Universität
Zürich^UZH

Universitätseinheit

Individual-level causes and population-level consequences of variation in fitness in an alpine rodent

Timothée Bonnet

Contents

1

General introduction

"Begin at the beginning," the King said, gravely, "and go on till you come to an end; then stop."

— Lewis Carroll, *Alice in Wonderland*

2

Successful by chance? The power of mixed models and neutral simulations for the detection of individual fixed heterogeneity in fitness components

If the talents I was born with are the right ones, I may someday achieve my goal. If not, I may go through life being as stupid as I am now.
— Eiji Yoshikawa, *Musashi* (1935)

Quand on veut comprendre une chose, on se place en face d'elle, tout seul, sans secours; tout le passé du monde ne pourrait servir de rien. Et puis elle disparaît et ce qu'on a compris disparaît avec elle.
— Jean-Paul Sartre, *La nausée* (1938)

Timothée Bonnet* and Erik Postma (2016) The American Naturalist 187(1):60-74

2.1 Abstract

Heterogeneity in fitness components consists of fixed heterogeneity due to latent differences fixed throughout life (e.g. genetic variation), and dynamic heterogeneity generated by stochastic variation. Their relative magnitude is crucial for evolutionary processes, as only the former may allow for adaptation. However, the importance of fixed heterogeneity in small populations has recently been questioned. Using neutral simulations (NS), several studies failed to detect fixed heterogeneity, thus challenging previous results from mixed models (MM). To understand the causes of this discrepancy, we estimate the statistical power and false positive rate of both methods, and apply them to empirical data from a wild rodent population. While MM show high false positive rates if confounding factors are not accounted for, they have high statistical power to detect real fixed heterogeneity. In contrast, NS are also subject to high false positive rates, but always have low power. Indeed, MM analyses of the rodent population data show significant fixed heterogeneity in reproductive success, whereas NS analyses do not. We suggest that fixed heterogeneity may be more common than is suggested by NS, and that NS are useful only if more powerful methods are not applicable and if they are complemented by a power analysis.

Keywords: *Chionomys nivalis*; individual-based model; generalized linear mixed model; simulations; snow vole; statistical power

Online enhancements: Online appendices. Data available from the Dryad Digital Repository: *http://dx.doi.org/10.5061/dryad.3cb61*.

2.2 Introduction

Within species, individual variation in lifetime reproductive success (LRS) is plentiful, with most individuals producing few or no offspring and a few individuals producing a large share of the next generation (Clutton-Brock 1988; Stearns 1992). Given their skewed and heterogeneous nature, LRS distributions are unlikely to be solely shaped by unstructured environmental stochasticity. Instead, individuals seem to differ in their probability of surviving or reproducing (Kendall et al. 2011).

Often, this individual heterogeneity in LRS is assumed to originate from latent individual differences which are fixed throughout an individual's life, i.e. that there is individual heterogeneity in frailty, quality or fitness (e.g. Vaupel, Manton, and Stallard 1979; Morris 1998; Cam and Monnat 2000). This is commonly referred to as fixed heterogeneity. Genetic variation is one source of fixed heterogeneity (e.g. Keller and Waller 2002; Ellegren and Sheldon 2008), but epigenetic, maternal and permanent environmental effects may also be important (Wolf and Wade 2009; Turner 2009). This fixed variation is usually measured retrospectively; in some cases it may have arisen at fertilization, but it may also be shaped by the environment an individual experiences throughout its life, for instance through variation in habitat choice or through gene by environment interactions. It is important to distinguish fixed heterogeneity as it is used here—that is, the repeatability of individual performance—from other sources of variation that are not due to the properties of individuals (e.g. climatic variations among years). Indeed, only fixed differences among individuals can be the target of selection and allow for adaptation, provided that these fixed differences are passed on to the next generation—be it through genes (Keller and Waller 2002), philopatry (Schauber et al. 2007) or other processes (Bonduriansky 2012).

Recent publications (Tuljapurkar, Steiner, and Orzack 2009; Steiner, Tuljapurkar, and Orzack 2010; Orzack et al. 2011; Steiner and Tuljapurkar 2012) have argued forcefully that invoking fixed differences among individuals (i.e. fixed heterogeneity) in fitness components is rarely required to explain the observed heterogeneity in LRS. Instead, they emphasize that due to the stochasticity of individual life histories, individual heterogeneity is expected even in populations of identical individuals (Caswell 2011). Indeed, if individuals take a random trajectory through the various life-history stages, and if these stages are associated with differential reproductive and survival rates, the population-level distribution of LRS may be skewed and heterogeneous. This type of heterogeneity is referred to as dynamic heterogeneity (Tuljapurkar, Steiner, and Orzack 2009). Crucially, dynamic heterogeneity originates from differences among life stages, whereas fixed heterogeneity originates from variation in the properties of individuals.

Given that most life-history traits are heritable to some degree (Mousseau and Roff 1987; Postma 2014), it is beyond doubt that some fixed heterogeneity is present in most wild populations. At the same time, the cumulative effects of individual histories on their realized lifespan and reproductive success are also unquestionable (Caswell 2011). What is subject to discussion, however, is the relative importance of fixed, versus dynamic, heterogeneity in shaping variation in LRS. Steiner and Tuljapurkar 2012 suggested that, at least in small populations, the drift generated by large life-history stochasticity is too large for fixed heterogeneity to play a significant role in shaping evolution and demography at the level of a single population. Instead, they have proposed dynamic heterogeneity as the null model to explain any observed heterogeneity. Only if this null model can be rejected should we consider an additional role for fixed heterogeneity in shaping variation in LRS or fitness components.

Tuljapurkar, Steiner, and Orzack 2009 have suggested that an appropriate tool to test for fixed heterogeneity is provided by neutral simulations (NS hereafter), which generate summary statistics describing the distribution of LRS and the pattern of life-stage transitions expected in the absence of fixed heterogeneity. These expectations can subsequently be compared to their observed counterparts to detect departures from neutrality due to the existence of fixed heterogeneity.

The application of NS to data for two sea bird populations (Steiner, Tuljapurkar, and Orzack 2010; Orzack et al. 2011), as well as to a compilation of 22 vertebrate populations (Tuljapurkar, Steiner, and Orzack 2009) has been unable to reject the null hypothesis of neutrality, leading to the conclusion that dynamic heterogeneity alone can explain the observed variation in life histories in most populations. Indeed, we are aware of only one study in which NS rejected neutrality, for one of three reproductive parameters in a roe deer population (Plard et al. 2012).

In contrast to studies relying on NS, studies employing linear mixed models (hereafter MM) commonly report evidence for fixed heterogeneity (e.g. Cam and Monnat 2000; Royle 2008; Chambert et al. 2013; Guillemain et al. 2013; Chambert, Rotella, and Higgs 2014). Interestingly, Cam et al. 2013 have provided evidence for fixed heterogeneity in a data set for which the existence of fixed heterogeneity had been dismissed based on NS (Steiner, Tuljapurkar, and Orzack 2010). However, MM and NS differ in how they deal with data: MM rely on repeated measurements of individuals, while NS use summary statistics aggregated at the population level. Compared to MM, NS are thus less data-demanding, but might be less sensitive to statistical signals at the individual level. On the other hand, aggregation might allow NS to detect effects that emerge only at the population level and are invisible to MM. More formally, the discrepancy between NS and MM suggests that they differ in either their type I (i.e. false positive) error rate, or in their type II error rate (i.e. power). For instance, the opposite conclusions reached by NS in Steiner, Tuljapurkar, and Orzack 2010 and MM in Cam et al. 2013 may be the result of the statistical power of the NS being too low, preventing the detection of fixed heterogeneity (i.e. a type II error). Alternatively, MM may have high rates of type I error, if the individual-level variances estimated by the MM are spurious, or they are unduly interpreted as the mark of fixed heterogeneity.

Applying both methods to data with known properties allows for the estimation of both types of error rates and thereby provides insight into the ability of both methods to detect fixed heterogeneity. Unfortunately however, fixed heterogeneity is the result of latent, unobservable traits, which cannot be inferred without a modeling step (Cam et al. 2013), and it is precisely the performance of this modeling step that we investigate here. Computer simulations provide a way around this problem, as they allow one to apply methods to data sets with known underlying properties (e.g. Villemereuil, Gimenez, and Doligez 2013; Brooks, McCoy, and Bolker 2013).

Here, we simulate a series of longitudinal, individual-based, data sets through an algorithm that introduces varying amounts of fixed and dynamic heterogeneity in survival and reproduction. For illustrative purposes, these simulations are parametrized to match a population of snow voles (*Chionomys nivalis*, Martins 1842) located in the Swiss Alps. In order to assess the type I and type II error rates of both NS and MM, we subsequently analyze the simulated data sets using both methods. In a final step, we use these results to interpret the results of the application of both methods to the real snow vole data set. Figure 2.1 shows a diagram summarizing our approach. Altogether, our results highlight the lack of statistical power of NS, but at the same time emphasize that MM output should be interpreted with care. We discuss the origin of the discrepancy between NS and MM, and what this tells us about the nature of biological variability.

