**Title:** Inferring structured vital rates from a time series of population sizes and structures: an inverse problem

**Running title:** Inferring structured vital rates

**Word count:**

**Authors:** Edgar J. González\*,1, Carlos Martorell2, Benjamin M. Bolker1,3

**Addresses:**

1: Department of Mathematics and Statistics, McMaster University, 1280 Main St W, L8S4L8, Hamilton, Ontario, Canada

2: Departamento de Ecología y Recursos Naturales, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, 04510, México, D.F., Mexico

3: Department of Biology, McMaster University, 1280 Main St W, L8S4L8, Hamilton, Ontario, Canada

**Corresponding author:**

***Name*:** Edgar J. González

***Address*:**Department of Mathematics and Statistics, McMaster University, 1280 Main St W, L8S4L8, Hamilton, Ontario, Canada

***E-mail*:**gonzaleze@math.mcmaster.ca

**Abstract**

1. Traditional demographic methods have relied on tracking marked individuals in a population over time to estimate its vital rates. In a structured population these vital rates change as a function of a variable, such as size, that explains most of the among-individual demographic variation. Projection models integrate these structured vital rates, generating a time series of population sizes and structures.

2. Here, we present an inverse problem: inferring the vital rates and their change according to a structuring variable from the time series that they produce. We used simulated data to show that this approach works for the case where the vital rates are structured by a continuous variable, remain constant through time, and the population is not in its stable state. We also explored cases where partial information on the vital rates is available, and well as scenarios of shorter time series and reduced per-year sample size.

3. Using a hybrid optimization approach to estimate model parameters, we show that this approach can provide accurate reconstructions of the vital rates in a scenario where no information on them is available. Better estimates are obtained when some additional information on the vital rates is provided. Parameter estimation is more difficult when short time series are provided, but per-year sample size can be largely reduced without having a large impact on the estimates.

4. Given the type of input that the method uses, it can be a useful tool for those systems where tracking has not been performed, or is costly or impractical.

**Key-words**

inverse model, population dynamics, integral projection model, integrated population model, demography, parameter estimation

**Introduction**

Population ecology is a time and money demanding discipline. The information required for the understanding of the dynamics of a population over time, particularly in long-lived organisms, requires the follow up of individuals for a sensible timespan to capture data on the birth, survival, trait change, and reproduction rates of its members. Depending on the species and system under study, some of these rates will be harder to measure than others, and cases can exist where no information on any rate is available.

Additionally, individuals differ because they display different traits (e.g., allele, size, weight, sex) or experience different environments (e.g, less/more predation, light intensity, rainfall, disturbance, etc.). By increasing variation among its individuals, we say that these variables *structure* the population, and thus that vital rates are structured (Tuljapurkar & Caswell, 1997; Merow et al. 2014). Although important to understanding the dynamics of a population, accounting for the effect of these variables generally increases costs.

Integrated population models (traditionally called IPMs, but here IPopMs) have been proposed as a single statistical framework to study the dynamics of populations. In IPopMs, data from different sources can be used together and in this way better inform unknown estimates on demographic parameters (Raftery, Givens & Zeh, 1995; Besbeas et al. 2002; Brooks et al. 2004; Schaub & Abadi, 2011; Kéry & Schaub 2012; Chandler & Clark, 2014). Therefore, this framework allows estimating partial information on vital rates to infer those rates that are harder to estimate.

Previous work with IPopMs has focused on scenarios where only partial information on the vital rates is available, usually working with vertebrate species (e.g. Schaub et al. 2007; Kéry & Schaub 2012; Maunder & Punt 2013; but cf. McMahon & Parker 2014). In addition to research explicitly stated as IPopMs, many authors have tried to infer unobserved vital rates from observed ones (REFS; McMahon & Parker 2013). This method can be traced back to inverse/backward projection models in human populations where the aim is to infer mortality and fertility rates, along with population age structure, from annual birth and death records (Lee 1974, 1985; Wrigley & Schofield 1981; Oeppen 1993).

When at least one of the variables that structure a population is continuous, an integral projection model (traditionally called IPM, but here IProjM) is the standard tool to analyse how structured vital rates interact. In an IProjM the user provides information on the vital rates, fits a function to each of these rates and obtains a time series of population structures describing the change of the population over time. Traditionally, the functional forms used to describe these rates have depended on the particular system under study as well as on the type of data available, but the general trend has been to fit some kind of linear model (generalized, additive, mixed, etc.) to the data on each vital rate separately (Metcalf et al. 2013; Merow et al. 2014). Using linear models allows for a relatively small number of parameters to describe mathematically each vital rate, greatly reducing parameter estimation problems when compared with matrix projection models (Easterling, Ellner & Dixon, 2000), while allowing traditional statistical hypothesis testing (Coulson 2012), and model selection to be performed at the level of each vital rate (Metcalf et al. 2013; Merow et al. 2014).

The reduced complexity of IProjMs, coupled with the statistical framework provided by IPopMs, allows for the study of scenarios where very few data are available on some vital rates, up to the point where none is available other than the outcome of their interplay: population size and structure. This last scenario is an inverse problem.

In a general sense, an inverse problem involves taking the results of a series of processes and attempting to estimate these processes. Inverse problems have appeared in many areas of research including astronomy, astrophysics, remote sensing and hydrology (Ambartsumian 1980; Natterer 1986; Brown 1995; Carrera et al. 2005). As can be deduced from their name, every inverse problem must strictly relate to a direct problem. In our case, the direct problem would be to relate the vital rates in a mechanistic way, so that the outcome of such relation are the population sizes and structures observed over time. This is what an IProjM does. Using matrix projection models, Fournier, Hampton & Sibert (1998) showed that an inverse problem approach works in the context of fisheries species, while Ghosh, Gelfand & Clark (2012) have used IProjMs, using population structures as input to a similar problem.

In this paper, we explore the inverse problem of inferring some or all of the structured vital rates associated to a population for which a time series of population sizes and structures is available. We use simulated data to evaluate how well this approach works for the case where an IProjM kernel is constant through time, i.e. the population dynamics remains constant, but the population structure has not reached its asymptotic state. We show that the less information on the vital rates is available, the harder it is to find the correct reconstruction under a likelihood approach. Also, we show that the length of the time series is more important than per-year individual measurements to provide reliable reconstructions. Finally, we discuss the problems we faced when trying to identify accurate estimates for the parameters of this challenging inverse problem.

