson, M & Rehfisch, M.M. (2007) Potential use of population viability analysis to assess the impact of offshore windfarms on bird populations. BTO Research Report No. 480 to COWRIE. BTO, Thetford.

Masden, E.A., Haydon, D.T., Fox, A.D. & Furness, R.W. (2010) Barriers to movement: Modelling energetic costs of avoiding marine wind farms amongst breeding seabirds. Marine Pollution Bulletin 60: 1085-1091

Masden, E.A., McCluskie, A., Owen, E. & Langston, R.H.W. (2015) Renewable energy developments in an uncertain world: The case of offshore wind and birds in the UK. Marine Policy, 51, 169-172

McCarthy, M. A, Andelman, S. J., & Possingham, H. P. (2001) Reliability of relative predictions in population viability analysis. Conservation Biology 17(4): 982-989.

Moray Offshore Renewables Ltd (2013) Environmental Statement: Ornithology population viability analysis outputs and review

Nadeem, K., Lele S. R. (2012) Likelihood based population viability analysis in the presence of observation error. *Oikos* **121**, 1656–1664.

Pe'er, G., Matsinos, Y.G., Johst, K., Franz, K.W., Turlure, C., Radchuk, V., Malinowska, A.H., Curtis, J.M.R., Naujokaitis-Lewis, I., Wintle, B.A. & Henle, K. (2013) A Protocol for Better Design, Application, and Communication of Population Viability Analyses. Conservation Biology, 27, 644-656

Plummer, M. (2003) JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/

Reed, J. M., Mills, L. S., Dunning, J. B., Menges, E. S., McKelvey, K. S, Frye, R., Beissinger, S. R., Anstett, M. C. & Miller, P. (2002). Emerging issues in population viability analysis. Conservation Biology 16(1): 7-19.

Ricker, W. E. (1954) Stock and Recruitment Journal of the Fisheries Research Board of Canada, 11(5): 559–623. doi:10.1139/f54-039.

Sanz-Aguilar, A., JM Igual, J.M., Tavecchia, G., Genovart, M. & Oro, D. (2016) When immigration mask threats: The rescue effect of a Scopoli's shearwater colony in the Western Mediterranean as a case study. Biological Conservation 198: 33-36

Scottish Government (2011) Habitats Regulations Appraisal of Draft Plan for Offshore Wind Energy in Scottish Territorial Waters: Appropriate Assessment Information Review (2011).
http://www.scotland.gov.uk/Publications/2011/03/04165857/15

Spendelow, J.A., Nichols, J.D., Nisbet, I.C.T., Hays, H. & Cormons, G.D. (1995) Estimating Annual Survival and Movement Rates of Adults within a Metapopulation of Roseate Terns. Ecology, Vol. 76, No. 8 (Dec., 1995), pp. 2415-2428

Stubben CJ & Milligan BG (2007). Estimating and Analyzing Demographic Models Using the popbio Package in R. *Journal of Statistical Software, 22*(11).

Thomas, G.E. (1993) Estimating annual total heron population counts. Appl. Statistics, 42, 473-486.

Trinder, M. (2014) Flamborough and Filey Coast pSPA Seabird PVA Final Report: Appendix N to the response submitted for deadline V. Report for SMart Wind.

Trinder, M. (2015) Flamborough and Filey Coast pSPA Seabird PVA Report: Appendix M to the response submitted for deadline IIA. Report for SMart Wind.

Trinder, M. & Furness, R. (2015) Flamborough and Filey Coast pSPA Seabird PVA Report: MacArthur Green Seabird PBA Report.

Wang, G.M., Hobbs, N.T., Slade, N.A., Merritt, J.F., Getz, L.L., Hunter Jr, M., Vessey, S.H., Witham, J. & Guillaumet, A. (2011) Comparative population dynamics of large and small mammals in the Northern Hemisphere: deterministic and stochastic forces. Ecography 36: 439–446.

Weimerskirch, H. (2001) Seabird demography and its relationship with the marine environment. In: Schreiber, E.A. & Burger, J. (eds) Biology of marine birds. CRC Press, Boca Raton, FL

Williams, G.C. (1966) Adaptation and Natural Selection. Princeton University Press, Princeton, USA

Winsor, C. P. (1932). The Gompertz Curve as a Growth Curve. Proc Natl Acad Sci U S A. 18(1): 1–8.

WWT (2012). Gannet Population Viability Analysis: Demographic data, population model and outputs. WWT Consulting.

Appendix A

Literature review of PVAs for seabirds

Grey literature used in PVAs for seabirds:

[1] MacArthur Green Seabird PBA Report – August 2015; Appendix M to the Response submitted for Deadline IIA; Application Reference: EN010053. Hornsea OWF Project 2. https://infrastructure.planninginspectorate.gov.uk/wp-content/ipc/uploads/projects/EN010053/EN010053-001275-Appendix%20M_MacArthur%20Green%20Seabird%20PVA%20Report%20-%20August%202015.pdf

[2] MacArthur Green 2014: Updated PVA Note. Appendix N to the Response submitted for Deadline V. Application Reference: EN010033. Hornsea OWF Project 1. https://infrastructure.planninginspectorate.gov.uk/wp-content/ipc/uploads/projects/EN010033/EN010033-001628-Appendix%20N.%20Updated%20PVA%20Note.pdf

[3] PVA Note. Appendix X to the Response submitted for Deadline IV. Application Reference: EN010033. Hornsea OWF Project 1. SmartWind. 2014. https://infrastructure.planninginspectorate.gov.uk/wp-content/ipc/uploads/projects/EN010033/EN010033-001579-Appendix%20X%20-%20PVA%20Note.pdf (same report as above)

[4] Reports – Habitats Regulations Assessment Report. Information to Support the Appropriate Assessment for Project One. PINS Document Reference: 12.6. APFP Regulation 5(2)(g). Kittiwake: Trinder and Furness. July 2013. Annex K of this report: https://infrastructure.planninginspectorate.gov.uk/wp-content/ipc/uploads/projects/EN010033/EN010033-000665-12.6%20Habitats%20Regulation%20Assessment.pdf

[5] Beatrice Offshore WF 2013. Arcus Consultancy Services. Section 7.5.3.