2.3 Material and methods

2.3.1 Data simulation

The simulation model matches the life cycle of the population of snow voles which we use in the empirical comparison of both methods. The monitoring of this population is discussed in some detail in Appendix 2.12. Only two age classes are modeled (non-reproducing juveniles and reproducing adults), and there are no sex-specific or spatio-temporal effects on fitness components, as the uncertainty with respect to the appropriate specification of these models would introduce an additional layer of complexity (see e.g. Cam et al. 2013). All simulated populations are monitored for 10 years. For every individual, we have perfect knowledge of survival and reproduction during the study period, but their fate beyond this period is unknown. Every year, a new cohort of 100 juveniles appears. After one year, these juveniles become adults and start reproducing. Every year, adults can reproduce once; the number of offspring produced by an individual is labeled annual reproductive success (ARS). In the real snow vole population, there is no apparent senescence in survival and the maximum age observed is four years old. Accordingly, in the simulations, adult survival probability does not vary with age until the fourth year, but all individuals still alive at that point die during the next winter. Mortality events occur after birth for juveniles and after reproduction for adults. A single sex is simulated, as the two sexes are generally analyzed separately in NS, and in MM sex differences in the mean are accounted for by fitting sex as a fixed factor.

We define a scenario as a collection of simulation parameters. For each scenario, 1000 data sets were simulated, that is 1000 putative populations with the same underlying properties. In an attempt to detect evidence for fixed heterogeneity, each data set was then analyzed using MM and NS. Note the potential for confusion between the simulation of the data sets on the one hand, and the neutral simulation method on the other. The latter is always referred to as NS. Simulations were carried out using a C++ program (available at `https://github.com/timotheenivalis/FixDynHet`), using the pseudo-random number generator `Mersenne Twister` (Matsumoto and Nishimura 1998) and a command file procedure following that of `IBDsim` (Leblois, Estoup, and Rousset 2009). The analyses of the simulation output were all conducted in `R 3.1.0` (R Core Team 2014), using the package `lme4` (version 1.1-7) (Bates et al. 2014).

Due to demographic stochasticity (sensu Fox and Kendall 2002), all simulated data sets contain a baseline level of dynamic heterogeneity. Indeed, according to Tuljapurkar, Steiner, and Orzack 2009, the presence of dynamic heterogeneity results in the "scaled sequence entropy of the transition matrix between reproductive stages" (hereafter simply referred to as entropy), being greater than zero, which is always the case here. Entropy measures the rate at which the diversity of life-history trajectories increases with their length, which is due to random transitions between stages with different survival probabilities and reproductive outcomes (Tuljapurkar, Steiner, and Orzack 2009).

Beyond this baseline level of dynamic heterogeneity, heterogeneity in fitness components is introduced either as explicit fixed heterogeneity, or through a Markovian process. For the simulation of fixed heterogeneity, at birth, each individual receives a fixed quality as reproducer and survivor. These fixed qualities do not change over the course of its life. Therefore, some individuals intrinsically have a high probability to perform well, and some individuals have a high probability to perform poorly, irrespective of their past performance, as in a classic frailty model (Vaupel, Manton, and Stallard 1979). In contrast, for the simulations using a Markovian process, an individual's probability to survive and to achieve a certain ARS is not fixed, but changes at each time step and depends solely on its ARS the time step before. Therefore, these data contain dynamic heterogeneity only. However, some of this mimics fixed heterogeneity because individual performances can persist over time. Generalized linear mixed models were used to check that the properties of the simulated data sets matched the model and the parameters used to generate them (see Appendix 2.8).

Simulations with explicit fixed heterogeneity At birth, every individual receives a quality as reproducer $q_{\rho,i}$, which is normally distributed with a mean of 0 and a variance equal to σ_ρ^2, i.e. $q_{\rho,i} \sim \mathcal{N}(0, \sigma_\rho^2)$. Individuals also receive a quality as survivor $q_{\phi,i}$, with $q_{\phi,i} \sim \mathcal{N}(0, \sigma_\phi^2)$. These qualities are fixed for the lifetime of an individual. Because trade-offs between survival and reproduction are not considered here, the two qualities are drawn independently for each individual. The variances σ_ρ^2 and σ_ϕ^2 represent the amount of fixed heterogeneity in reproduction and survival, respectively.

If individual i is an adult at time t, its annual reproductive success, $\rho_{i,t}$, is drawn from a Poisson distribution,

$$\rho_{i,t} \sim \mathcal{P}(\exp(\log(\mu_\rho) + q_{\rho,i})), \tag{2.1}$$

where μ_ρ is the mean annual reproductive success. For an individual with $q_{\rho,i} = 0$, i.e. the average individual in a population with fixed heterogeneity, the parameter of the Poisson distribution $(\exp(\log(\mu_\rho) + q_{\rho,i}))$ reduces to the population mean ARS (μ_ρ). The qualities for reproduction ($q_{\rho,.}$) are normally distributed on the log-transformed scale of ARS.

The survival outcome of an individual i at time t, $\phi_{i,t}$, is zero (death) if the individual is four years old, and otherwise is drawn from a Bernoulli distribution:

$$\phi_{i,t} \sim \mathcal{B}(\text{logit}^{-1}(\text{logit}(\mu_\phi + j_{i,t}\beta_j) + q_{\phi,i})), \tag{2.2}$$

where $\text{logit}(p) = \log(\frac{p}{1-p})$ and its inverse function $\text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$, where $j_{i,t}$ is a Boolean variable equal to 0 for adults and 1 for juveniles, and where β_j is the difference between the mean survival probability of juveniles and adults. For an individual with $q_{\phi,i} = 0$, the probability of survival $(\text{logit}^{-1}(\text{logit}(\mu_\phi + j_{i,t}\beta_j) + q_{\phi,i}))$ reduces to $(\mu_\phi + j_{i,t}\beta_j)$, the age-specific mean survival probability. The qualities for survival ($q_{\phi,.}$) are normally distributed on the logit-transformed scale.

The mean of a log (or a logit) distribution is in general not equal to the log (or the logit) of the mean of this distribution (i.e. $\overline{\log(x)} \neq \log(\bar{x})$). Hence, Gaussian variance in individual qualities introduces a bias on the log or logit scale in the mean realized ARS and survival. If not corrected for, this bias causes the distributions of ARS and survival to deviate from their neutral expectations, which could be interpreted as evidence for fixed heterogeneity. To this end, the median individual qualities, $\tilde{q_\rho}$ and $\tilde{q_\phi}$, were iteratively modified so that the realized population means do not depend on the variances in individual qualities.

Because they are fixed for life, the individual qualities are the target of selection. Indeed, selection, i.e. the individual-level covariance between quality and relative LRS, increases with increasing variances (σ_ρ^2 and σ_ϕ^2) (Appendix 2.10). It could thus be argued that in response to this selection, mean latent qualities should increase and their variances decrease over time. However, here we chose not to simulate a trans-generational response to selection, as this introduces an unnecessary layer of complexity: First, a phenotypic response to selection on components of fitness is not necessarily expected. For example, environmental deterioration, which may be the result of an increase in mean competitiveness (Fisher 1958; Hadfield, Wilson, and Kruuk 2011), may mask a genetic change. Second, only the additive genetic part of the variation can respond to selection, and genetic variation may be renewed through migration, mutations and balancing selection (Fisher 1958; Charlesworth 2015). Therefore, simulating a response to selection would require much more complicated simulations and many more assumptions (e.g. an explicit genetic architecture for fitness, mechanisms to maintain genetic variation, competitive interactions). Finally, both MM and NS are blind to temporal variation, as they compute statistics averaged over the whole data set, and even if a response to selection were apparent, it would have little effect on their performance.

The simulation framework outlined above closely matches the structure of the MM later used to analyze the simulated data. Although we believe this simulation framework to be closest to biological reality, it could be argued that this may result in an overestimation of the ability of MM to deal with real data. Therefore, two alternative simulation structures not exactly matching the structure of MM were used. In the first, fixed heterogeneity was introduced on the original, rather than transformed, scale of survival probability and expected reproductive success. The results from this first alternative simulation structure did not differ qualitatively from the results obtained with the standard simulation structure, so they are presented in Appendix 2.11. The second alternative structure considers identical individuals, that is there is no explicit fixed heterogeneity, and a Markovian process with structured transition probabilities between reproductive stages and survival probabilities (see below).

Simulations with a Markovian process Simulations were carried out as previously described, except that ARS and survival probabilities depended on their previous state and not on fixed individual qualities. This matches the structure of the NS as proposed by Tuljapurkar, Steiner, and Orzack 2009 and is referred to as the “full dynamic model” in Plard et al. 2012. Note that in this model, as shown in Plard et al. 2012, the non-random transition probabilities of the Markovian process can be interpreted either as the result of fixed heterogeneity (if successful animals have a higher than average probability of remaining successful because of their individual properties, such as genetic quality) or of dynamic heterogeneity (if the persistence of success comes from the properties of reproductive stages rather than individuals, e.g. only individuals that have a territory can reproduce and these individuals are more likely than non-reproducers to have a territory next year). Indeed, for short lived species, a Markovian process produces among-individual variance because there are only a few observations per individual, and the first outcome of a Markov chain can have a big influence on the mean individual outcome. In long-lived species, on the other hand, mean individual performances will asymptotically approach the population mean.