**Materials and methods**

***Model***

Easterling, Ellner & Dixon (2000) presented the first version of an IProjM. It this paper, a single variable structures the population, size (*x*), and the vital rates are assumed to remain constant over time. Under these circumstances, the following equation relates the population vital rates with the population structure (*n*) at any point in time:

*nt*+1(*y*) = ∫[*s*(*x*)⋅*g*(*y*,*x*) + *f*1(*x*)⋅*f*2(*y*,*x*)]⋅*nt*(*x*)*dx,* eqn. 1

where *nt*(*x*) is the proportion of individuals of size *x* at time *t*; *nt*+1(*y*), the proportion of individuals of size *y* at time *t*+1; and the vital rates are: *s*(*x*), the survival probability of extant *x-*size individuals; *g*(*y*,*x*), the probability an individual has of changing from size *x* to *y* from one unit time to the next; *f*1(*x*), the number of newborns produced by an *x-*size individual each time unit, and *f*2(*y*,*x*), the size distribution of newborns produced by *x-*size individuals.

As eqn. 1 shows, an IProjM relates the vital rates with a time series of population structures (*n*1, *n*2, ...), and the time series of population sizes given by the integration of each population structure over the observed size range (*N*1 *=* ∫ *n*1(*x*)*dx*, *N*2 *=* ∫ *n*2(*x*)*dx*, ...). The traditional approach has been to use as input to the model the former. Here, we use the latter as input, thus posing an inverse problem.

Many alternative structures to describe which and how vital rates interact to produce the size structure exist. As a case study, we use an IProjM with the following structure, close to Easterling, Ellner & Dixon (2000) original model:

*s*(*x*) = logit(*β*0 + *β*1⋅*x + β*2⋅*x*2), eqns. 2

*g*(*y*,*x*) ~ Normal(*μ* = *β*3 + *β*4⋅*x*, *σ* = exp(*β*5)),

*f*1(*x*) = exp(*β*6 + *β*7⋅*x*),

*f*2(*y*) ~ Normal(*μ* = *β*8, *σ* = exp(*β*9)).

Note that the vital rates are determined by the value of 10 low-level parameters (*β*0, ..., *β*9), and thus that the values of these parameters are sufficient to model the vital rates. Thus the inverse problem consists of estimating these parameters using as data a time series of observed population sizes and structures.

Additionally, we explored simpler scenarios where information on some vital rate(s) is available. In many cases, population ecologists have limited information on the vital rates and they would like to infer the vital rates that are missing. Therefore, we used as input to the model those observed vital rates and a time series of population sizes and structures. We explored all possible scenarios where one, two or three vital rates are unknown, with the inverse problem being the limiting scenario where no information on the vital rates is available.

***Parameter estimation***

To assess whether any given set of low-level IProjM parameter values is able to reproduce the observed time series, we substituted these in eqns 2. We then calculated the time series of population structures through the iteration of equation 1, and the time series of population sizes by integrating the structures (*Ni =* ∫ *ni*(*x*)*dx*). We used as initial population structure, *n*0(*x*), the first observed size structure.

The observed and estimated time series were compared through a composite log-likelihood function. We have two log-likelihood functions (see Appendix 1 for details): one that measures the goodness of fit between the estimated and observed population structures, *ln*, and another that measures the goodness of fit between the estimated and observed population sizes, *lN*. Therefore, to have an overall goodness of fit, we used the composite log-likelihood function:

*l* = *w*⋅*ln* + (1 – *w*)⋅*lN*, eqn. 3

where *w* is a weighting factor of the relative importance of fitting the former vs. the latter. Since we did not want to favor one fit over the other, we set *w* = *T*/(*T +* Σ*Ni*). By doing this, we account for the fact that *T* observed population sizes relate to Σ*Ni* measured individuals.

We used a hybrid approach for the maximization of *l*. This approach used two optimization algorithms, one heuristic and one derivative-based. The heuristic part consisted of a generalized simulated annealing algorithm (GenSA; Xiang & Gong 2000) to coarsely explore the likelihood function within a large hypercube of possible parameter values. This large hypercube was selected to incorporate as many conceivable population dynamics as possible (see Appendix 1 for details). However, within this large hypercube infinite the likelihood function can reach infinite values in those scenarios where the population grows too quickly (white areas in Fig. 2). This happens because we are dealing with an exponential model. Thus, a derivative-based algorithm is not useful since the derivative has to be finite. However, we wanted to keep the advantage of a derivative algorithm, since GenSA has the caveat that it can spend a lot of time exploring a local basin. Therefore, we used GenSA for a limited time and used the values provided by this algorithm as the starting point for a derivative-based algorithm (ADMB; Fournier et al. 2012) to further improve parameter estimation within a smaller region. This hybrid approach highly increases the probability that the global maximum is found. However, if the starting point where the optimization is started is too far from the actual maximum, a limited time may not be enough to reach the global maximum basin. To guarantee that this basin was reached within a reasonable time frame (~ 1.5 h) and to evaluate success ratio in reaching the global maximum, we ran the model in parallel with 100 starting points (selected with a Latin hypercube sampler; Fig.2, starting point of the cyan segments) in this large hypercube and restricted the GenSA part of the optimization algorithm to run for 1 h.

The model was coded in C++ and integrated into the R environment (R Core Team 2014) using Rcpp (Eddelbuettel et al. 2011). We used the GenSA package (Xiang et al. 2013) to run the heuristic part of the optimization procedure and the solution provided by this package was then used as input to ADMB through the R2admb package (Bolker, Skaug & Laake 2012). We used the lhs package (Carnell 2012) to obtain the starting points.

For purposes of result visualization we performed a principal component analysis (PCA) having as its centre the solution found by ADMB when starting at the known parameter values and as scale the standard deviations of the solutions found by the GenSA+ADMB algorithm. We used first two principal components (explained variance = 88.12%). This allowed visualizing the solutions in the plane of highest variation. Likelihoods were calculated on this plane to show likelihood contours.