[6] McKenzie IBM 2011 Technical Report PVA model (June 2011) Race Bank OWF and Docking Shoal OWF.

[7] McKenzie IBM 2009

[8] MORL 2013 (Moray Firth)

[9] Inch Cape Volume 2F Appendix 15B: PVA

[10] Gannet PVA 2012 MacArthur Green

Peer-reviewed literature used in PVAs for seabirds and other species:

[1] Hernández-Camacho CJ, Bakker VJ, Aurioles-Gamboa D, Laake J, Gerber LR (2015) The Use of Surrogate Data in Demographic Population Viability Analysis: A Case Study of California Sea Lions. PLoS ONE 10(9): e0139158. https://doi.org/10.1371/journal.pone.0139158

[2] Maclean, I.M.D., Frederikson, M and Rehfisch, M.M. (2007) Potential use of population viability analysis to assess the impact of offshore windfarms on bird populations. BTO Research Report No. 480 to COWRIE. BTO, Thetford.

[3] McCarthy, M. A., et al. 2001. “Testing the Accuracy of Population Viability Analysis.” Conservation Biology, vol. 15, no. 4, 2001, pp. 1030–1038. JSTOR, www.jstor.org/stable/3061322

[4] Brook et al 2000. Predictive accuracy of PVA in conservation biology. Nature 404:385-387 .

[5] Reed et al 2002. Emerging issues in PVA. Conservation Biology 16: 7-19.

[6] Finkelstein ME, Wolf S, Goldman M, Doak DF, Sievert PR, Balogh G, et al. The anatomy of a (potential) disaster: Volcanoes, behavior, and population viability of the short-tailed albatross (Phoebastria albatrus). Biological Conservation. 2009; 143:321–331.

[7] Dennis, B., P. L. Munholland, and J. M. Scott. 1991. Estimation of growth and extinction parameters for endangered species. Ecological Monographs 61:115–143.

[8] Taylor 1995. Reliability of using PVA for risk classification of species. Conservation Biology 9:551-558.

[9] Doxa, A., Besnard, A., Bechet, A., Pin, C., Lebreton, J.-D. and Sadoul, N. (2013), Inferring dispersal from local demography. Anim Conserv, 16: 684-693. doi:10.1111/acv.12048.

[10] Peery and Henry 2010. Recovering marbled murrelets via corvid management: a population viability analysis approach. Biological Conservation, Volume 143, Issue 11, November 2010, Pages 2414-2424.

[11] Cook, A. S. C. P., & Robinson, R. A. (2017). Towards a framework for quantifying the population-level consequences of anthropogenic pressures on the environment: The case of seabirds and windfarms. Journal of Environmental Management, 190. https://doi.org/10.1016/j.jenvman.2016.12.025

[12] Sandvik Hanno, Tone K. Reiertsen,Kjell Einar Erikstad, Tycho Anker-Nilssen, Robert T. Barrett, Svein-Håkon Lorentsen, Geir Helge Systad, Mari S. Myksvol. 2014. The decline of Norwegian kittiwake populations: modelling the role of ocean warming. Climate Research Vol. 60: 91–102.

[13] J. Douglas Steventon, Glenn D. Sutherland, and Peter Arcese. 2006. A population-viability-based risk assessment of Marbled Murrelet nesting habitat policy in British Columbia. Can. J. For. Res. 36: 3075–3086.
[14] Nadeem and Lele 2012. Likelihood based population viability analysis in the presence of observation error. Oikos Volume121, Issue10: Pages 1656-1664.
[15] Arnold et al 2006. ALBATROSS POPULATIONS IN PERIL: A POPULATION TRAJECTORY FOR BLACK-BROWED ALBATROSSES AT SOUTH GEORGIA. Ecological Applications 16(1), 2006, pp. 419–432
[16] Oro et al 2004. Modelling demography and extinction risk in the endangered Balearic Shearwater. Biological Conservation 116(1):93-102. DOI: 10.1016/S0006-3207(03)00180-0
[17 Goyert et al. 2017. Density dependence and changes in the carrying capacity of Alaskan seabird populations. Biological Conservation 209:178-187. DOI: 10.1016/j.biocon.2017.02.011
[18] Jaffré, M., Le Galliard, J. Population viability analysis of plant and animal populations with stochastic integral projection models. Oecologia 182, 1031–1043 (2016). https://doi.org/10.1007/s00442-016-3704-4
[19] Edwards, K.L., Walker, S.L., Dunham, A.E. et al. Low birth rates and reproductive skew limit the viability of Europe's captive eastern black rhinoceros, Diceros bicornis michaeli . Biodivers Conserv 24, 2831–2852 (2015). https://doi.org/10.1007/s10531-015-0976-7.
[20] Hostetler JA, Kneip E, Van Vuren DH, Oli MK. Stochastic population dynamics of a montane ground-dwelling squirrel. PLoS One. 2012;7(3):e34379. doi:10.1371/journal.pone.0034379.
[21] Shiang-Lin Huang, Leszek Karczmarski, Jialin Chen, Ruilian Zhou, Wenzhi Lin, Haifei Zhang, Haiyan Li, Yuping Wu. 2012. Demography and population trends of the largest population of Indo-Pacific humpback dolphins. Biological Conservation, Volume 147, Issue 1, Pages 234-242.
[22] Ruete, A., Wiklund, K. and Snäll, T. (2012), Hierarchical Bayesian estimation of the population viability of an epixylic moss. Journal of Ecology, 100: 499-507. doi:10.1111/j.1365-2745.2011.01887.x

Appendix B

Detailed results of Forth-Tay evaluation

In this appendix we present detailed results for the Forth-Tay. The comparison between population modelling approaches within this dataset should be interpreted with considerable caution, hence we present them in this appendix, rather than the main text. This is because the number of test years available for the Forth-Tay comparison is typically very low, and may not provide a representative sample of years across populations within the Forth-Tay area. In addition, the results of the comparisons for different years and SPAs are unlikely to be independent, with the result that the effective number of independent comparisons is likely to be lower than the number of apparent comparisons (e.g. the comparisons are likely to contain less information than would be applied by the sample size).