In these simulations, the ARS of individual i at time t, $\rho_{i,t}$, follows:

$$\begin{aligned}\rho_{i,t} &\sim \mathcal{P}(\mu_\rho);\text{ for second year individuals,}\\ \rho_{i,t} &\sim \mathcal{P}(\mu_\rho + m(\rho_{i,t-1} - \mu_\rho));\text{ for older individuals,}\end{aligned}$$

where $\rho_{i,t-1}$ is the ARS of the focal individual the year before, μ_ρ is the mean ARS of the population and m controls the strength of the Markovian process, i.e. the degree to which current reproductive success depends on the previous reproductive success. Only positive values of m were used in order to produce an individual persistence of ARS, which may mimic latent fitness (see below).
Similarly, the survival outcome of individual i at time t, $\phi_{i,t}$, follows:

$$\begin{aligned}\phi_{i,t} &\sim \mathcal{B}(\mu_\phi + \beta_j);\text{ for juveniles}\\ \phi_{i,t} &\sim \mathcal{B}(\text{logit}^{-1}(\text{logit}(\mu_\phi) + c(\rho_{i,t-1} - \mu_\rho)));\text{ for adults,}\end{aligned}$$

where μ_ϕ is the mean adult survival, β_j is the difference between the mean survival of juveniles and adults, and c controls the correlation between reproduction and survival. Survival probability at time t depends on ARS at time $t - 1$ rather than on previous survival, as the latter is always 1 for surviving individuals. Again, only positive values of c were used to simulate persistence of the individual propensity to survive. The positive correlation between successive survival probabilities arises indirectly through the positive correlation between successive ARS, combined with the positive correlation between ARS and survival.

In the presence of allocation trade-offs between different life-history traits, or between successive expressions of the same life-history trait, negative correlations (i.e. $m < 0$) and autocorrelations (i.e. $c < 0$) could be expected. However, phenotypic correlations between life-history traits are often positive (Stearns 1992, chapter 4). This discrepancy is the result of the variance in resource acquisition, which is related to variance in latent fitness, being larger than the variance in resource allocation (Noordwijk and Jong 1986). Based on this, positive values of c and m are in line with the presence of variation in latent fitness. Indeed, a positive correlation between survival and reproduction is observed in the snow vole data (correlation between observed variation in survival and reproduction: Pearson's correlation, 0.097, 95%CI $[-0.007; 0.198]$. For the correlation between the latent propensities to survive and to reproduce, see Appendix 2.14

Simulation parameters The simulated mean survival probability from year t to year $t + 1$ was 0.4 for juveniles and 0.2 for adults (observed means in snow voles: 0.403 and 0.219, respectively). ARS, averaged over adults, was set to 3, 10 or 50 offspring. For the real snow vole population, mean ARS values of 3 (resulting in a decreasing population size) and 10 (i.e. increasing population size) are within the range observed among years (noting that we include offspring of both sexes in ARS, while we analyze vital rates for only one sex), while the value 50 aimed at confirming the direction of the trend in test performance with respect to mean ARS. The variance in individual quality, either on the original scale or on a transformed scale, σ^2_ϕ and σ^2_ρ, took the values 0, 0.1, 0.5, 1 or 2. In simulations without fixed heterogeneity, the m parameters took the values 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 or 1, while the c parameters took the values 0, 0.5 or 1. We had no a priori expectations for the heterogeneity parameters (σ^2_ϕ, σ^2_ρ, m and c) in the real snow vole population and thus selected the non-null values in a range from small to large relative to the mean survival and ARS.

2.3.2 Testing for fixed individual heterogeneity

Neutral Simulations (NS) NS were carried out following Tuljapurkar, Steiner, and Orzack 2009, but we used the "full stochastic model" proposed by Plard et al. 2012. Compared to the original formulation of NS, the "full stochastic model" better isolates dynamic heterogeneity by making future states independent of the current state. Thereby it removes the non-stochastic component of transition probabilities and allows testing whether "a given lifetime reproductive metric distribution is generated only by dynamic heterogeneity" (Plard et al. 2012).

Briefly, individual life histories, starting as juveniles, are simulated by producing a sequence of ARS values, with the probability of each value of ARS corresponding to its frequency in the focal data set. Mortality events, with an age-specific probability estimated from the data set, are mapped to these individual trajectories. Subsequently, properties of the resulting LRS distribution, as well as of the transition matrix between life stages, are compared between the focal data set and that obtained using NS.

Here it is crucial to highlight some differences between the NS and the way in which we simulated the data sets to which they are applied. First and foremost, in NS the propensity to reproduce and to survive is identical for all individuals and never depends on the previous reproductive success. Second, in our simulations, ARS follows a Poisson distribution—all positive integers are possible values—whereas in NS, ARS are drawn from the ARS values observed in the focal data set, which can follow any distribution, and for instance may have gaps, multiple modes or extreme skewness. Third, in our simulations, mean survival probability is always 0.4 for juveniles and 0.2 for adults, while in NS these age-specific probabilities are the age-specific frequencies of survival that are realized in the focal data set. To sum up, our simulations are parametric and follow well defined distributions, while NS use empirical distributions and thereby stick to the data.

To test for a deviation from the neutral expectation, LRS distributions were compared using both Kolmogorov-Smirnov tests (used in Steiner, Tuljapurkar, and Orzack 2010) and χ^2 tests (used in Plard et al. 2012). Additionally, we calculated mean LRS, the variance in LRS, as well as the persistence of the reproductive stage transition matrix and its entropy following Plard et al. 2012. Observed values greater than the 95% quantile—or smaller than the 5% quantile in the case of entropy, because more fixed heterogeneity should decrease entropy (Tuljapurkar, Steiner, and Orzack 2009)—of the neutral distribution were considered significantly different. The proportion of data sets for which a test is significant in the absence of simulated fixed heterogeneity gives the type I error rate, whereas the proportion of data sets for which a given test is not significant in the presence of simulated fixed heterogeneity gives the type II error rate. The NS method is computationally intensive, so to minimize computational time, we used the minimal number of NS per simulated data set beyond which statistical power did not change (Appendix 2.9).

Mixed Models (MM) Generalized linear mixed models (GLMMs) were used to estimate the variance in reproduction and survival attributable to fixed individual heterogeneity, as well as to test for its statistical significance. Significance of the variance components was assessed using Likelihood Ratio Tests (LRT) (see e.g. Pinheiro and Bates 2000; Crainiceanu and Ruppert 2004), assuming that the statistic follows an even mixture of χ^2_1 and χ^2_0 (Self and Liang 1987). For survival, first a logistic model not allowing for individual-level heterogeneity was fitted:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} \,, \tag{2.3}$$

where μ_ϕ denotes the intercept and $\text{Age}_{i,t}$ denotes the effect of age (juvenile or adult) of individual i at time t. In order to model individual-level heterogeneity, this model was subsequently expanded with an individual random intercept:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} + z_{\phi,i} \;;\text{with } z_\phi \sim \mathcal{N}(0, \hat{\sigma_\phi}^2) \,. \tag{2.4}$$

Model (2.4) estimated the individual-level heterogeneity in survival probability, $\hat{\sigma_\phi}^2$. Moreover, a LRT comparing model (2.4) to model (2.3) tested for the significance of $\hat{\sigma_\phi}^2$.

Similarly, for ARS a first Poisson model without individual-level heterogeneity was fitted:

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} \,, \tag{2.5}$$

where μ_ρ denotes the intercept and $\text{Age}_{i,t}$ denotes the effect of age. Subsequently, an individual random intercept was included to model individual-level heterogeneity:

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} + z_{\rho,i} \;;\text{with } z_\rho \sim \mathcal{N}(0, \hat{\sigma_\rho}^2) \,. \tag{2.6}$$

Model (2.6) estimated the individual-level heterogeneity in reproductive ability, $\hat{\sigma_\rho}^2$. Moreover, a LRT comparing model (2.5) to model (2.6) tested for the significance of $\hat{\sigma_\rho}^2$.

In addition, for the analyses of data simulated by means of a Markovian process not including any explicit fixed heterogeneity, the models (2.5) and (2.6) were refitted while adding past reproductive success $\rho_{i,t-1}$ as a covariate. The estimated variance $\hat{\sigma_\rho}^2$ and the LRT comparing these two new models tests the significance of fixed heterogeneity while accounting for a Markovian process.

2.3.3 Analysis of the snow vole data set

A snow vole population, located in the Swiss Alps near Churwalden, at 2000m above sea level, has been monitored continuously since 2006. Analyses presented here are based on data collected until 2013. Individual recapture probability is virtually equal to 1.0, which facilitates the modeling of survival. For more information on the study site and data collection, see Appendix 2.12. NS were applied to the real snow vole data set exactly in the same way as to the simulated data sets, separately for males and females. For MM, starting from the models for ARS and survival used for the simulated data sets, we added sex and the sex by age interaction as additional fixed factors, as well as a random effect accounting for variation among years and an observation-level random effect. The latter accounts for overdispersion (see e.g. Atkins et al. 2013) and quantifies the overdispersion due to sources of heterogeneity not included in the model. In a second step, models also including ARS in the previous year were fitted in order to test for the presence of fixed heterogeneity after accounting for variation introduced by Markovian processes. Confidence intervals for all parameters were computed through 1000 parametric bootstraps, using the `confint` function in `lme4`. In a final step, the correlation between the propensity to survive and to reproduce was estimated using a bivariate GLMM in `MCMCglmm` (version 2.21) (Hadfield 2010). This model is detailed in Appendix 2.14.