***Simulations***

We simulated a population subject to structured vital rates constant over time. Starting with known values for the parameter estimates (*βi*'s in eqns. 2; Fig. 1a, b, d, e), we simulated a population that, starting with 10,000 individuals, followed the dynamics given by the modelled vital rates (eqns. 2) over 100 years (Fig. 1c). We used as the starting population structure the one obtained after iterating once eqn. 1 using a uniform distribution as *n*0. We chose the population dynamics so that the population did not reach its stable state over the timespan under consideration. The time series of population sizes and structures produced each year by this population was used as input to the model (Fig. 1c, f). We ran the 15 possible scenarios where three, two, one or all of the vital rates are unknown by having fixed the corresponding parameters at the known values. We also explored the impact on parameter estimation of reducing the number of individuals measured on each year (50, 90 and 99% reduction in per-year data), as well as the number of years for which data were gathered (50-, 20- and 10-year time series). To see how these scenarios affected parameter estimation, we estimated confidence intervals through likelihood profiling (*lprof* command in ADMB; Normal approximation) and used an Markov chain Monte Carlo (MCMC) approach to estimate the marginal distributions of the parameters within the hypercube (*mcmc* command in ADMB). Finally, we calculated confidence intervals for the mean vital rates using Normal approximation.

***Real data***

***TO BE DONE IF TIME ALLOWS...***

**Results**

***Inverse problem: no information on the vital rates***

Using the hybrid maximization approach to find the global maximum of the likelihood function in a scenario of no information on the vital rates allowed for a 91.7% success ratio in finding such maximum (Fig. 2). Starting the hybrid algorithm at 100 random starting points (Fig. 2; start point of cyan segments), 33 out of 100 reached the known solution (Fig. 2, orange segments reaching the green point; Table 1, first row); 61 reached some parameter bound and stopped there (Appendix 3); for three, ADMB estimated a non-positive definite Hessian and stopped the optimization process; and three starting points finished at different local maxima (Appendix 3). As can be seen on Fig. 2, GenSA can leave regions where the likelihood function is not finite (Fig. 2, white areas) and usually takes a starting point to a parameter bound or a to a local maximum basin (Fig. 2, cyan lines). Starting from a local basin, ADMB quickly reaches a local maximum (Fig. 2, orange lines on inset). Since points that reached a bound or stopped where the Hessian was non-positive definite are easy to identify and discard, we have that 36 starting points reached some maximum and 33 out of those were the global one. We therefore had a 33/36 = 0.917 success ratio (Table 1; first row).

The global maximum found by the algorithm closely matches the known vital rates from which the data were simulated (Fig. 1, black vs grey lines; Table 2). However, as the confidence bands show (Fig. 1; dashed grey lines), some parameters are easier to estimate than others (Table 2). Narrow confidence intervals were obtained for the parameters associated to the distribution of newborn sizes, the mean of the distribution of adult sizes, and the intercept and linear slope of the survival function. Fecundity was the vital rate that was harder to estimate, with normal confidence intervals been poor approximations since they went beyond their marginal distribution limits (cf. normal and marginal distribution limits in Table 2). The same happened with the estimation of the quadratic term in the survival function. Grow variation was also poorly estimated, although the confidence intervals remained within sensible values, possibly because the likelihood function is normal locally (Appendix 4). This difficulty in accurately estimating these parameters is because it is in these parameters that the likelihood function presents more than one local maximum and the marginal distributions are not normal (Appendices 3, 4).

***Scenarios of limited information on the vital rates***

When limited information on the vital rates is available, better success ratios are achieved, ranging from 92 to 100% (Table 1). As expected, the more information is available on the vital rates, the more likely it is that any starting point will reach the global maximum, to the point that, when only one vital rate is estimated, a single maximum exists. This is also reflected in the reduced width of the normal confidence intervals as more information is provided to the model (Appendix 2).

For the scenarios where two vital rates are missing, the number of starting points reaching the global maximum decreases. This occurs in the cases where information on growth is missing, where the number of starting points that reach parameter bounds or end up in flat regions or local maxima increases. This is highly due to the difficulty of estimating growth variation around the mean, the only parameter that is hard to estimate for this vital rate.

When information on only one vital rate is available, success ratios, similar to those scenarios where two vital rates are missing, are obtained. This is because the proportion of starting points that end up at a parameter bound or flat region increases, but not the number of local maxima. Again, success ratios decrease when no information on growth is available.

***Scenarios of reduced information on the time series***

We found that reducing the number of years under study had a larger impact on accurate parameter estimation than per-year sample size (Table 3). When a whole 100-year time series was available, 2600 individual measurements proportionally distributed along the time series were enough to correctly estimate the vital rates with a 100% success ratio. On the other hand, a 10-year time series did not provide enough information on the vital rates to correctly estimate them, even when the population started having a large population size and the starting point were the parameter values from which the data were simulated.

A time series of 20 years gave limit results. When size data was available for a large population (138,000 measurements over the entire time series), the model correctly reconstructed the vital rates with a 61% success ratio. This rate decreased as sample size decreased, up to the point where the model could not estimate the vital rates when only 1367 size measurements were available for those 20 years. This is also reflected in the widening of the confidence intervals for the parameter estimates under the different scenarios (Appendix 2).

**Discussion**

IProjMs have shown wide applicability (e.g. the examples presented in this special issue). However, following enough individuals over a sensible timespan is a demanding activity in both monetary and human resources. Our aim here was to provide a method that both makes use of all the available information on the population, much in tune with what IPopMs seek, taking this approach to its limit: no information on the vital rates. To evaluate how such an approach would behave, the use of simulated data is the best strategy: by knowing beforehand the parameter values we want to reach and controlling sample size and sampling error, we can examine whether or not the estimated vital rates match those used to generate the data (Bolker et al. 2013). In our simulations, the fact that a third of the 100 starting points reached the solution shows that this inverse modelling approach can estimate parameters in a short time when running the model in parallel. Discarding all problematic endpoints, we have a 91.7% success ratio in identifying the correct reconstruction of the vital rates. With simulated data we were also able to show that an IProjM, when used as the function describing the population dynamics in an IPopM, is a useful procedure to estimate unobserved vital rates in a population structured by a continuous variable. Additionally, by changing the length of the time series and the per-year sample size, we found that time series length is more important that the number of organisms measured each year for accurate parameter estimation.