Atlantic puffins

Methods could be compared at only one SPA in the Forth-Tay for this species, so all modelling methods were compared to observed counts from the Forth Islands SPA in 2013, with a test period starting in 2013. In terms of mean and median estimates for the predicted abundance, all models performed well, with the exception of the SIPM model, where the unstable trend in puffin counts caused the model to over-predict the observed abundance, although the 95% credible interval did encompass the true value (Figure 30; see Freeman *et al.* 2014 for full details of challenges modelling Forth/Tay puffins using SIPMs). All other modelling methods slightly underestimated the observed count, but where they could be generated, confidence intervals did encompass the true value (Figure 30) Regional pooling methods for demographic data resulted in little change to the predicted value, or confidence limits (Figure 30).

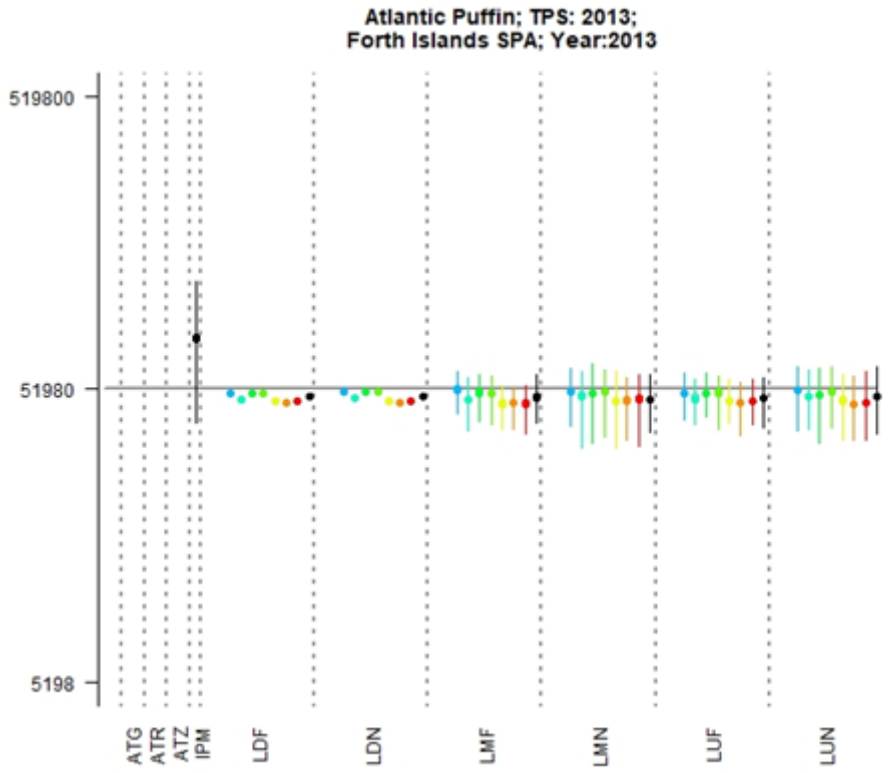


Figure 30: Atlantic Puffin results: performance of population modelling methods for the Forth-Tay SPAs for specific combinations of SPA, year and test period start (TPS). Each graph shows the median (open circle), mean (closed circle) and 95% confidence interval (vertical line) associated with each modelling method. Statistical methods are shown in blocks (separated by dotted grey lines), and the regional pooling methods within these, colour coded as: R0 (purple), R1 (dark blue), R2 (blue), R3 (light blue), R4 (green), R5 (light green), R6 (yellow), R7 (orange), R8 (red) and R9 (black). If some methods are omitted from a plot it is because they could not be applied for this combination. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix deterministic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

Black-legged kittiwakes

Across the four SPAs for this species there were annual 12 counts in the test period that could be compared against modelled values. In general, the SIPM performed very well, with the predicted mean or median being very close to the observed abundance in almost all instances, and the 95% credible interval capturing the

observed abundance in all cases except one (Figure 31). Two of the time series methods – the simple growth and Ricker models – also performed well, when they could be applied, with predicted means and medians falling close to the observed value, and the 95% confidence intervals capturing the observed value in most instances (Figure 31). However, the time series Gompertz model performed more poorly, often overestimating the observed abundance by a considerable amount, and producing 95% confidence intervals that were very wide, and therefore of little use (Figure 31). The various Leslie matrix methods all performed similarly, with predicted abundances that were reasonably close to the observed abundance in many cases, but with a notable underestimation of uncertainty, whereby on many occasions the 95% confidence intervals were very narrow, and often did not include the observed value (Figure 31). In several instances, the Leslie Matrix methods tended to result in an increasing underestimation of the observed value as the level of regional pooling increased from regions R1-R9 (Figure 31).

