**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

**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

**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 the painstaking tracking of marked individuals in a population 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 modelling approach: we infer the structured vital rates from the time series. We simulated different population dynamics 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 explore cases where partial information on the vital rates is available. We also apply the method to data from a plant system.

3. We show that this approach provides good reconstructions of the vital rates. The more information on the vital rates is provided the easier it is to find a correct reconstruction.

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, is costly or even impracticable.

**Key-words**

inverse modelling, vital rates, population distribution, population density, integral projection model, integrated population model

**Introduction**

The number and distribution of individuals within a population observed at any point in time are the result of several demographic processes interacting through time (REF). Such processes involve, at the very least, the birth, growth, reproduction and death of individuals. Usually, individuals do not behave homogeneously due to differences in some qualitative/quantitative traits or environmental factors, which increase variation among individuals and thus structure the population (Tuljapurkar & Caswell, 1997; Merow et al. 2014).

When the variable that structures a population is continuous, an integral projection model (IPM1) is the standard tool to analyse how vital rates interact. Easterling, Ellner & Dixon (2000) presented the first version of an IPM1 described as

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

where *n*(*x*,*t*) is the distribution of individuals according to a structuring variable *x*, and *s*(*x*), *g*(*y*,*x*), *f*1(*x*) and *f*2(*y*,*x*) are the survival probability and growth of extant individuals, and the number and sizes of newborns, respectively. Traditionally, the particular functional structure of these latter functions has depended on the system under study as well as on type of data that was gathered, but the general trend has been to take each vital rate separately and fit some kind of linear model (generalized, additive, mixed, etc.) to the available data (Metcalf et al. 2013; Merow et al. 2014). Using linear models allows for a limited number of parameters to determine the structure of each vital rate, largely 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 at the level of each vital rate (Metcalf et al. 2013; Merow et al. 2014).

In a parallel line of research, integrated population models (IPM2s) have been proposed as a single statistical framework were data from different sources can be used to better inform demographic parameters (Raftery, Givens & Zeh, 1995; Besbeas et al. 2002; Brooks et al. 2004; Schaub & Abadi, 2011; Kéry & Schaub 2012). An additional advantage is that they allow accounting for parameter and model uncertainty in our conclusions, an element that has largely been overlooked when fitting IPM1s.

Within an IPM2 framework previous work has been done where only partial information on the vital rates is available. However, the species under study have not been considered as having their populations structured by a continuous variable and have mostly restricted to vertebrate species (e.g. Schaub et al. 2007; Kéry & Schaub 2012; Maunder & Punt 2013; but cf. McMahon & Parker 2014), which have their own modelling peculiarities.

Although not stated as IPM2s, a lot of research has been done to infer information on unobserved vital rates from limited observed vital rates (REFS, McMahon & Parker 2013). This method can be traced back to inverse/backward projection models in human populations where we try to infer mortality and fertility rates, along with population age structures, from annual births and deaths records (Lee 1974, 1985; Wrigley & Schofield 1981; Oeppen 1993).

The reduced complexity of IPM1s, coupled with the statistical framework provided by IPM2s, allowed us to explore the viability of an inverse modelling approach. An inverse model, in a general sense, takes the results of a series of processes and attempts 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 its name, every inverse model has to strictly relate to a direct model. In our case, the direct model would be an IPM1 where the vital rates are used as input and the population sizes and structures are its output. Using matrix projection models, Fournier, Hampton & Sibert (1998) showed that this inverse approach works in the context of fisheries species, while Ghosh, Gelfand & Clark (2012) used IPM1s, but only using population structures as input.

In this paper, we explore inverse modelling to infer some or all 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 show that this approach works for the case where an IPM1 kernel is constant through time, i.e. the population dynamics remains constant, but the population structure has not reached its asymptotic state. We show the method is capable of correctly reconstructing the vital rates, but, as would be expected with any statistical problem, the less information on the vital rates is available, the harder it is to find the correct reconstruction. Also, as expected, the less information you have (in terms of sample size and length of the time series), the harder it is to provide reliable reconstructions.

**Materials and methods**

***Model***

As can be seen in equation 1, an IPM1 relates the population vital rates (*s*, *g*, *f1*, *f2*) with a time series of population structures (*n*(*x*,1), *n*(*x*,2), ...), and the time series of population sizes given by the addition (integration) of the population structures (*N*1 *=* ∫*n*(*x*,1)*dx*, *N*2 *=* ∫*n*(*x*,2)*dx*, ...). The traditional approach has been to use as input to the model the former. Here, we use the latter as input, thus having an inverse problem.

In their simplest structure, the functions describing the relation between the vital rates and the structuring variables can be assumed to be constant over time and that a single variable structures the population (Crouse, Crowder & Caswell 1987; Easterling, Ellner & Dixon 2000). As a case study, we used an IPM1 with such structure, using the following equations to describe the vital rates as a function of a continuous variable *x*:

*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*), and

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

Note that the vital rates are determined by the value of 10 parameters (*β*0, ..., *β*9).