***Likelihood diagnosis***

One general aspect of inverse problems is that, because they are less constrained than traditional statistical modelling approaches, they can easily incorporate large amounts of biological complexity, possibly outrunning the capability of the data to inform parameter estimation. As the capabilities of estimation using complex ecological models has exploded in the past few decades, ecologists have had to confront the problems of parameter identifiability and estimation, such as the existence of multiple maxima, i.e., qualitatively different sets of parameter values, each fitting the data better than any parameters in their respective neighbourhoods (Bolker 2008). Despite the repeated cautions from sophisticated modellers on these topics (REF), there are still few concrete, worked examples showing general ecologists how they can diagnose and mitigate such problems (e.g. by simplifying or constraining models) should they arise.

Here, we performed different analyses to visualize and understand the likelihood surface associated to this inverse problem. Since we are dealing with a 10-parameter model, exploring how these parameters interact in terms of their fit to the data, visualization tools become important to understand how an optimization procedure is performing. Pairwise likelihood slices allow visually identifying the regions where the likelihood function takes infinite values, and where the different starting points end up after the optimization procedure. This allows identifying problems with points reaching parameter bounds, flat regions or local maxima.

MCMC sampling allows visualizing the likelihood surface integrated over all but one parameter. Although usually used within the context of Bayesian inference, MCMC sampling produces, when no priors are imposed, the marginal distribution of the parameters. This allows determining if the parameter has a unimodal distribution and how big variation around this mode is. The parameters that are harder to identify are those that present more than one mode and/or present a plateau rather than a bell shape. They also allow identifying if the proposed bounds can be reduced for that particular model and set of data and give an idea of the impact a prior would have on parameter estimation (as was the case of the parameter associated to growth variation).

***Model optimization***

Efficient nonlinear optimization is a challenging task (Bolker 2008; Bolker et al. 2013). Even with a limited number of parameters, finding the values that maximize the model likelihood is difficult. A derivative approach (e.g., automatic differentiation) is usually a good alternative when the likelihood function is differentiable everywhere and the Hessian is positive defined everywhere, i.e., when the likelihood function is smooth and has no flat regions (Bolker 2008). In such cases, the approach is able to find a global maximum in a relative short time frame (Laake, Johnson & Conn 2013; Raue et al. 2013).

On the other hand, a heuristic approach, such as simulated annealing (Kirkpatrick et al. 1983; Bohachevsky et al. 1986; Suman & Kumar 2006), is more suited for problems where the likelihood function is not differentiable everywhere. The caveat of such approach is that it is time consuming. Therefore, the model is usually run with several starting points, taking as solution the one with the maximum likelihood; however, this does not guarantee that the global solution is found, and large time frames may be involved (Maunder & Punt, 2013).

With our particular problem, the likelihood function presents combinations of parameter values that produce non-finite values for the likelihood (numeric overflow), rendering in practical terms the function not differentiable for those values. Also, our likelihood function presents flat regions, making the Hessian non-positive defined. Therefore, the chances of finding a global maximum with a derivative approach were very low (1000 starting points may be required so that a single one reaches a known global maximum). This called for a different approach.

Our hybrid algorithm benefits from the advantages associated to each approach. By using GenSA to explore fast, but coarsely, a large area of possible parameter values for a reasonable time, we avoid those regions where a derivative approach is impractical. Once a smooth and non-zero-slope region is reached, ADMB quickly finds the local maximum. Running in parallel the model with 100 starting points for an hour greatly increased the changes of finding the global optimum.

Obviously, other optimization algorithms can be used in conjunction with a derivative algorithm. Better ones will probably be required for more complex IProjMs, for which the global maximum will be even harder to find. With a previous version of the model, particle swarm optimization was used with some level of success (González 2008), though the particular platform used at the time (Matlab) limited computational speed. Furthermore, a conjunction of derivative-free algorithms could also be used. As McMahon et al. (2014) show, combining GenSA with a Nelder-Mead algorithm (which, as ADMB, is good for exploring a local minimum basin) produces good results.

***Integrated integral projection models***

Although Schaub & Abadi (2011) clearly pointed towards the application of IPopMs to IProjMs, such work has, to our knowledge, not been performed yet. However, the idea of using multiple sources of information to inform the estimation of vital rates is not new. Partial information is a constant in population ecology, and IPopMs allows for the use of all possible sources of data into the estimation of the vital rates, which are the usually the harder part of any population study.

Here we explored the scenarios where one, two, or three vital rates were unknown, showing how our method performs with every possible scenario of restricted information on these rates. Results show that estimating one vital rate is straightforward, since one starting value will be enough to find the correct parameter values associated to the missing vital rates.

When information on two or three vital rates is missing, the identity of the one that is missing becomes relevant. The model behaves well when information on growth is available. When this information is missing, which is usually the case in those studies where organisms are not marked, the likelihood surface becomes complex and flat regions appear, as well as potentially non-realistic maxima. The same happens with the scenario of no information on the vital rates.

Under the scenario of no information on the vital rates, the model requires multiple starting points since the changes of ending on a flat region or non-realistic maximum are high. By putting biologically realistic bounds we can easily discard those starting points going onto unrealistic regions in the likelihood surface. Flat regions cannot be avoided, and large regions will require increasing the number of starting points. Fortunately, ADMB tags points that have reached such a region, so they can be discarded a posteriori. Therefore, the problem of find the maximum becomes a problem of computer time, and not of identifiability.

With our simulated data we had a success ratio of 91.7% in accurately estimating the known vital rates. This is in line with previous work we have done with a more complex IProjM (González & Martorell 2013). In such case a 90% success rate was obtained, although these rates are not strictly comparable as the latter refers to the number of simulated population dynamics for which good reconstructions were obtained. Nonetheless, the fact that the inverse problem works with different IProjMs and different population dynamics further increases our confidence in the approach.

***Reduced information on the time series***

However, a high success ratio depends on having a good time series of population structures and sizes, a feature expressed by its performance when shorter/smaller databases are provided. Our results show that parameter non-identifiability will appear in the case where short time series are available. Surprisingly, 10-year time series were not long enough to correctly estimate parameters and this may explain, at least to some extent, why the previous implementation of the approach with real data for a time series with 10 observed time points poorly estimated some parameters (González & Martorell 2013).