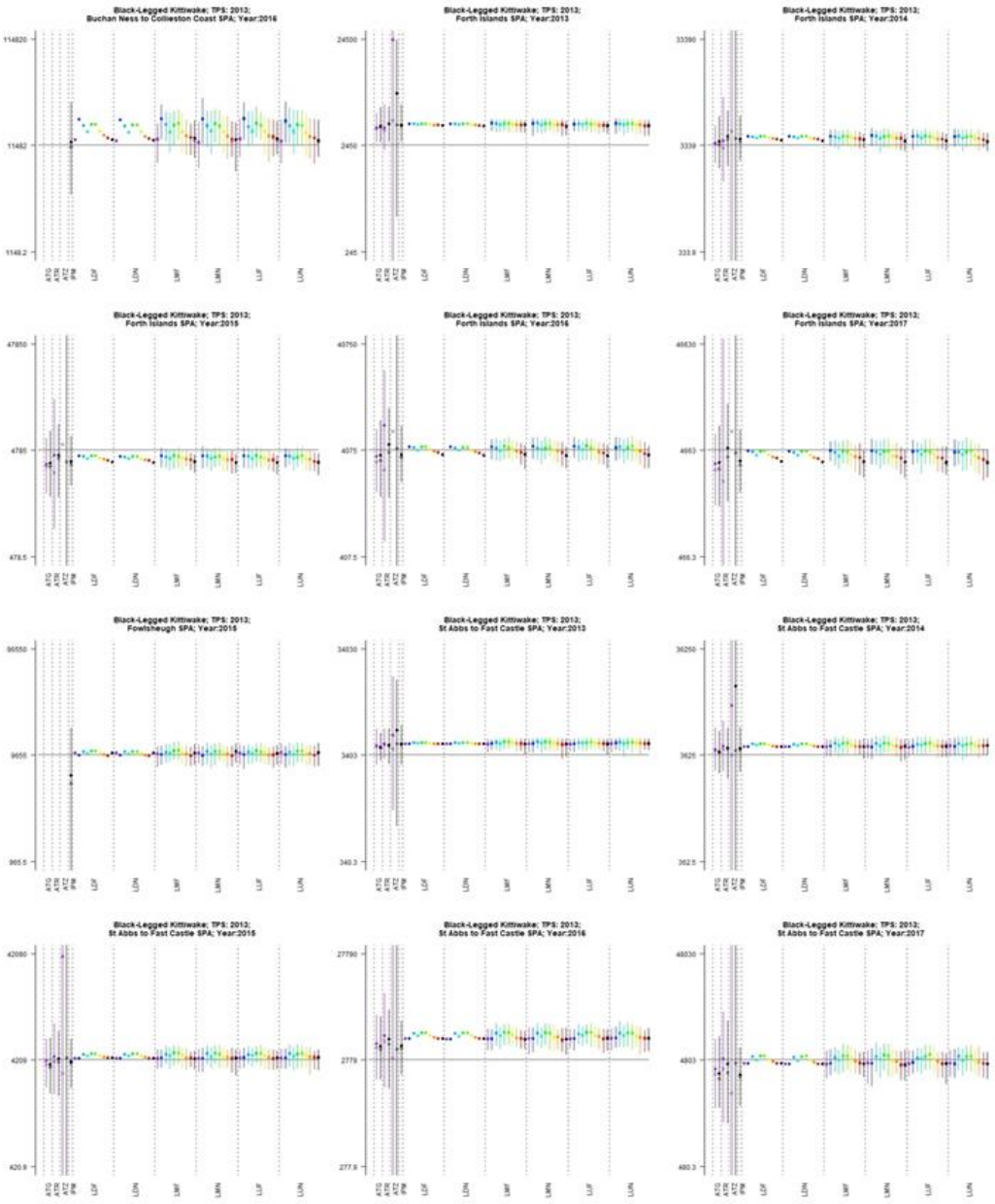


Figure 31: Black-legged kittiwake results: performance of population modelling methods for the Forth-Tay SPAs for specific combinations of SPA, year and test period start (TPS). Each graph shows the median (open circle), mean (closed circle) and 95% confidence interval (vertical line) associated with each modelling method. Statistical methods are shown in blocks (separated by dotted grey lines), and the regional pooling methods within these, colour coded as: R0 (purple), R1 (dark blue), R2 (blue), R3 (light blue), R4 (green), R5 (light green), R6 (yellow), R7 (orange), R8 (red) and R9 (black). If some methods are omitted from a plot it is because they could not be applied for this combination. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix determinisitic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

Common guillemot

Overall, modelling methods could be compared at nine SPA by year by training period combinations for this species, and results were very similar to those described above for black-legged kittiwakes. In general, the SIPM performed very well, with the predicted mean or median being very close to the observed abundance in almost all instances, and the 95% credible interval capturing the observed abundance in all cases except one (Figure 32). Two of the time series methods – the simple growth and Ricker models – also performed well, when they could be applied, with predicted means and medians falling close to the observed value, and the 95% confidence intervals capturing the observed value in most instances (Figure 32). However, the time series Gompertz model performed more poorly, often overestimating the observed abundance by a considerable amount, and producing 95% confidence intervals that were very wide, and therefore of little use (Figure 32). The various Leslie matrix methods all performed similarly, with predicted abundances that were reasonably close to the observed abundance in many cases, but with a notable underestimation of uncertainty, whereby on many occasions the 95% confidence intervals were very narrow, and in approximately half the test cases, did not include the observed value (Figure 32).In contrast to the results for black-legged kittiwakes, there was no consistent trend in bias for predicted values as the level of regional pooling increased from regions R1-R9 (Figure 32).