We explored the scenarios where information on some vital rates 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, being the inverse problem the limit scenario where no information on the vital rates is available.

***Parameter estimation***

For the set of low-level IPM1 parameters, *β*0, ..., *β*9, and an observed time series of population structures, *n*(*x*,1), *n*(*x*,2), ..., *n*(*x*,*T*), and sizes, *N*1, *N*2, ... *NT*, we have a composite likelihood function that measures the goodness of fit of the model to the data. We have two log-likelihood functions: 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*. We measure the overall goodness of fit through 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*).

We used a hybrid approach for the maximization of *l*. This approach used two optimization algorithms: one heuristic and one gradient-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, becoming in practical terms an unbounded problem. Based on previous work, we knew that the likelihood function presented many local maxima and thus an entirely gradient-based algorithm would not work efficiently. However, we wanted to keep the advantage of a gradient-based algorithm since GenSA has the caveat that it is time consuming to explore a local basin. Therefore, once a local maximum was identified, we used a gradient-based algorithm (ADMB; Fournier et al. 2012) to further improve parameter estimation within such basin.

This hybrid approach highly increases the probability that the global maximum is found, although it may take a lot of time to find it (e.g., starting at a random point in this large hypercube, it took this approach up to 6.3e7 evaluations of the model, ~15 h, to find a known global maximum). To reach the correct solution within a reasonable time frame (< 1 h), we ran the model in parallel with 100 starting points (selected with a Latin hypercube sampler) in this large hypercube and restricted the GenSA part of the optimization algorithm to run up to 0.5 h.

The model was coded in C++ and used within 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 GenSA 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.

***Simulations***

We simulated a population subject to a static population dynamics not in its stable state. Starting with known values for the parameter estimates (*β*i's in eqns. 2; Fig. 1A), we simulated a population that, starting with 10,000 individuals, followed the dynamics given by the modelled vital rates (eqns. 2) over 100 years. 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. 1B). 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, as well as the number of years for which data were gathered. To see how this affected parameter accuracy we estimated confidence intervals through likelihood profiling (*lprof* option in ADMB; Normal approximation).

***Real data***

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

**Results**

***TO BE WRITTEN...***

**Discussion**

Being an IPM2, this model makes a number of assumptions that should be fulfilled to ensure an accurate and unbiased estimation of parameters (Schaub & Abadi, 2011).

We assume that a static IPM1 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 constant through time. However, we consider that showing that the method works with a simple IPM1 will serve as a starting point to explore more complex IPM1 structures.

***Simultaneous model fitting***

In a traditional IPM1, vital rates are estimated by separately fitting generalized linear models (GLM) to individual data (Merrow et al. 2014). Our inverse model can be seen as estimating all vital rates at once using a single dataset.

***Model optimization***

Efficient nonlinear function optimization is a challenging task. Even with a limited number of parameters, finding the values that maximize the model likelihood is difficult. Using a gradient approach (automatic differentiation) the identification of a global maximum is not guaranteed. Therefore, usually the model is ran with several starting points, taking as solution the one with the maximum likelihood, but this does not guarantee that the global solution is found (Maunder & Punt, 2013). With our particular problem (10 parameters in the version of the model presented here, but 18 in a more complex version), the likelihood function is filled with local maxima and the chances of finding the global one are very low (1000 starting points may be required so that a single one reaches the global maximum). This called for a different approach.

Simulated annealing has been shown to be a simple algorithm to find the global maximum of a non-linear function (Kirkpatrick et al. 1983; Bohachevsky et al. 1986; Suman & Kumar 2006). In its generalized version...

However, other heuristic optimization algorithms could be used to explore the likelihood function, especially for more complex applications of this method where 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 2012), though the particular platform used at the time (Matlab) limited time efficiency.

***Integrated integral projection models***

Although Schaub & Abadi (2011) clearly pointed towards the application of IPM2s to IPM1s, 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 the use of

Here we explored the scenarios where one, two, or three vital rates were known...

Results show that the estimation...

***Population-level models***

Fitting observed population structures to an IPM1 structure has been seen as a strategy 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 captures 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). This approach, presented by Ghosh, Gelfand & Clark (2012), converges to a similar statistical method to infer the vital rates. Under such population-level conception, we believe that population size is also another piece of information that would be useful to incorporate into the estimation of the parameters of an IPM1. This is supported by previous work done with matrix projection models (Fournier, Hampton & Sibert 1998).

***Limitations of the method***

As with all statistical models, confidence in the estimated parameter values depends on data availability. The fewer data the more difficult will be to have confident 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. This

Also, the structure of the IPM1 kernel is important. If the functions that constitute the kernel do not conform to how actually the population behaves, the reconstructed vital rates might not reflect how the population dynamics actually is. This, again, is a problem of 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. This problem would be partially overcome by the fitting of different kernel structures and performing traditional model selection/averaging to get better parameter estimates. However, given the nonlinearity of the problem, careful evaluation of the reconstructed vital rates given by each model should be done during this selection process.