Unfortunately, short time series of small populations is the scenario most population ecologists have as their study case. In this case, data (or sensible estimates) on some vital rates will be required. Apparently, information on growth variation will be the estimate that will probably improve parameter estimation. Therefore, focus on providing some kind of information on this variation will prove rewarding.

Finally, the fact that the model performs well when limited information on per-year sample sizes is available becomes an asset in the context of conservation ecology, because small population sizes are the norm when dealing with endangered or endemic species.

***Population-level models***

Fitting observed population structures to an IProjM structure has been seen as a way to use population-level information to infer population-level vital rates. The idea behind this approach is that individual-level information gathered through the follow-up of the life cycle of individuals in a population does not necessarily capture the processes that occur at the population level. To remedy this, population structures have been seen as a better source of information from where to infer the vital rates (Ghosh, Gelfand & Clark 2012; Gelfand, Ghosh & Clark 2013). We recognize that this approach converges to a similar statistical method to the one we use to infer the vital rates.

Under such population-level conception, we believe that population size is also another piece of valuable information that would be useful to incorporate into the estimation of the parameters of an IProjM. This is supported by the use of population size to infer vital rates using matrix projection models (Fournier, Hampton & Sibert 1998).

In our model, we incorporated both sources of information through a composite likelihood in which a weight (*w* in eqn. 3) control the relative importance of fitting the size structures vs that of the population sizes. González & Martorell (2013) showed that weight values close to zero give the best results. This is accordance with the *w*-value we used here, which is based on a logical argument: *T* observed population sizes relate to Σ*Ni* measured individuals.

***Limitations of the method***

As with all statistical models, confidence in the estimated parameter values depends on data availability. The smaller the data set, the more difficult will be to provide accurate reconstructions of the vital rates. Data availability relates to both the number of years for which population size and structure were recorded, as well as on the per-year number of individuals for which the structuring variable was measured. Here we show that the first seems to be more important for accurate reconstructions than the latter.

The structure of the IProjM kernel is also important. If the functions that constitute the kernel do not conform to how actually the population behaves, the reconstructed vital rates will not reflect how the population dynamics really is. This is a problem shared by all statistical models: if the model that one is proposing does not reflect the actual structure of the system, low confidence can be put on the conclusions we derive from it. By using different equations to describe the relation between the vital rates and the structuring variable, and performing traditional model selection/averaging could potentially overcome this problem. However, given the nonlinearity of the problem, the reconstructed vital rates produced by each model should be evaluated visually to confirm that the models under consideration are not associated to grossly dissimilar reconstructions.

Here we assume that a static IProjM is a correct model to describe the dynamics of the population for which data are available. This is usually not the case, as the environment is hardly ever constant over time. However, exploring the feasibility of reconstructing the vital rates using a simple IProjM, as well as establishing which vital rates are harder to estimate, can serve as a starting point to explore more complex (and realistic) IProjMs.

***Future research***

Previously, we have found that the global maximum found by the optimization algorithm may not be biologically realistic (González & Martorell 2013). Visual examination of the solutions to discard unrealistic ones is always a reasonable approach, but we believe that constraining the parameter space to only realistic solutions is a more promising alternative. This can be seen, in an optimization framework, as maximizing a multiobjective problem (Suman & Kumar 2005) and, in a Bayesian framework, as imposing weakly informative priors. This approach has already been explored using ADMB, but such priors create flat areas, which, as stated before, make the exploration of the likelihood surface difficult. The question still remains on how to do this without introducing considerable bias in parameter estimation. This would be the case of the parameter describing growth variation. As shown by its marginal distribution (Appendix 4), the basin of the global maximum is relative small and the rest of the surface is mainly flat. Imposing a prior that does not have its mode on this basin would bias the estimation of this parameter. Nonetheless, given that a Bayesian framework has proved its usefulness with other inverse problems (Heisey, et al. 2010; Stuart 2010; Ghosh, Gelfand & Clark 2012), it makes sense trying to implement it with our problem.

**Acknowledgements**

EJG thanks Steve Walker and Guillaume Blanchet for their incredible help while understanding many of the statistical elements associated to this problem. Funding for this research was provided by the Mexican Consejo Nacional de Ciencia y Tecnología through a postdoctoral grant and by BMB...

**Data accessibility**

**References**

Ambartsumian, V.A. (1980). On some trends in the development of astrophysics. *Annual Review of Astronomy and Astrophysics*, 18, 1-13.

Besbeas, P., Freeman, S.N., Morgan, B.J. & Catchpole, E.A. (2002). Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, 58, 540–547.

Bolker, B.M. (2008). Ecological models and data in R. Princeton University Press, Princeton, NJ.

Bolker, B., Skaug, H. & Laake, J. (2012). R2admb: ADMB to R interface functions. R package version 0.7.9. http://CRAN.R-project.org/package=R2admb

Bolker, B.M., Gardner, B., Maunder, M., Berg, C.W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T., de Valpine, P., Ford, J., Gimenez, O., Kéry, M., Kim, E.J., Lennert-Cody, C., Magnusson, A., Martell, S., Nash, J., Nielsen, A., Regetz, J., Skaug, H. & Zipkin, E. (2013). Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, 4, 501–512.

Brown, J.C. (1995). Overview of topical issue on inverse problems in astronomy. *Inverse Problems*, 11, 635.

Carnell, R. (2012). lhs: Latin Hypercube Samples. R package version 0.10. http://CRAN.R-project.org/package=lhs [accessed 4 May 2015]

Carrera, J., Alcolea, A.S., Medina, A., Hidalgo, J. & Slooten, L.J. (2005). Inverse problem in hydrogeology. *Hydrogeology Journal*, 13, 206–222.

Coulson, T. (2012). Integral projections models, their construction and use in posing hypotheses in ecology. *Oikos*, 121, 1337–1350.

Easterling, M.R., Ellner, S.P. & Dixon, P.M. (2000). Size-specific sensitivity: applying a new structured population model. *Ecology*, 81, 694–708.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D. & Ushey, K. (2011). Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, 40, 1–18.