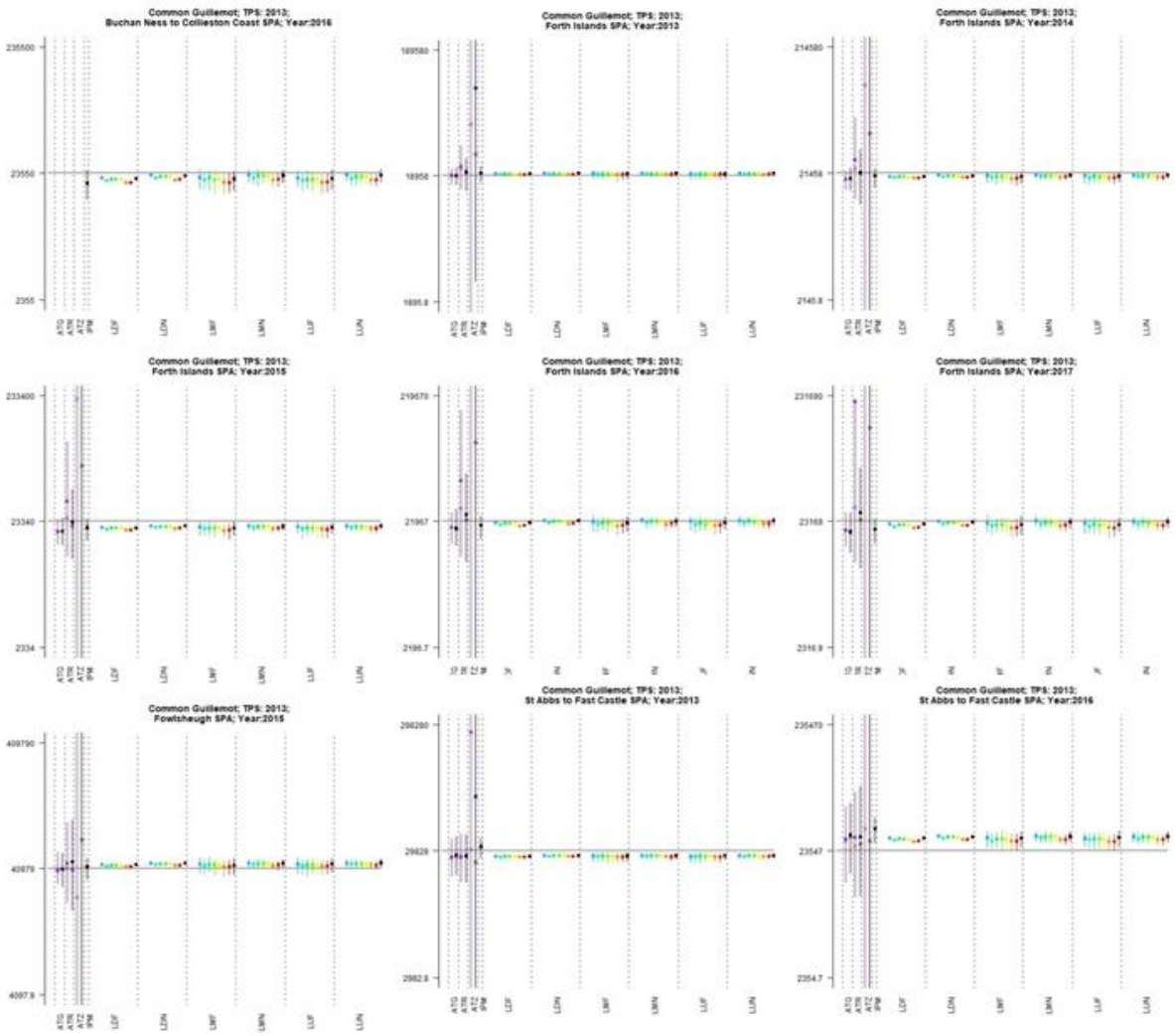


Figure 32: Common guillemot results: performance of population modelling methods for the Forth-Tay SPAs for specific combinations of SPA, year and test period start (TPS). Each graph shows the median (open circle), mean (closed circle) and 95% confidence interval (vertical line) associated with each modelling method. Statistical methods are shown in blocks (separated by dotted grey lines), and the regional pooling methods within these, colour coded as: R0 (purple), R1 (dark blue), R2 (blue), R3 (light blue), R4 (green), R5 (light green), R6 (yellow), R7 (orange), R8 (red) and R9 (black). If some methods are omitted from a plot it is because they could not be applied for this combination. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix deterministic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

Herring gull

Comparisons for herring gulls could only be made for a single SPA, for five years. *The SIPM* performed particularly well for this species, in each of the five years, with the predicted mean or median being very close to the observed abundance in all instances, and the 95% credible interval capturing the observed abundance in all cases (Figure 33). Two of the time series methods – the simple growth and Ricker models – also performed reasonably well, when they could be applied, with predicted means and medians falling reasonably close to the observed abundance, although with a tendency for overestimation, and 95% confidence intervals capturing the observed abundance, but tending to generate large confidence intervals (Figure 33). As with the previous species, the Gompertz time series methods performed more poorly, often overestimating the observed abundance by a considerable amount, and producing 95% confidence intervals that were very wide, and therefore of little use (Figure 33). For this species, the Leslie matrix methods also performed poorly, overestimating the observed abundance in all five years, with very narrow 95% confidence intervals that did not include the observed abundance in almost all instances (Figure 33). The LMN and the LUM performed better than the other Leslie matrix methods in terms of generating marginally wider 95% confidence intervals that occasionally included the observed abundance (Figure 33). There was some tendency for the overestimation of the Leslie matrix methods to decrease as the pooling region increased from R2-R9 (Figure 33).