***Future research***

IPM1s have mushroomed into a panoply of applications (e.g. the many examples in this special issue). However, following enough individuals over a sensible timespan is a demanding activity in both monetary and human resources. My aim here is to provide a method that both makes use of all the available information on the population, much in tune with what IPM2 seek, taking this approach to its limit: no information on the vital rates.

A foreseeable challenge may arise if the global maximum sought by the optimization algorithm is not biologically realistic. Visual examination of the solutions to discard unrealistic ones is always a possibility (González & Martorell 2013), 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) and, in a Bayesian framework, as imposing weakly informative priors. This approach has already been explored using a gradient-based approach but such priors create flat areas, which in turn make difficult the exploration of the likelihood surface. Hopefully, the hybrid optimization presented here will not suffer from this problem. The question still remains on how to do this without introducing considerable bias in parameter estimation, but it makes sense to use a Bayesian framework, as it has proved its usefulness with other inverse problems (Calvetti, Kaipio, Somersalo 2014).

**Acknowledgements**

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

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

Calvetti, D., Kaipio, J.P. & Somersalo, E. (2014). Inverse problems in the Bayesian framework. **30**, 110301.

Carnell, R. (2012). lhs: Latin Hypercube Samples. R package version 0.10. http://CRAN.R-project.org/package=lhs

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.

Crouse, D.T., Crowder, L.B. & Caswell, H. (1987). A stage-based population model for loggerhead sea turtles and implications for conservation. *Ecology*, **68**, 1412–1423.

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. & 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.

Kéry, M. & Schaub, M. (2012). *Bayesian population analysis using WinBUGS: a hierchical perspective*. Academic Press, Waltham, MA.

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 (M. Rees, Ed.). *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.

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.

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 vital rates. Success rate is the number of starting points, out of a 100, that reached the sought solution within 1 h. Vital rates: *s*, survival; *g*, growth; *f1*, fecundity; and *f*2, newborn sizes.

|  |  |  |
| --- | --- | --- |
| Unknown | Known | Success rate |
| *s* | *g, f*1, *f*2 | 100 |
| *g* | *s, f*1, *f*2 | 100 |
| *f*1 | *s*, *g, f*2 | 100 |
| *f2* | *s*, *g, f*1 | 100 |
| *s*, *g* | *f*1, *f*2 | 67 |
| *s, f*1 | *g, f*2 | 100 |
| *s, f*2 | *g, f*1 | 100 |
| *g, f*1 | *s, f*2 | 84 |
| *g, f*2 | *s, f*1 | 77 |
| *f*1, *f*2 | *s*, *g* | 100 |
| *s*, *g, f*1 | *f2* | 84 |
| *s*, *g, f*2 | *f*1 | 14 |
| *s, f*1, *f*2 | *g* | 13 |
| *g, f*1, *f*2 | *s* | 14 |
| *s, g, f*1, *f*2 | *–* | 5 |

Table 2. Estimates and confidence intervals obtained with the simulated data. Initial: values of the parameters from which the data were generated; Estimate: estimated values of the parameters using the model were all the vital rates were assumed unknown; Confidence interval: obtained through likelihood profiling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vital rate | Parameter | Initial | Estimate | Confidence interval |
| survival | *β0* | 1.9 | 1.8846 | 1.86585, 1.90549 |
| *β1* | 0.4632 | 0.69035 | 0.669349, 0.71216 |
| *β2* | 0.0046 | 0.010426 | 0.00302327, 0.0175478 |
| growth | *β3* | -1.3 | -1.326 | -1.37161, -1.28206 |
| *β4* | 0.025 | 0.025816 | 0.0240798, 0.0274866 |
| *β5* | 0.99 | 0.98962 | 0.989016, 0.990194 |
| fecundity | *β6* | -2 | -1.7085 | -4.75406, 1.21813 |
| *β7* | 0.3 | 0.27812 | 0.22149, 0.332621 |
| newborn sizes | *β8* | -0.95 | -0.9372 | -0.947687, -0.927102 |
| *β9* | -2 | -1.988 | -1.99189, -1.98435 |

Table 3. Performance of the model under different per-year sample sizes and number of years available. Here, all vital rates are assumed to be unknown.

***WORKING ON THIS...***

**Figures**

Fig. 1. Simulated data and reconstructed vital rates. The population followed constant vital rates through time (A), which produced, over a 100 years, a time series of population structures (B) and sizes (C). D: Reconstructed vital rates when no information on them is provided (confidence intervals in grey).

***WORKING ON THIS...***

Fig. 2. Likelihood slices for pairs of parameters.

***WORKING ON THIS...***