Fournier, D.A., Hampton, J. & Sibert, J.R. (1998). MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. *Canadian Journal of Fisheries and Aquatic Sciences*, 55, 2105–2116.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A. & Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27, 233–249.

Gelfand, A.E., Ghosh, S. & Clark, J.S. (2013). Scaling Integral Projection Models for Analyzing Size Demography. *Statistical Science*, 28, 641–658.

Ghosh, S., Gelfand, A.E. & Clark, J.S. (2012). Inference for size demography from point pattern data using integral projection models. *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 678–681.

González, E.J. (2008). *Modelación de la dinámica poblacional de cactáceas endémicas del Valle de Tehuacán a lo largo de un gradiente de disturbio antropogénico crónico.* MSc thesis, Universidad Nacional Autónoma de México, México, D.F.

González, E.J. & Martorell, C. (2013). Reconstructing shifts in vital rates driven by long-term environmental change: a new demographic method based on readily available data. *Ecology and Evolution*, 3, 2273–2284.

Heisey, D.M., Osnas, E.E., Cross, P.C., Joly, D.O., Langenberg, J.A. & Miller, M.W. (2010). Linking process to pattern: estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. *Ecological Monographs*, 80, 221–240.

Kéry, M. & Schaub, M. (2012). *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press, Waltham, MA.

Laake, J.L., Johnson, D.S. & Conn, P.B. (2013). marked: an R package for maximum likelihood and Markov Chain Monte Carlo analysis of capture-recapture data. *Methods in Ecology and Evolution*, 4, 885–890.

Lee, R.D. (1974). Estimating series of vital rates and age structures from Baptisms and burials: a new technique, with applications to pre-industrial England. *Population Studies: a Journal of Demography*, 28, 495–512.

Lee, R.D. (1985). Inverse projection and back projection: A critical appraisal, and comparative results for England, 1539 to 1871. *Population Studies*, 39, 233–248.

Maunder, M.N. & Punt, A.E. (2013). A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142, 61–74.

McMahon, S.M. & Parker, G.G. (2014). A general model of intra-annual tree growth using dendrometer bands. *Ecology and Evolution*, 5, 243–254.

Merow, C., Dahlgren, J.P., Metcalf, C.J.E., Childs, D.Z., Evans, M.E.K., Jongejans, E., Record, S., Rees, M., Salguero-Gómez, R. & McMahon, S.M. (2013). Advancing population ecology with integral projection models: a practical guide. *Methods in Ecology and Evolution*, 5, 99–110.

Metcalf, C.J.E., McMahon, S.M., Salguero-Gómez, R. & Jongejans, E. (2012). IPMpack: an R package for integral projection models *Methods in Ecology and Evolution*, 4, 195–200.

Natterer, F. (1986). *The mathematics of computerized tomography*. Wiley, Chichester.

R Development Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org [accessed 15 April 2015]

Raftery, A.E., Givens, G.H. & Zeh, J.E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, 90, 402–416.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U. & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25, 1923–1929.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B.D., Theis, F.J., Klingmüller, U. & Timmer, J. (2013). Lessons Learned from Quantitative Dynamical Modeling in Systems Biology. *PLoS ONE*, 8, e74335.

Schaub, M., Gimenez, O., Sierro, A. & Arlettaz, R. (2007). Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. *Conservation Biology*, 21, 945–955.

Schaub, M. & Abadi, F. (2010). Integrated population models: a novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, 152, 227–237.

Stuart, A.M. (2010). Inverse problems: a Bayesian perspective. *Acta Numerica*, 19, 451–559.

Suman, B. & Kumar, P. (2005). A survey of simulated annealing as a tool for single and multiobjective optimization. *Journal of the Operational Research Society*, 57, 1143–1160.

Tuljapurkar, S. & Caswell, H. (1997). *Structured-population Models in Marine, Terrestrial, and Freshwater Systems*, Vol. 18. Chapman & Hall, New York.

Wrigley, E.A., & Schofield, R. S. (1981). *The Population History of England, 1541–1871: A Reconstruction.* Harvard University Press, Cambridge, MA.

Xiang, Y. & Gong, X.G. (2000). Efficiency of generalized simulated annealing. *Physical Review E*, 62, 4473–4476.

Xiang, Y., Gubian, S., Suomela, B. & Hoeng, J. (2013). Generalized simulated annealing for global optimization: the GenSA Package. *R Journal*, 5, 13–28.

**Tables**

Table 1. Performance of the model under different scenarios of availability of information on the vital rates. Out of 100 starting points, some reach the global maximum (Global maximum), others a parameter bound (Bound), others a region where the Hessian was non-positive definite (Hessian problem), and others a different maximum (Other maxima). Success ratio is the percentage of starting points, out of those that reached some maximum, which reached the global one. Vital rates: *s*, survival; *g*, growth; *f1*, fecundity; and *f*2, newborn sizes (as in eqn. 1).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Unknown | Known | Global maximum | Bound | Hessian problem | Other maxima | Success ratio |
| *s, g, f*1, *f*2 | *–* | 33 | 61 | 3 | 3 | 91.67 |
| *g, f*1, *f*2 | *s* | 61 | 27 | 7 | 5 | 92.42 |
| *s, f*1, *f*2 | *g* | 100 | 0 | 0 | 0 | 100 |
| *s*, *g, f*2 | *f*1 | 62 | 23 | 13 | 2 | 96.87 |
| *s*, *g, f*1 | *f2* | 90 | 3 | 5 | 2 | 97.83 |
| *f*1, *f*2 | *s*, *g* | 100 | 0 | 0 | 0 | 100 |
| *g, f*2 | *s, f*1 | 88 | 0 | 4 | 8 | 91.67 |
| *g, f*1 | *s, f*2 | 90 | 5 | 3 | 2 | 97.83 |
| *s, f*2 | *g, f*1 | 100 | 0 | 0 | 0 | 100 |
| *s, f*1 | *g, f*2 | 100 | 0 | 0 | 0 | 100 |
| *s*, *g* | *f*1, *f*2 | 62 | 3 | 35 | 0 | 100 |
| *f2* | *s*, *g, f*1 | 100 | 0 | 0 | 0 | 100 |
| *f*1 | *s*, *g, f*2 | 100 | 0 | 0 | 0 | 100 |
| *g* | *s, f*1, *f*2 | 100 | 0 | 0 | 0 | 100 |
| *s* | *g, f*1, *f*2 | 100 | 0 | 0 | 0 | 100 |