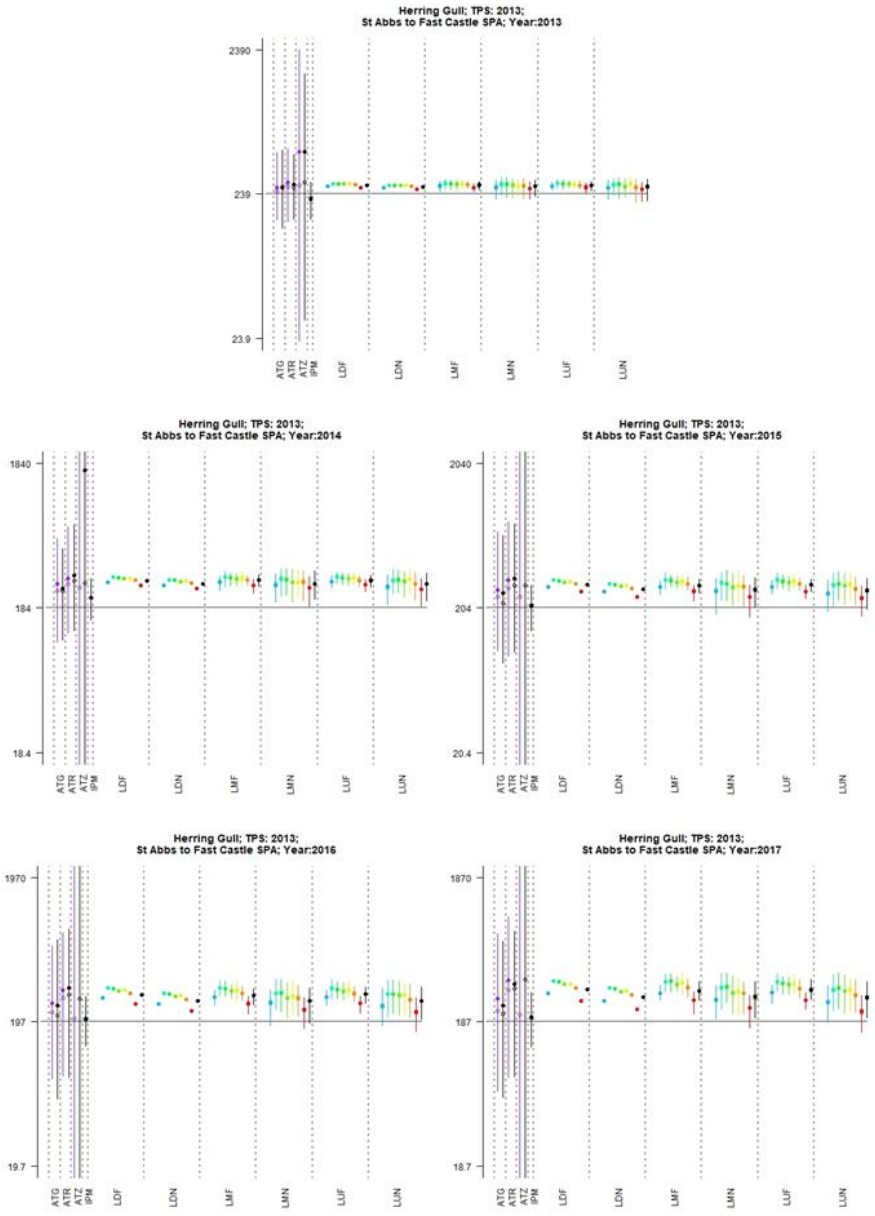
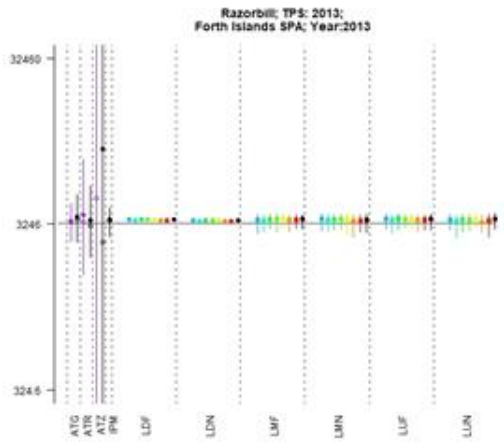
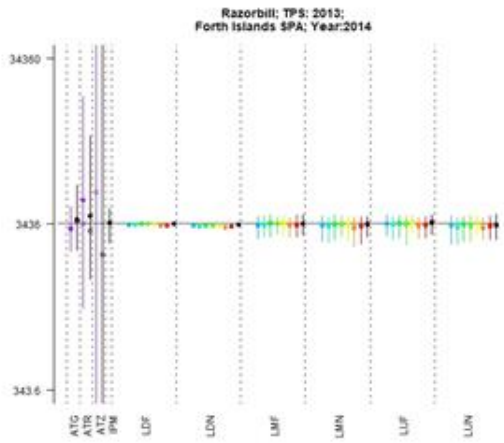


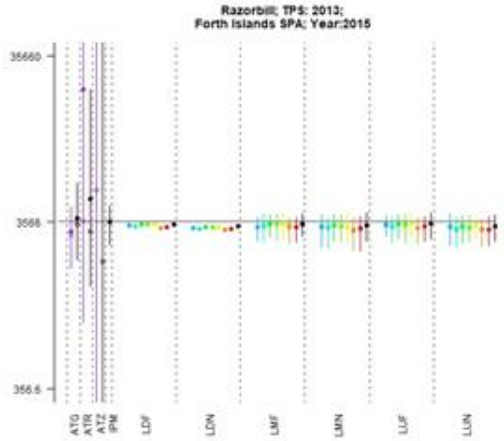
Figure 33: Herring gull results: performance of population modelling methods for the Forth-Tay SPAs for specific combinations of SPA, year and test period start (TPS). Each graph shows the median (open circle), mean (closed circle) and 95% confidence interval (vertical line) associated with each modelling method. Statistical methods are shown in blocks (separated by dotted grey lines), and the regional pooling methods within these, colour coded as: R0 (purple), R1 (dark blue), R2 (blue), R3 (light blue), R4 (green), R5 (light green), R6 (yellow), R7 (orange), R8 (red) and R9 (black). If some methods are omitted from a plot it is because they could not be applied for this combination. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix deterministic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

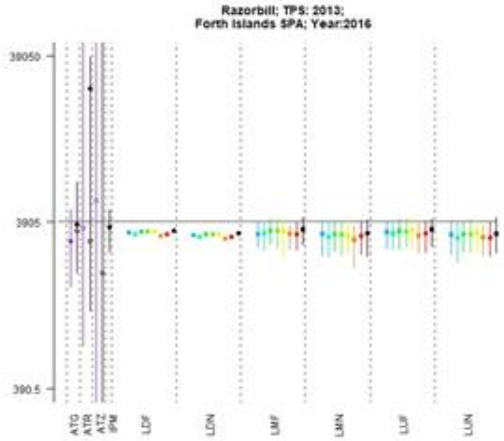
Razorbill

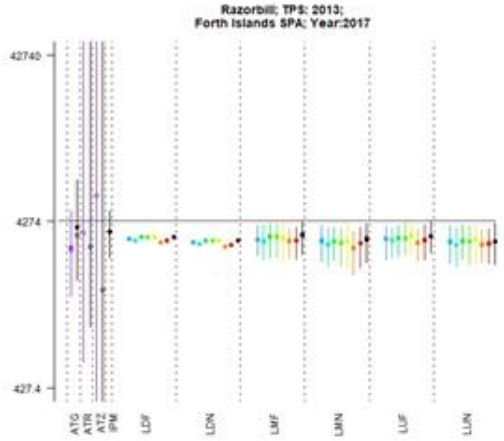
Comparisons for razorbills could be made in eight SPA by year by training period combinations. The SIPM again performed particularly well for this species, in all eight test cases, with the predicted mean or median being very close to the observed abundance in all instances, and the 95% credible interval capturing the observed abundance in all cases (Figure 34). Two of the time series methods – the simple growth and Ricker models – also performed reasonably well, when they could be applied, with predicted means and medians falling reasonably close to the observed abundance, and 95% confidence intervals capturing the observed abundance, but tending to generate large confidence intervals (Figure 34). As with the other species, the time series Gompertz method performed more poorly, occasionally overestimating the observed abundance by a considerable amount, and producing 95% confidence intervals that were very wide, and therefore of little use (Figure 34). The various Leslie matrix methods all performed similarly, with predicted abundances that were reasonably close to the observed abundance in many cases, but with a notable underestimation of uncertainty, such that on many occasions the 95% confidence intervals were very narrow, and in three of the eight test cases, did not include the observed value (Figure 34).