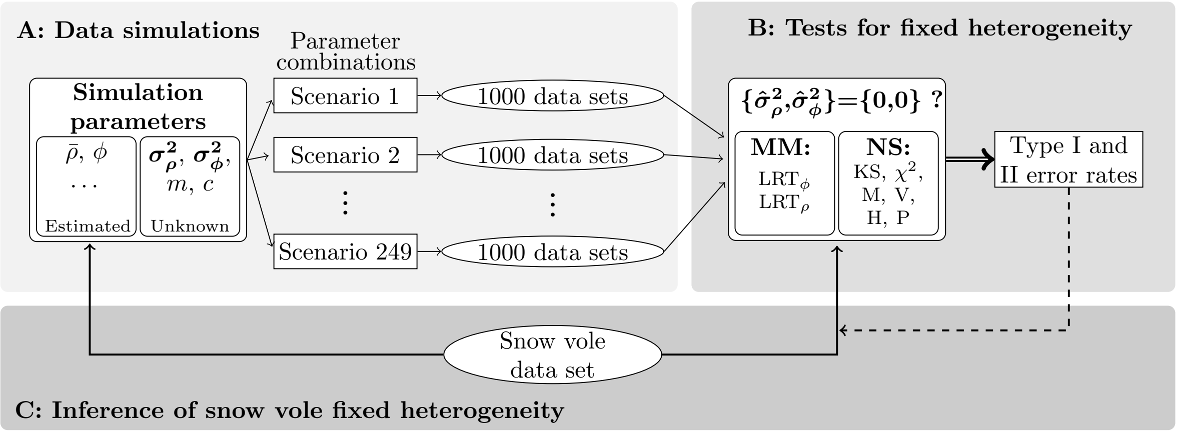

Figure 2.1: Illustration of the simulation and testing process. (A) Data simulation: The simulation model is parametrized using the life cycle and vital rates of a snow vole population, along with additional, unknown, parameters introducing fixed heterogeneity (σ^2_ϕ and σ^2_ρ) and dynamic heterogeneity (m and c). Different combinations of these simulation parameters define 249 scenarios. For each scenario, 1000 data sets are simulated. (B) Tests for fixed heterogeneity: Each simulated data set is tested for the presence of fixed heterogeneity with both mixed models (MM) using likelihood ratio tests (LRT) on survival (ϕ) and reproduction (ρ), and neutral simulations (NS), using six different tests (see main text). Because σ^2_ϕ and σ^2_ρ are known for each simulated data set, we can estimate the type I and type II error rates under each scenario. (C) Analysis of the snow vole data: Both MM and NS are applied to the real snow vole data set, and the outcome is interpreted in the light of the estimated error rates of each test.

2.4 Results

Mean ARS had no effect on the error rates of any test, so we merged together the the scenarios differing only by mean ARS. Therefore, all error rates are estimated based on 3000 tests (1000 data sets per scenario, times three mean ARS values).

2.4.1 Type I error rates

In the absence of simulated individual fixed heterogeneity and non-random transition probabilities between successive stages, all tests have a low rate of null-hypothesis rejection (table 2.1). This means that any discrepancy between NS and MM must come from a difference in type II rather than type I error rates.

Table 2.1: Type I error of tests used in the MM and NS approaches, when applied to data sets without underlying fixed heterogeneity and with fully random transition probabilities

	Mixed models		Neutral simulations					
	LRT_ρ	LRT_ϕ	KS	χ^2	H	P	M	V
estimate	0.042	0.000	0.000	0.021	0.018	0.039	0.000	0.000
95% CI	0.039;0.054	0;0.001	0;0.001	0.016;0.027	0.014;0.023	0.033;0.047	0;0.001	0;0.001

Note: Type I error rates are estimated as the proportion of simulated data sets, generated without fixed heterogeneity nor Markovian process, for which a test provides a p-value below 0.05. Hence, each proportion is estimated from 3,000 tests. The 95% CI (confidence intervals) are Wilson score intervals. LRT_ρ and LRT_ϕ refer to the Likelihood Ratio Tests of the variance associated with the individual random intercept in reproductive success and survival, respectively. KS refers to a Kolmogorov-Smirnov test, and χ^2 to a χ^2 test, both of which compare the Lifetime Reproductive Success (LRS) distribution in a focal data set to the distribution of LRS distributions obtained through neutral simulations (NS). The four other tests are based on the distribution of values obtained by NS compared to the value in the focal data set (mean (M) and variance (V) of the LRS distribution; and entropy (H) and persistence (P) of the transition matrix between successive annual reproductive successes.

2.4.2 Type II error rates

Simulations with explicit fixed heterogeneity

Neutral simulations (NS) The Kolmogorov-Smirnov test comparing LRS distributions is significant for only one simulated data set (pertaining to the scenario $\{\sigma^2_\rho = 1, \sigma^2_\phi = 2, \bar{\rho} = 50\}$) out of the 72,000 data sets with explicit fixed heterogeneity on the transformed scale. For the parameter range simulated, this test has thus effectively null power. Nevertheless, p-values decrease with increasing σ^2_ρ and σ^2_ϕ (for $\{\sigma^2_\rho = 0, \sigma^2_\phi = 0, \bar{\rho}.\}$ $\overline{p\text{-value}} = 0.998$, SE $= 0.001$; for $\{\sigma^2_\rho = 2, \sigma^2_\phi = 2, \bar{\rho}.\ \}$ $\overline{p\text{-value}} = 0.776$, SE $= 0.032$), showing that the extremely low power is not the result of a complete calculation failure. Similar to the results of Plard et al. 2012, the χ^2 test is more powerful than the Kolmogorov-Smirnov test. Nevertheless, statistical power remains below 0.8 for moderately sized simulated variances, and its maximal value is 0.89 for the highest simulated variances (figure 2.2(A)).

Tests based on mean LRS are non-significant for all data sets and every scenario. The power of tests based on the variance in LRS increases with increasing σ^2_ϕ, while the power peaks at intermediate values of simulated σ^2_ρ and decreases again for higher σ^2_ρ (figure 2.2(B)). The non-monotonic shape might be the result of the simultaneous increase in both the real observed-expected difference and the sampling variance: As the simulated variances go up, the LRS distribution becomes wider and flatter. Keeping the number of NS constant, this results in a less extensive sampling of the LRS distribution and a reduced power.

Tests based on the entropy of transition matrices display a pattern that is similar to that for χ^2 tests, albeit with lower statistical power, this time peaking at 0.57 (figure 2.2(C)). Tests based on the persistence of transition matrices have high statistical power (≈ 0.8) for $\sigma_\rho^2 \geq 1$, while increases in σ_ϕ^2 result only in a slight increase in statistical power (figure 2.2(D)). While they reach higher statistical power than the χ^2 tests, they have lower power than the χ^2 at intermediate σ_ρ^2 values.

Mixed models (MM) In contrast to NS, the power of the likelihood ratio test for ARS (LRT_ρ) is almost perfect for $\sigma_\rho^2 \geq 0.1$. Even though fixed heterogeneity in reproduction and survival are simulated independently, the power to detect fixed heterogeneity in reproduction is marginally influenced by the value of σ_ϕ^2 (figure 2.2(E) and, more clearly, Appendix 2.11 figure 2.112.6(E)). This is because a higher variance in latent survival probability increases the proportion of individuals that reach the maximal age, which provides more successive observations of reproduction and thereby increases the power to detect variance in reproductive quality. Overall, σ_ρ^2 is slightly underestimated ($\hat{\sigma_\rho^2} = 0.972\sigma_\rho^2$; adjusted R^2=0.9997).

The LRT_ϕ is never significant, even for $\sigma_\phi^2 = 2$. Moreover the estimation of σ_ϕ^2 is always close to zero (average of the median values 0.029) and does not increase with increasing σ_ϕ^2 (slope and SE: -0.0016 ± 0.0006). The failure of this model illustrates the intrinsic difficulty in estimating random effects for binary traits, especially when there are few repeated measurements per individual (e.g. Albert and Anderson 1984; Hosmer, Lemeshow, and Sturdivant 2013, chapter 9), as is the case in our short-lived simulated animals.