Table 2. Estimates for the parameters of the model obtained with the simulated data. Known: values of the parameters from which the data were generated; Estimate: estimated parameter values; CV: coefficient of variation; Normal CI: confidence interval obtained through normal approximation of the likelihood profile; Marginal distributions were obtained through a Markov chain Monte Carlo procedure. Here all vital rates were assumed unknown.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Vital rate | Par | Known | Estimate | CV | Normal 95% conf. intervals | Hypercube intervals | Marginal distr. 95% limits |
| survival | *β0* | 1.9 | 1.884 | 0.144 | 1.343, 2.395 | -18, 22 | 0.72, 15.7 |
|  | *β1* | 0.4632 | 0.69 | 0.372 | 0.176, 1.184 | -9, 11 | -0.035, 4.89 |
|  | *β2* | 0.0046 | 0.01 | 15.1 | -0.292, 0.299 | -1.01, 0.99 | -0.16, 0.20 |
| growth | *β3* | 0.025 | 0.026 | 0.692 | -0.015, 0.065 | -4, 4 | -0.70, 2.145 |
|  | *β4* | 0.99 | 0.99 | 0.006 | 0.975, 1.003 | -0.5, 1.5 | 0.94, 1.50 |
|  | *β5* | -1.3 | -1.326 | -0.367 | -2.402, -0.293 | -11.513, 0.693 | -11.24, 0.05 |
| fecundity | *β6* | -2 | -1.708 | -0.805 | -4.754, 1.218 | -13.816, 13.816 | -13.14, 0.42 |
|  | *β7* | 0.3 | 0.278 | 2.176 | -1.061, 1.563 | 0, 2.878 | 0.050, 2.875 |
| newborn sizes | *β8* | -2 | -1.988 | -0.021 | -2.079, -1.901 | -11.513, 0 | -9.950, -0.010 |
| *β9* | -0.95 | -0.937 | -0.12 | -1.175, -0.690 | -4, 0 | -2.770, -0.005 |

Table 3. Performance of the inverse model under different scenarios of data availability: per-year sample size and time series length. Success ratio: percentage of starting points, out of those that reached some maximum, which reached the global one (i.e., those that reached parameter bounds or could not estimate the Hessian are not considered); total number of individual measurements in the entire time series is in parenthesis. Total time units: length of the time series. Percentage of size measurements per year: the population, starting from 10000 individuals, decreases over time (Fig.1). Here all vital rates were assumed unknown.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Success ratio  (ΣN*i*) | | Total time units (*T*) | | | |
| 100 | 50 | 20 | 10 |
| Percentage of size measurements per year | 100% | 91.67 (264208) | 75 (219713) | 61.29 (137692) | - (83309) |
| 50% | 96.39 (132077) | 79.41 (109842) | 52 (68841) | - (41652) |
| 10% | 98.87 (26383) | 82.76 (21953) | 32 (13761) | - (8327) |
| 1% | 100 (2596) | 100 (2173) | -  (1367) | -  (829) |

**Figures**

Fig. 1. Simulated data and reconstructed vital rates. The population followed constant vital rates through time (black lines; a: survival, b: growth, d: fecundity, e: newborn size distribution), which produced, over a 100 years, a time series of population sizes (c) and structures (f). Reconstructed vital rates (grey solid lines) when they are assumed unknown; Normal confidence intervals in dashed.