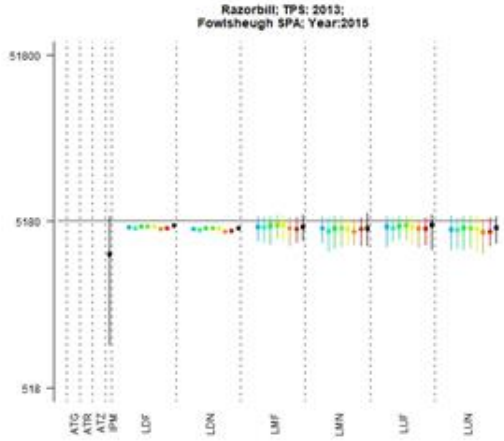

Razorbill; TPS: 2013;
Forth Islands SPA; Year:2013
32460
3246
324.6
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN

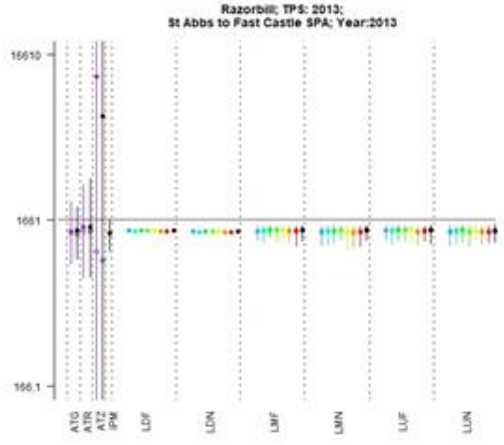

Razorbill; TPS: 2013;
Forth Islands SPA; Year:2014
34360
3436
343.6
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN


Razorbill; TPS: 2013;
Forth Islands SPA; Year:2015
35660
3568
356.6
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN


Razorbill; TPS: 2013;
Forth Islands SPA; Year:2016
39050
3905
390.5
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN


Razorbill; TPS: 2013;
Forth Islands SPA; Year:2017
42740
4274
427.4
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN


Razorbill; TPS: 2013;
Fowlsheugh SPA; Year:2015
51800
5180
518
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN


Razorbill; TPS: 2013;
St Abbs to Fast Castle SPA; Year:2013
16610
1661
166.1
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN

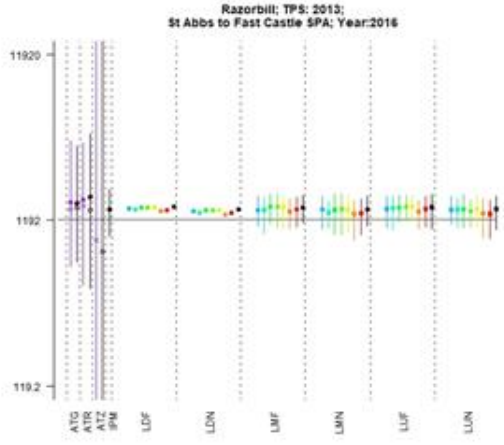

Razorbill; TPS: 2013;
St Abbs to Fast Castle SPA; Year:2016
11920
1192
119.2
ATG
ATR
ATZ
IPM
LDF
LDN
LMF
LMN
LUF
LUN

Figure 34: Razorbill results: performance of population modelling methods for the Forth-Tay SPAs for specific combinations of SPA, year and test period start (TPS). Each graph shows the median (open circle), mean (closed circle) and 95% confidence interval (vertical line) associated with each modelling method. Statistical methods are shown in blocks (separated by dotted grey lines), and the regional pooling methods within these, colour coded as: R0 (purple), R1 (dark blue), R2 (blue), R3 (light blue), R4 (green), R5 (light green), R6 (yellow), R7 (orange), R8 (red) and R9 (black). If some methods are omitted from a plot it is because they could not be applied for this combination. Pooling regions were: R0: site level; R1: SMP regions; R2: ICES regions; R3: JNCC regional seas; R4: Cook & Robinson Abundance; R5: Cook & Robinson Breeding Success; R6: MSFD; R7: OSPAR; R8: Global (all colonies in England, Northern Ireland, Scotland, Wales, Channel Islands and Isle of Man). Modelling methods were: ATG: simple time series growth model; ATR: Ricker model; ATZ: Gompertz model; IPM: Semi-Integrated Population Model; LDF: Leslie Matrix deterministic model parameterised using national rates; LDN: Leslie Matrix deterministic model parameterised using Forth-Tay rates; LMF: Leslie Matrix stochastic model with constrained productivity parameterised with Forth-Tay rates; LMN: Leslie Matrix Stochastic model with constrained productivity parameterised with National rates; LUF: Leslie Matrix stochastic model with unconstrained productivity parameterised with Forth-Tay rates; LUN: Leslie Matrix Stochastic model with unconstrained productivity parameterised with National rates.

Tables

Table 0-1: Adult and immature survival data for each of the 15 species of seabirds used in assessments of PVA methods. Age (min) and Age (max) define the age range over which survival has been estimated. Standard deviation in survival rates denoted by 'SD'. For three species, no data were available on the standard deviation of survival rates: little tern, great black-backed gull and arctic skua (missing immature survival rate standard deviation only). For these three species we used standard deviations from the nearest closely related species: common tern, herring gull, and for immature great skuas we used the standard deviation for adult great skuas.