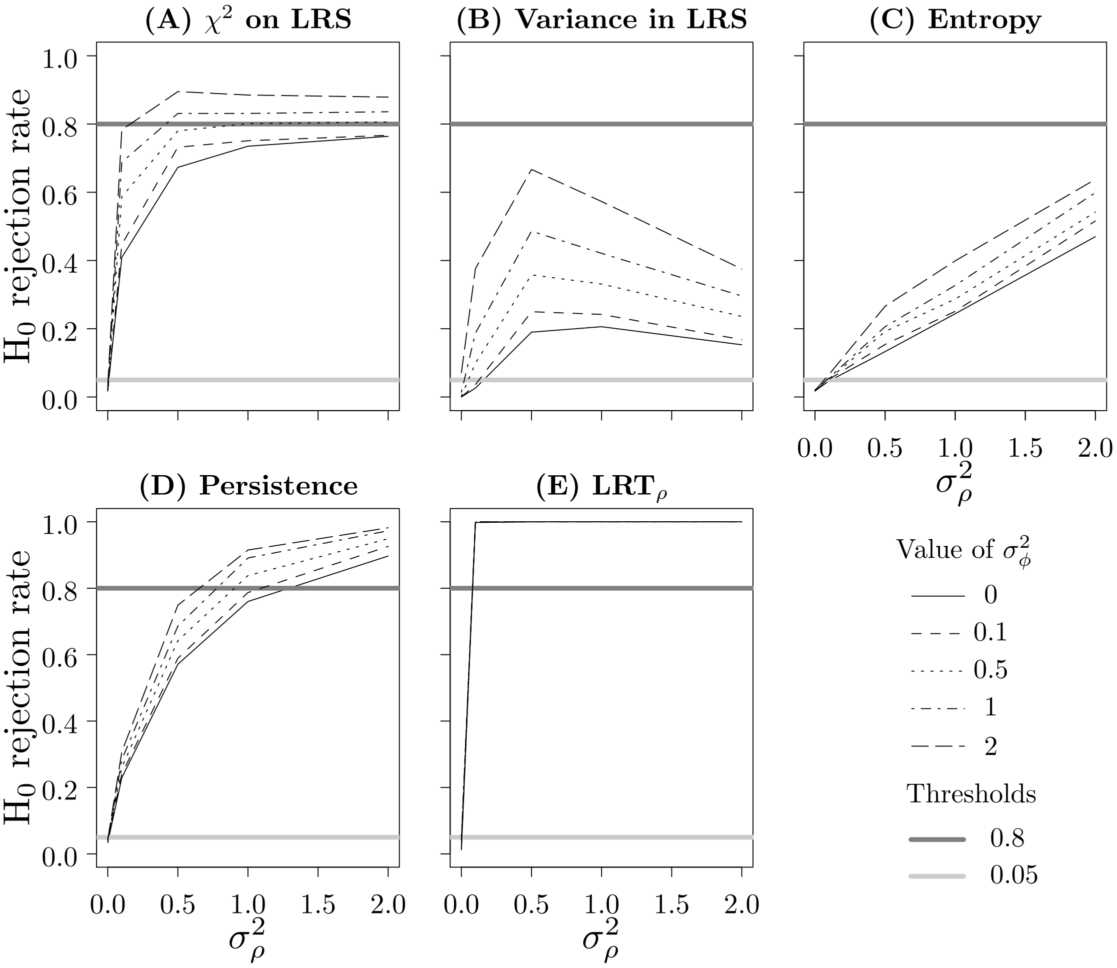

Figure 2.2: Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, as a function of the variance in reproductive propensity, σ^2_ρ, and survival propensity, σ^2_ϕ, when these variances are introduced on the transformed scales. The methods are: (A) a χ^2 test comparing the LRS distribution in a focal data set to the distribution of LRS distributions obtained through the neutral simulation approach (NS); tests based on proportion of values obtained by NS greater or equal to the value in the focal data set for (B) the variance in LRS, (C) the entropy of the transition matrix between successive annual reproductive success and (D) the persistence of this matrix; (E) a LRT for the significance of the individual random intercept in reproductive success. When $\sigma^2_\rho = \sigma^2_\phi = 0$, the null-hypothesis rejection rates are equal to the type I error rates, which is expected to be 0.05 (light gray line). When $\sigma^2_\rho \neq 0$ or $\sigma^2_\phi \neq 0$, the null-hypothesis rejection rates give (1-type II error rate), i.e. statistical power. The dark gray line indicates the 0.8 threshold. (A)-(D) are related to NS, (E) is related to MM.

2.4.3 Simulations with a Markovian process

Although data sets simulated using a Markovian process do not contain explicit fixed heterogeneity, both MM and NS reject the null hypothesis of an absence of fixed heterogeneity in most of the cases (figure 2.3).

The LRT$_\rho$, testing for fixed heterogeneity in ARS (based on MM), rejects the null hypothesis with a high probability, except for the lowest values of c and m (figure 2.3(E)). When $m > 0$, current ARS is influenced by past ARS, which in turn introduces variance in the propensity to reproduce. When $c > 0$, current survival probability is positively influenced by current ARS. As a consequence, successful reproducers live longer, resulting in more ARS values for these individuals, which improves the ability of the MM to detect individual-level variance. The LRT_{ϕ} is never significant for $c = 0$, but rejects the null hypothesis at a high rate for $c \geq 0.5$, and this increases as m increases (figure 2.3(G)). This pattern was expected as c controls the correlation between survival and reproduction, and indirectly makes the probability to survive in the current time step dependent on the probability to survive in the previous time step. Increasing values of m further strengthen this correlation.

Both the Kolmogorov-Smirnov test on the LRS distribution, and the test based on mean LRS, are non-significant for any data set with Markovian process. Furthermore, the χ^2 test rejects the null hypothesis with near certainty when $c > 0$, and, when $c = 0$, with probabilities going from low to moderate with increasing m (figure 2.3(A)). Given the absence of explicit fixed heterogeneity in these data, the χ^2 test can therefore be considered to have very high type I error rates (but see the discussion). The tests based on the variance in LRS, entropy and persistence follow a similar pattern of increasing probability of null-hypothesis rejection when m and c increase, but the test based on entropy does not reach a probability higher than 0.65, while the two other tests are close to 1 for the highest values of the parameters (figures 2.3(B)-(D)).

Based on these findings, it could be argued that both MM and Plard's version of NS (Plard et al. 2012) have a very high type I error rate when the transitions between stages are structured. We examine this interpretation in more detail in the discussion. However, the rejection rate of the LRT_{ρ} for fixed heterogeneity in ARS is drastically reduced by the inclusion of the past ARS ($\rho_{i,t-1}$) in the two mixed models that are being compared, i.e. with and without the individual random effect (compare figure 2.3(E) and figure 2.3(F)). The type I error rate is greater than the alpha threshold of 5% only when both $m > 0.8$ and $c > 0$ (figure 2.3(F)). Moreover, the estimates of the variance in reproductive propensity are reduced by the inclusion of $\rho_{i,t-1}$ in the models: over all the scenarios, the mean is $\hat{\sigma_{\rho}^2} = 0.004$, SE=0.002, with a maximal estimate of 0.144, whereas without including $\rho_{i,t-1}$, the mean is 0.050, SE=0.008, and the maximum 0.459. The former estimate is closer to zero, i.e. the individual-level variance that is explicitly simulated.

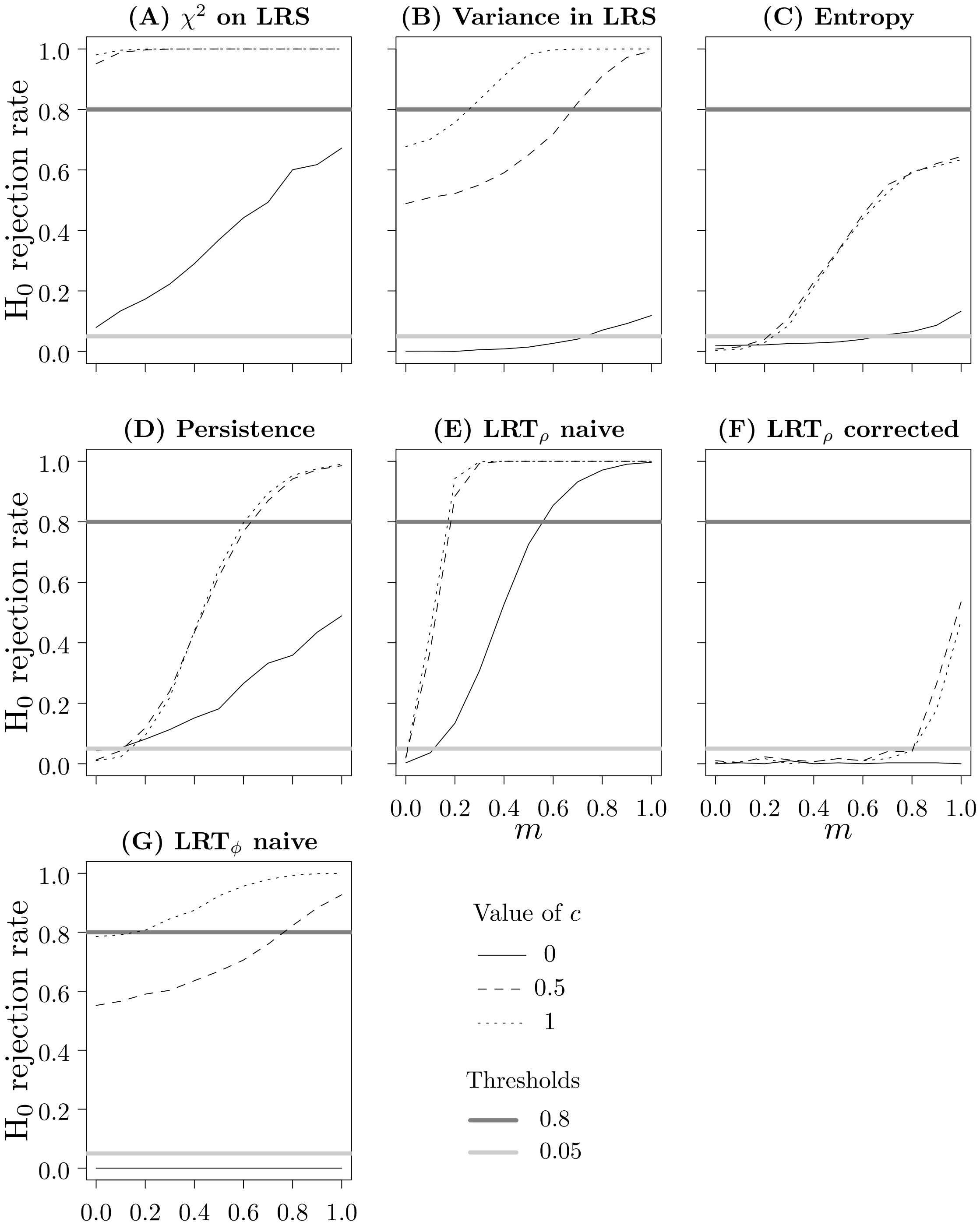

Figure 2.3: Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, when none is explicitly simulated, depending on the parameter *m*, controlling the structure of transitions between successive annual reproductive successes, and on the parameter *c*, controlling the dependency between survival probability and reproductive success (see the method section "Simulations of a Markovian process" for details). The methods are: (A) a χ^2 test comparing the LRS distribution in a focal data set to the distribution of LRS distributions obtained through NS; tests based on proportion of values obtained by NS greater or equal to the value in the focal data set for (B) the variance in LRS, (C) the entropy of the transition matrix between successive annual reproductive success and (D) the persistence of this matrix; (E) a LRT for the significance of the individual random intercept in reproductive success, using models that do not account for a Markovian process, or (F) that do account for a Markovian process; (G) a LRT for the significance of the individual random intercept in survival. For survival we did not try to account for the Markovian process. Assuming that the simulated Markovian process cannot be related to fixed heterogeneity, the null-hypothesis rejection rates represent type I error rates for all values of the *c* and the *m* parameters. (A)-(D) are related to the NS framework. (E)-(G) are related to the MM framework