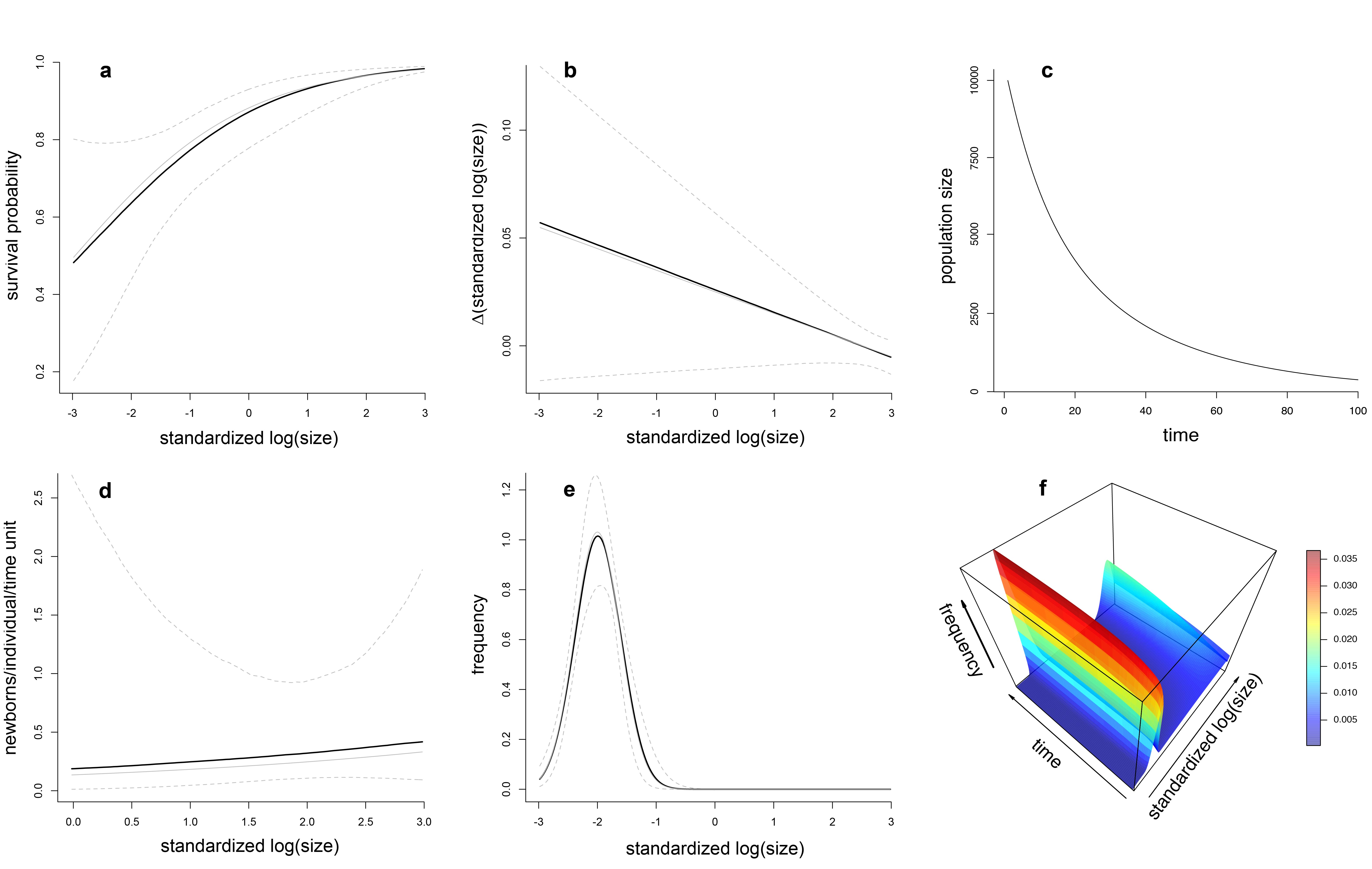
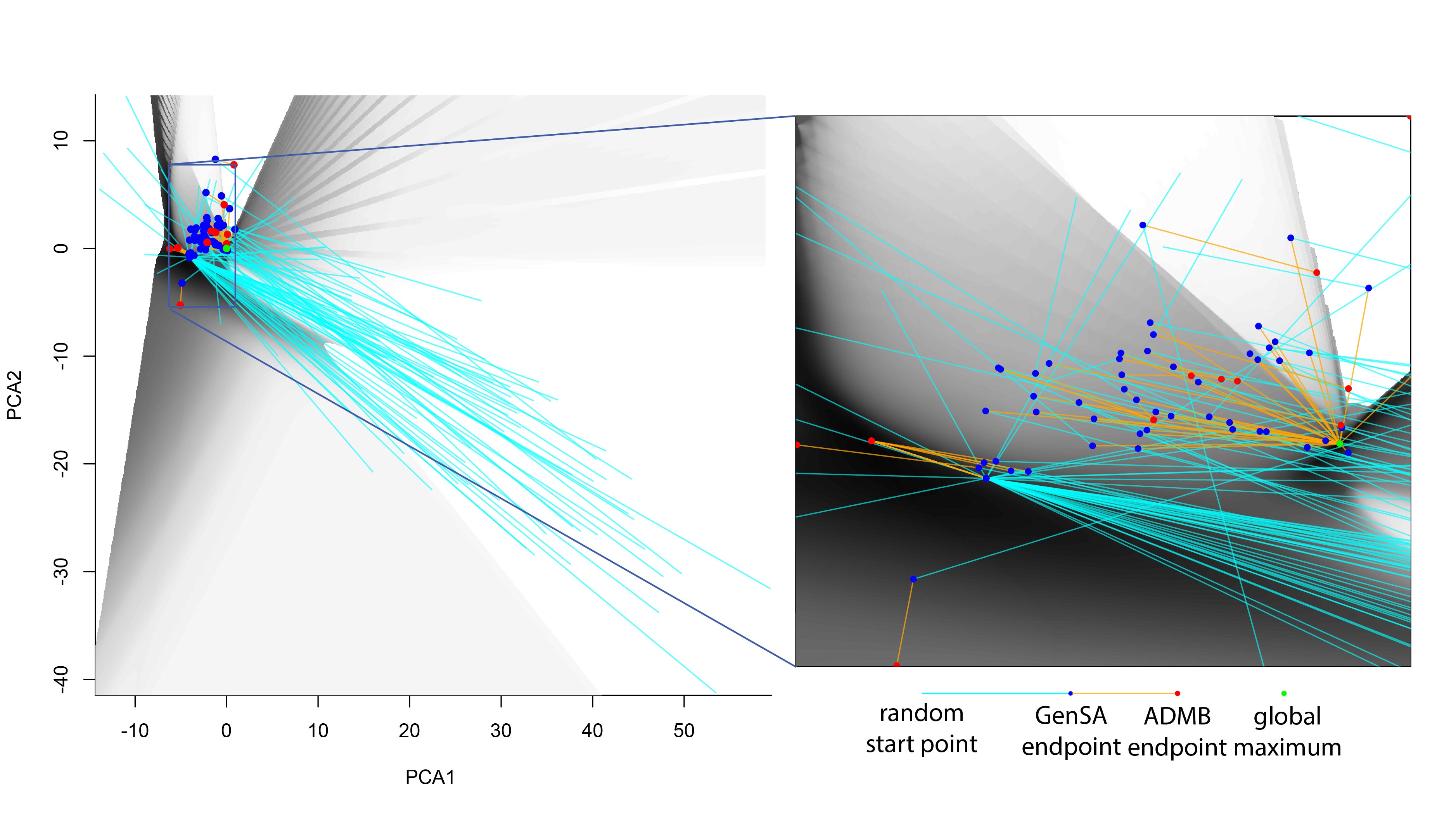


Fig. 2. Segments connecting starting and endpoints reached by the hybrid algorithm. Axes are the two principal components associated to the endpoints (centred on the global maximum and scaled to the standard deviations of these points) and the contours correspond to the likelihood values in these axes. The heuristic algorithm starts at a random point and reaches a point after 1h; the derivative algorithm takes this point and reaches a better solution within a local basin. This allowed 33 out of 100 starting points to reach the global maximum (green). White regions are where the likelihood takes non-finite values. Because starting points were selected randomly in the hypercube of parameter values (see Appendix 3), they do not necessarily look random in this distorted representation.



**Appendix 1**

Description of the likelihood approach used to evaluate goodness of fit between the estimated time series of population sizes and structures and the observed one.

**Appendix 2**

Normal confidence intervals for the estimated parameters under the 15 scenarios of data availability on the vital rates, and the nine scenarios of data availability on the per-year individual measurements and on the time series length.

**Appendix 3**

Likelihood slices for each pair of parameters in the model.

**Appendix 4**

Marginal distributions of the parameters in the model.