Species	Life stage	Age (min)	Age (max)	Mean survival	SD
Arctic Skua	Immature	0	4	0.346	0.038
Arctic Skua	Adult			0.91	0.038
Atlantic Puffin	Immature	0	3	0.709	0.022
Atlantic Puffin	Immature	3	4	0.76	0.019
Atlantic Puffin	Immature	4	5	0.805	0.017
Atlantic Puffin	Adult			0.906	0.083
Common Guillemot	Immature	0	1	0.56	0.013
Common Guillemot	Immature	1	2	0.792	0.034
Common Guillemot	Immature	2	3	0.917	0.022
Common Guillemot	Adult			0.939	0.015
Common Tern	Immature	0	3	0.441	0.004
Common Tern	Immature	3	5	0.85	0.014
Common Tern	Adult			0.883	0.014
Fulmar	Immature	0	8	0.26	0.15
Fulmar	Adult			0.936	0.055
Great Black-Backed Gull	All birds combined			0.93	0.034
Great Cormorant	Immature	0	1	0.54	0.09
Great Cormorant	Adult			0.868	0.055
Herring Gull	Immature	0	1	0.798	0.092
Herring Gull	Adult			0.834	0.034
Kittiwake	Immature	0	1	0.79	0.051
Kittiwake	Adult			0.854	0.051
Lesser Black-Backed Gull	Immature	0	1	0.82	0.022
Lesser Black-Backed Gull	Adult			0.885	0.022
Little Tern	All birds combined			0.8	0.014
Northern Gannet	Immature	0	1	0.424	0.007
Northern Gannet	Immature	1	2	0.829	0.004
Northern Gannet	Immature	2	3	0.891	0.003
Northern Gannet	Immature	3	4	0.895	0.003
Northern Gannet	Adult			0.919	0.042
Razorbill	Immature	0	2	0.63	0.209
Razorbill	Adult			0.895	0.067
Sandwich Tern	Immature	0	2	0.358	0.219
Sandwich Tern	Immature	2	5	0.741	0.206
Sandwich Tern	Adult			0.898	0.029
Shag	Immature	0	1	0.513	0.256
Shag	Immature	1	2	0.737	0.181
Shag	Adult			0.858	0.194

Table 0-2: Summary of the amount of data – the number of colony-by-year combinations - available for abundance and breeding success for each species.

Species	Abundance	Breeding success
Arctic skua	1561	381
Atlantic puffin	1262	114
Common guillemot	2240	309
Common term	5546	1824
Fulmar	6382	975
Great black-backed gull	5895	975
Great cormorant	3739	217
Herring gull	8046	1233
Kittiwake	3683	1423
Lesser black-backed gull	3770	361
Little tern	3558	14
Northern gannet	495	186
Razorbill	2450	163
Sandwich tern	1768	376
Shag	3953	536

Table 0-3: Regional classifications used in pooling data.

Regional classification		Number of regions
R0	Site (i.e. no regional pooling)	6383
R1	SMP	113
R2	ICES	11
R3	Regional Seas	8
R4	CRA	7
R5	CRB	4
R6	MSFD	3
R7	OSPAR	3
R8	Global	1
R9	Forth-Tay SPAs	5

Table 0-4: Description of the empirical inputs needed for generating PVAs using deterministic or stochastic Leslie matrix approaches.

	Input	**Description**	**Value determined by**
I1	Breeding success	Mean, and, for stochastic version only, SD	Species & pooling region
I2	Survival	Mean, and, for stochastic version only, SD	Species & age class
I3	Age at first breeding	Value	Species
I4	Initial count	Value, year associated with the value	Species & target colony

Table 0-5: Summary of PVA methods, and minimum data requirements for each method. "TP" denotes the training period.

Method	Model type	Specific model	Type of data required	Minimum data requirements	Survival rates
ATG	Abundance time series models	Simple growth model	abundance	10 years+ in TP for which abundance data are available in both current and previous year	Not relevant
ATR		Ricker	Abundance		
ATZ		Gompertz	Abundance		
LDN	Leslie matrix models	Deterministic	Demographic rates	1+ years breeding success data in TP, and 1+ years abundance data in TP	National
LDF					Forth-Tay
LMN		Stochastic – constrained productivity	Demographic rates	2+ years breeding success data in TP, and 1+ years abundance data in TP	National
LMF			Demographic rates		Forth-Tay
LUN		Stochastic – unconstrained productivity	Demographic rates		National
LUF			Demographic rates		Forth-Tay
IPM	Semi-integrated population model	Freeman et al. (2014)	Abundance and demographic rates	See Freeman et al. (2014)	Forth-Tay

Table 0-6: Criteria used in the evaluating the performance of PVA methods.

Criterion	Quantified by	Good performance indicated by
C1. Ability to use	Percentage of situations in which it is possible to apply the method	High values – values close to 100% are ideal.
C2. Occurrence of highly implausible results.	Percentage of situations in which $\lvert r_{ij} \rvert < 2$	High values – values close to 100% are ideal.
C3. Lack of systematic bias	Mean value of r_{ij}	Values close to zero.
C4. Lack of error	Mean value of $\lvert r_{ij} \rvert$	Low values – values close to zero are ideal.
C5. Accurate quantification of uncertainty	Percentage of situations in which observed count lies within 95% prediction intervals	Values close to the nominal level (95%) are ideal.
C6. Level of uncertainty	Width of 95% confidence interval	Low values – values close to zero are ideal
C&. Ease of Computation	Computer time (seconds) to run the method	Low values – values close to zero are ideal

Table 0-7: Number of species-colony combinations for which an observed count exists within the test period, for each definition of the test period, and the mean number of years with counts in the test period for each species-colony combination.

Test period	Number of species-colony combinations for which evaluation is possible	Mean number of years of counts within test period
1998-2017	2186	5.72
2003-2017	2350	4.49
2008-2017	2197	3.22
2013-2017	1869	1.93

Marine Scotland Science
Marine Laboratory
375 Victoria Road
Aberdeen
AB11 9DB

Copies of this report are available from the Marine Scotland website at
www.gov.scot/marinescotland