2.4.4 Application to the snow vole data set

Neutral simulations (NS)

For males, none of the six tests carried out within the NS framework are significant. Neither the LRS distribution, nor the transition matrix between successive values of ARS, are distinguishable from those generated using NS (table 2.2). For females, out of the six tests, two are significant: there is more persistence and more variance than expected under neutrality; and the test on mean LRS is close to being significant. However, the tests on the complete LRS distribution (Kolmogorov-Smirnov and χ^2) are far from significant (table 2.2). The latter is unsurprising as a graphical examination of the observed and the simulated neutral LRS distribution shows that the two distributions are almost indistinguishable (figure 2.4). According to the authors of the NS framework, the comparison of LRS distributions, either through a Kolmogorov-Smirnov test (in Steiner and Tuljapurkar 2012) or a χ^2 test (in Plard et al. 2012), is the gold standard when testing for the presence of fixed heterogeneity with NS (Steiner 2013, pers. comm. November 25th). Based on these NS results, there is thus no evidence for fixed heterogeneity in either of the sexes, although the results are more equivocal in females.

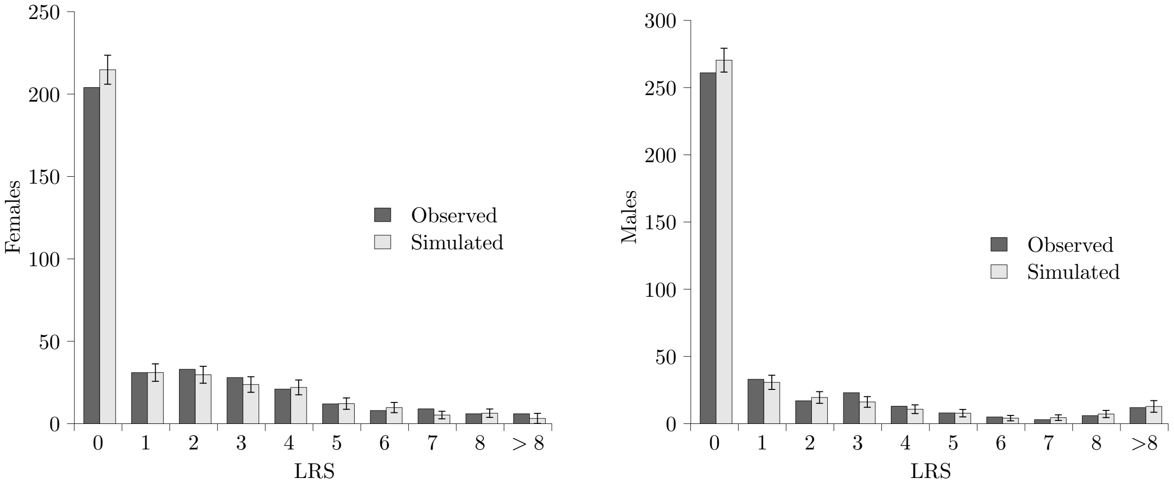

Figure 2.4: Distribution of lifetime reproductive success in the real snow vole data set, observed (dark bars) and simulated through 1000 neutral simulations (light bars with black error bars showing ± standard deviation), for 2.4.4 females and 2.4.4 males.WRONG

Mixed models (MM)

The GLMM for survival identifies significant between-years variance (5.622; 95% CI [1.133; 13.158]), but estimates a latent individual-level variance of 0 (95% CI [0; 0.248]) (see supplementary table 2.1 for all the estimates of this model).

The GLMM for ARS estimates variances among individuals (0.371; 95%CI [0.151; 0.475]) as well as among years (0.101; 95%CI [0.026; 0.452]) that are different from zero, and LRTs for both variances are highly significant. The random effect accounting for overdispersion does not significantly differ from zero, although its bootstrapped confidence interval includes positive values (table 2.2 for all the estimates of this model). When the individual random effect is not included, this overdispersion variance is highly significant, and the sum of squared Pearson residuals divided by the estimated residual degrees

Table 2.2: Outcomes of the various tests within the NS framework when applied to the real snow vole data set, for males and females separately

test	KS		χ^2			H	P	V	M
	D	p-value	χ^2	df	p-value	p-value	p-value	p-value	p-value
Males	0.025	0.969	8.33	15	0.909	0.629	0.646	0.395	0.378
Females	0.030	0.902	5.50	8	0.70	0.624	**0.035**	**0.031**	0.057

Note: KS refers to the Kolmogorov-Smirnov test, and χ^2 to the χ^2 test, comparing the Lifetime Reproductive Success (LRS) distribution in a focal data set to the distribution of LRS distributions obtained through NS. The four other tests are based on the proportion of values obtained by NS greater than the value in the focal data set for the mean (M) and variance (V) of the LRS distribution, and for the entropy (H) and persistence (P) of the transition matrix between successive annual reproductive success. The p-values $\leq$5% are shown in bold.

of freedom is approximately 2, while it falls to 1 with individual as a random effect. The estimation of residual degrees of freedom in GLMMs is a complex issue (Pinheiro and Bates 2000), but this approach seems to indicate that the overdispersion in the distribution is largely due to differences between individuals.

Excluding individuals reproducing for the first time, we fitted a GLMM that includes the previous reproductive success ARS_{t-1} and sex as fixed effects, and year as the only random effect. This model indicates a significant positive relationship between successive values of ARS (slope=0.0949; SE $= 0.0213$; p-value=8×10^{-06}). Nevertheless, adding individual as a random effect greatly improved the fit of the model ($\Delta\text{AIC} = 87$; LRT: $p\text{-value} < 10^{-16}$), providing evidence for the existence of significant individual-level variance ($\hat{\sigma}^2_{id} = 0.341$, bootstrapped 95% CI $[0.189; 0.453]$). Including ARS_{t-1} had little effect on the estimate of $\hat{\sigma}^2_{id}$ (see table 2.2), but now ARS_{t-1} no longer reached significance (slope=0.0210; SE $= 0.0275$; p-value=0.445).

Finally, the latent correlation between the propensities to survive and to reproduce was estimated as 0.32 (95% CI [-0.68;0.97]) and appears in the best model selected by DIC (see Appendix 2.14).

2.5 Discussion

2.5.1 Overview

Based on extensive simulations, we have shown that in the presence of fixed heterogeneity, NS have much less statistical power than MM, even when the model simulating the data does not match the structure assumed by the MM. In particular the Kolmogorov-Smirnov test, advocated in the earlier version of NS, has virtually no statistical power. In contrast, MM have low type I error rates and are not misled by the presence of dynamic heterogeneity, which in all data sets is non-zero if it is measured as entropy (Tuljapurkar, Steiner, and Orzack 2009). This finding directly contradicts the claim "[. . .] that random effect models will always detect unobservable fixed effects" Steiner, Tuljapurkar, and Orzack 2010. Second, in the absence of fixed heterogeneity, Markovian transitions between successive reproductive success and survival probabilities can induce high type I error rates, both in MM and NS sensu Plard et al. 2012. However, inclusion of previous reproductive success in the MM for reproduction substantially reduces these errors. Third, when applied to a real data set for a wild population of snow voles, NS only detect ambiguous deviations from neutrality and only for females. Moreover, the main tests of the framework, based on the total distribution of LRS, fail to reject the null hypothesis in both sexes. In striking contrast, MM show strong evidence for individual latent variance in reproductive success, even when a Markovian process is accounted for. In addition, MM give some indication of the presence of individual latent variance in survival, and of a positive correlation between survival and reproduction. However, the latter two parameters are estimated with substantial uncertainty.

2.5.2 Use of simulations

Testing methods on simulated data can be difficult because the specific simulation process used can differently match the assumptions and structures of the different methods. We tried to overcome this issue by using three different simulation models. Moreover, the rejection rates of MM and NS observed in our simulations are similar to those observed when the methods are applied to real data. Indeed, in the present work we applied both methods to a snow vole data set and found that the MM approach detected individual fixed heterogeneity, while the NS approach did not detect a significant deviation from the neutral expectation. This was also the case for the other data sets to which both methods were applied (MM by Cam et al. 2013; NS by Steiner, Tuljapurkar, and Orzack 2010). On the whole we are aware of only a single case in which NS led to the rejection of neutrality (Plard et al. 2012), whereas MM commonly find evidence for significant individual fixed heterogeneity, either by estimation of positive variance components, model selection (Cam et al. 2013) or posterior predictive checks (Chambert, Rotella, and Higgs 2014). Although there is some possibility of publication bias, this pattern is consistent with our power analysis.

2.5.3 Low power of Neutral Simulations

The low power of NS probably stems from the fact that they aggregate data on vital rates, and that they do so twice: first over the lifetime of individuals, and then they aggregate individuals into population-level statistics. Thereby they first discard the repeatability of individuals, which has been shown to blur heritable differences among individuals (Vaupel 1988). Second, population-level statistics can be produced by an infinite number of different mixtures of individual types (for instance, a mean probability of 0.5 can be the result of a population consisting only of individuals with a latent probability of 0.5, or from a uniform distribution of individual probabilities between 0 and 1). Therefore, some patterns of among-individual differences are indistinguishable at the population level. Individual-level data are naturally better at identifying the causes of variation at that level (Clutton-Brock and Sheldon 2010), and the ability to use non-aggregated data, for instance longitudinal information on marked individuals, further increases this power (Brooks, McCoy, and Bolker 2013). While a method such as Plard's NS could be valuable in the absence of such data, alternative methods making use of non-aggregated information, such as MM, should be preferred whenever possible.

Importantly, within a strict null-hypothesis testing framework, the failure to reject a null hypothesis cannot be interpreted as a proof of the null hypothesis. The absence of significance in most implementations of the NS (Steiner, Tuljapurkar, and Orzack 2010; Orzack et al. 2011; Tuljapurkar, Steiner, and Orzack 2009; Plard et al. 2012) is therefore not informative with respect to the presence and the biological significance of fixed heterogeneity. The null-hypothesis testing framework can partially be relaxed by an a priori power analysis. Although comparisons of simulated data sets with and without heterogeneity were indeed presented in Steiner and Tuljapurkar 2012, there fixed heterogeneity (assumed to be genetic) was modeled as two groups of homogeneous individuals, which except for clonal organisms is biologically unrealistic. In addition, the absence of significant differences between the data sets with and without fixed heterogeneity was not interpreted as a sign of a lack of statistical power, but as evidence that fixed heterogeneity has little effect on LRS distributions.

2.5.4 Effect of Markovian transitions

When no fixed heterogeneity was explicitly simulated, both MM and NS rejected the null hypothesis that fixed heterogeneity is absent. This was to be expected for MM, given that Markovian transitions mimic individual-level variance, and MM do not model population-level transition probabilities. It is more surprising that also NS had a high rate of false positives. However, we here used the "full random model" re-formulation of NS (Plard et al. 2012), and not the "full dynamic model" (Tuljapurkar, Steiner, and Orzack 2009). The latter simulates individual trajectories using a Markovian process, similar to the way data sets were simulated here, while the former simulates individual trajectories without taking into account the previous state. Hence, "full dynamic NS" would not reject the null hypothesis, and one could consider this in this case to be correct. However, as latent individual quality will necessarily produce a pattern that is consistent with a Markovian process, this formulation does not allow for a complete separation of fixed and dynamic heterogeneity (Plard et al. 2012). Observing a Markovian process is therefore in itself not informative with respect to the mechanisms shaping life histories. Hence, although they have a low type I error rate, "full dynamic NS" always have low statistical power.

We acknowledge that a Markovian process that is not due to fixed differences between individuals does mimic fixed heterogeneity, and thereby can bias estimates of between-individual variance based on full random NS and on MM. Therefore, a naive MM detects individual-level heterogeneity, irrespective of whether it is due to a population-level Markovian process or to individual-level differences. However, the type I error of MM can be substantially reduced by including previous reproductive success in the model (Rotella 2008; Cam et al. 2013). Although this is not a universal solution that accounts for all confounding factors, it highlights the flexibility of the MM framework, which allows for the incorporation of any factor that is perceived as potentially confounding based on knowledge of the study system.

2.5.5 Genetic variation as a source of fixed heterogeneity

In cases where the evidence for the presence of fixed heterogeneity is equivocal, for instance because the effects of Markovian processes and individual-level fixed differences are confounded, the use of genetic information and quantitative genetic methods has the potential to tease apart latent genetic quality from other sources of performance persistence, including stochastic transitions. Indeed, although other sources of variation may also generate fixed heterogeneity, the existence of significant additive genetic variation implies significant fixed heterogeneity, by definition determined at fertilization. Interestingly, estimates of additive genetic variation for fitness components are often large, even in small populations (for reviews see Mousseau and Roff 1987; Postma 2014). As a matter of fact, when standardized by the mean (i.e. evolvability) rather than the variance (i.e. heritability), fitness components appear to have higher additive genetic variation than other types of traits (Hansen, Pélabon, and Houle 2011; Postma 2014). In addition to our findings, this provides further support for fixed heterogeneity being more common than suggested by NS.

2.5.6 Interpretation of the snow vole results

Because they are similar in structure, our simulated data sets can shed light on the results from the analysis of the real snow vole data set. For example, it is unsurprising that the MM fails to detect individual heterogeneity in snow vole survival probabilities. The LRT_{ϕ} has no statistical power for simulated data sets with simulated $\sigma^2_{\phi} \leq 2$, while confidence and credibility intervals indicate that the possible values of σ^2_{ϕ} lay between 0 and 1 at most (supplementary tables 2.1 and 2.4). Unlike heterogeneity in individual survival probability, heterogeneity in individual reproductive success is easily detected and quantified by MM applied to simulated data sets (figure 2.2(E)). Accordingly, the analysis of the real data set identifies an individual variance in the propensity to reproduce that is significantly different from zero, and is estimated to be more than three time larger than the variance among years. Finally, given the estimate of the variance σ^2_{ρ}, we can get an estimate of the statistical power of the other tests to detect fixed heterogeneity in the real snow vole data set: a significant test seems possible for the χ^2 test (figure 2.2(A)), but quite unlikely for the test based on entropy (figure 2.2(C)).

A positive correlation between individual-level variation in reproduction and survival would provide further support for fixed heterogeneity. However, as mentioned above, the estimation of individual-level variance in survival is difficult because this is a binary trait, and because due to their short lifespan there are few observations per individual. Hence there is a lot of uncertainty in the estimation of this correlation parameter. Nevertheless, the most likely values are positive (Appendix 2.14).

2.5.7 Fixed heterogeneity and the concept of fitness

The debate surrounding the biological significance of fixed heterogeneity appears to stem at least partly from different concepts of fitness. On the one hand, proponents of the neutral theory of life histories consider fitness to be a property of a category of individuals, and consider variation in reproductive success among individuals to be mostly due to dynamic heterogeneity, rather than due to variation in latent individual properties (Steiner and Tuljapurkar 2012). On the other hand, researchers in the field of evolutionary ecology often see fitness as a latent property of individuals (Cam and Monnat 2000), that is, an expected value defined at the individual level that cannot be measured directly (Brandon and Beatty 1984; Price 1996; Krimbas 2004). As the mean value of a group is also the expected value of an individual belonging to this group, the two views are not fundamentally different. In sexual organisms however, each individual is unique, which makes it difficult to assign it to a hypothetical group made of identical individuals. If stochastic variation underlies most of the realized reproductive success and there are no fitness differences between individuals, as adherents of the neutral theory of life histories advocate, then it is useless to define fitness at the individual level. However, if there exists significant fixed heterogeneity, individual performances carry some information about their latent properties, for example due to their genetic makeup. In the presence of fixed heterogeneity it therefore seems useful to use an individual-level definition of fitness, differing from both group-level fitness and realized reproductive success.

2.6 Conclusions

Using extensive simulations, we have demonstrated that NS are uninformative with respect to the biological significance of fixed heterogeneity. Based on the work of Plard et al. 2012 and our power analysis, we conclude that the observation of a Markovian process in stage-transition probabilities does in itself not provide any biological insights. Within the NS framework, the full random model (Plard et al. 2012) should be preferred over the full dynamic model (Tuljapurkar, Steiner, and Orzack 2009), and the χ^2 test should be preferred over the Kolmogorov-Smirnov test. In addition, any use of NS should be complemented by an a priori power analysis, or otherwise be restricted to a strict null-hypothesis testing framework, where failure to reject the null hypothesis does not allow any conclusions regarding the null hypothesis being true, and/or the alternative hypothesis false. However, even when these improvements are included in the NS framework, we recommend that its use is restricted to data sets where individuals are not identified.

Instead, we show that MM are more powerful, but not more susceptible to type I error. Although MM can be mislead by confounding factors, given a good knowledge of the biological system, it is possible to account for these confounding factors, in which case MM have a very low type I error rate.

Finally, the confrontation of our power analysis with the analysis of the real snow vole data set supports the presence of fixed heterogeneity in fitness components in this population. Further research is being carried out to identify what traits can be related to this latent heterogeneity, and how genetic and maternal effects shape these differences.

On the whole, this work supports the idea that fixed heterogeneity is more common than suggested by the studies based on NS.

2.7 Acknowledgments

We thank Arpat Ozgul, Yannis Michalakis, Benjamin M. Bolker, Gordon A. Fox and an anonymous reviewer for helpful comments on this manuscript. Thanks to Ulrich K. Steiner for discussions on an earlier version of this work and for help with the implementation of neutral simulations. Thanks also to Lukas Keller, Josh Van Buskirk and Pirmin Nietlisbach for inspiring discussions on this topic. The snow vole monitoring was authorized by the *Amt für Lebensmittelsicherheit und Tiergesundheit*, Chur, Switzerland. Funding was provided by the Swiss National Science Foundation project grant (`31003A_141110`).

References

Albert, A, and J.A. Anderson. 1984. "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika* 71 (1): 1–10. *http://biomet.oxfordjournals.org/content/71/1/1.abstract*.

Atkins, David C, Scott A Baldwin, Cheng Zheng, Robert J Gallop, and Clayton Neighbors. 2013. "A tutorial on count regression and zero-altered count models for longitudinal substance use data." *Psychology of Addictive Behaviors* 27, no. 1 (March): 166–77.

Barnett, Adrian G., Nicola Koper, Annette J. Dobson, Fiona Schmiegelow, and Micheline Manseau. 2010. "Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology." *Methods in Ecology and Evolution* 1, no. 1 (March): 15–24.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2014. *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-7.

Bonduriansky, R. 2012. "Rethinking heredity, again." *Trends in Ecology & Evolution* 27, no. 6 (June): 330–6. ISSN: 0169-5347. *http://www.ncbi.nlm.nih.gov/pubmed/22445060*.

Brandon, Robert, and John Beatty. 1984. “The propensity interpretation of ′ fitness ′. No interpretation is no substitute.” *Philosophy of Science* 51 (2): 342–347.

Brooks, Mollie E, Michael W McCoy, and Benjamin M Bolker. 2013. “A method for detecting positive growth autocorrelation without marking individuals.” *PloS One* 8, no. 10 (January): e76389.

Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn.* Springer-Verlag, New York.

Cam, Emmanuelle, Olivier Gimenez, Russell Alpizar-Jara, Lise M. Aubry, Matthieu Authier, Evan G. Cooch, David N. Koons, et al. 2013. “Looking for a needle in a haystack: inference about individual fitness components in a heterogeneous population.” *Oikos* 122, no. 5 (May): 739–753.

Cam, Emmanuelle, William a Link, Evan G Cooch, Jean-Yves Monnat, and Etienne Danchin. 2002. “Individual covariation in life-history traits: seeing the trees despite the forest.” *The American Naturalist* 159, no. 1 (January): 96–105.

Cam, Emmanuelle, and Jean Yves Monnat. 2000. “Stratification based on reproductive state reveals contrasting patterns of age-related variation in demographic parameters in the kittiwake.” *Oikos* 90, no. 3 (September): 560–574.

Caswell, Hal. 2011. “Beyond R_0: Demographic models for variability of lifetime reproductive output.” *PLoS One* 6, no. 6 (June): e20809.

Chambert, Thierry, Jay J Rotella, and Megan D Higgs. 2014. “Use of posterior predictive checks as an inferential tool for investigating individual heterogeneity in animal population vital rates.” *Ecology and Evolution* 4, no. 8 (April): 1389–97.

Chambert, Thierry, Jay J Rotella, Megan D Higgs, and Robert A Garrott. 2013. “Individual heterogeneity in reproductive rates and cost of reproduction in a long-lived vertebrate.” *Ecology and Evolution* 3, no. 7 (July): 2047–60.

Charlesworth, Brian. 2015. “Causes of natural variation in fitness : Evidence from studies of Drosophila populations.” *Proceedings of the National Academy of Sciences* 112 (6): 1662–1669.

Clutton-Brock, Tim H. 1988. “Reproductive success in male and female red deer.” In *Reproductive success. Studies of individual variation in contrasting breeding systems.* Edited by Tim H Clutton-Brock, 325–343. University of Chicago Press, Chicago / London.

Clutton-Brock, Tim H, and Ben C Sheldon. 2010. “Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology.” *Trends in Ecology & Evolution* 25 (10): 562–573.

Crainiceanu, Ciprian M., and David Ruppert. 2004. “Likelihood ratio tests in linear mixed models with one variance component.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, no. 1 (February): 165–185.

Ellegren, Hans, and Ben C Sheldon. 2008. “Genetic basis of fitness differences in natural populations.” *Nature* 452:169–175.

Fisher, R. 1958. *The genetical theory of natural selection.* Second. Dover Publications, New York.

Fox, Gordon A, and Bruce E Kendall. 2002. “Demographic stochasiticy and the variance reduction effect.” *Ecology* 83 (7): 1928–1934. ISSN: 0012-9658.

Guillemain, Matthieu, Andy J. Green, Géraldine Simon, and Michel Gauthier-Clerc. 2013. "Individual quality persists between years: individuals retain body condition from one winter to the next in Teal." *Journal of Ornithology* 154, no. 4 (May): 1007–1018.

Hadfield, Jarrod D. 2010. "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33 (2): 1–22.

Hadfield, Jarrod D, Alastair J Wilson, and Loeske E B Kruuk. 2011. "Cryptic evolution: does environmental deterioration have a genetic basis?" *Genetics* 187, no. 4 (April): 1099–113.

Hadfield, J.D., D.S. Richardson, and T. Burke. 2006. "Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework." *Molecular Ecology* 15, no. 12 (January): 3684.

Hansen, Thomas F., Christophe Pélabon, and David Houle. 2011. "Heritability is not evolvability." *Evolutionary Biology* 38:258–277.

Haring, E, B Herzig-Straschil, and F Spitzenberger. 2000. "Phylogenetic analysis of Alpine voles of the *Microtus multiplex* complex using the mitochondrial control region." *Journal of Zoological Systematics and Evolutionary Research* 38:231–238.

Hosmer, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression, third edition.* Hoboken, New Jersey: Wiley. ISBN: 0471356328.

Janeau, G, and S Aulagnier. 1997. "Snow vole - *Chionomys nivalis* (Martins 1842)." *IBEX Journal of Mountain Ecology* 4:1–11.

Keller, LF, and DM Waller. 2002. "Inbreeding effects in wild populations." *Trends in Ecology & Evolution* 17 (5): 19–23.

Kendall, Bruce E., Gordon A. Fox, Masami Fujiwara, and Theresa M. Nogeire. 2011. "Demographic heterogeneity, cohort selection, and population growth." *Ecology* 92 (10): 1985–1993.

Krimbas, Costas B. 2004. "On fitness." *Biology & Philosophy* 19, no. 2 (March): 185–203.

Leblois, R., A. Estoup, and F. Rousset. 2009. "IBD Sim: A computer program to simulate genotypic data under Isolation by Distance." *Molecular Ecology Resources* 9:107–109.

Marshall, TC, J Slate, Loeske E.B. Kruuk, and Josephine M Pemberton. 1998. "Statistical confidence for likelihood-based paternity inference in natural populations." *Molecular Ecology* 7:639–655.

Matsumoto, M., and T. Nishimura. 1998. "Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator." *ACM Transactions on Modeling and Computer Simulation* 8 (1): 3–30.

Morris, DW. 1998. "State-dependent optimization of litter size." *Oikos* 83:518–528.

Mousseau, Timothy A, and Derek A Roff. 1987. "Natural selection and the heritability of fitness components." *Heredity* 59, no. 2 (October): 181–197.

Noordwijk, Arie J van, and G. de Jong. 1986. "Acquisition and allocation of resources: Their influence on variation in life history tactics." *The American Naturalist* 128 (1): 137–142.

Orzack, Steven Hecht, Ulrich K. Steiner, Shripad Tuljapurkar, and Paul Thompson. 2011. "Static and dynamic expression of life history traits in the northern fulmar *Fulmarus glacialis*." *Oikos* 120, no. 3 (March): 369–380.

Pinheiro, Jose C., and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS.* Statistics and Computing Series. Springer-Verlag, New York, NY.

Plard, F., C. Bonenfant, D. Delorme, and J.M. Gaillard. 2012. “Modeling reproductive trajectories of roe deer females: Fixed or dynamic heterogeneity?” *Theoretical Population Biology* 82, no. 4 (December): 317–328.

Postma, Erik. 2014. “Four decades of estimating heritabilities in wild vertebrate populations: improved methods, more data, better estimates?” In *Quantitative genetics in the wild,* edited by Anne Charmentier, Dany Garant, and Loeske E. B. Kruuk. Oxford: Oxford University Press.

Price, P. W. 1996. *Biological evolution.* Saunders College Publishing, Philadelphia, PA.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Robertson, A. 1966. “A mathematical model of the culling process in dairy cattle.” *Animal Production* 8:95–108.

Rotella, Jay J. 2008. “Estimating reproductive costs with multi-state mark-recapture models, multiple observable states, and temporary emigration.” In *Modeling Demographic Processes In Marked Populations. Series Environmental and Ecological Statistics, Vol 3.* Edited by David L Thomson, Evan G Cooch, and Michael J Conroy, 157–172. Springer.

Royle, J Andrew. 2008. “Modeling individual effects in the Cormack-Jolly-Seber model: a state-space formulation.” *Biometrics* 64 (2): 364–70.

Schauber, Eric M, Brett J Goodwin, Clive G Jones, and Richard S Ostfeld. 2007. “Spatial selection and inheritance: applying evolutionary concepts to population dynamics in heterogeneous space.” *Ecology* 88 (5): 1112–1118.

Self, Steven G., and K Y Liang. 1987. “Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions.” *Journal of the American Statistical Association* 82 (398): 605–610.

Stearns, Stephen C. 1992. *The evolution of life histories.* Oxford University Press, Oxford.

Steiner, Ulrich K, Shripad Tuljapurkar, and Steven Hecht Orzack. 2010. “Dynamic heterogeneity and life history variability in the kittiwake.” *Journal of Animal Ecology* 79, no. 2 (March): 436–44.

Steiner, Ulrich Karl, and Shripad Tuljapurkar. 2012. “Neutral theory for life histories and individual variability in fitness components.” *Proceedings of the National Academy of Sciences* 109 (12): 4684–4689.

Tuljapurkar, Shripad, Ulrich K Steiner, and Steven Hecht Orzack. 2009. “Dynamic heterogeneity in life histories.” *Ecology Letters* 12, no. 1 (January): 93–106.

Turner, Bryan M. 2009. “Epigenetic responses to environmental change and their evolutionary implications.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364, no. 1534 (November): 3403–18.

Vaupel, J W. 1988. “Inherited frailty and longevity.” *Demography* 25 (2): 277–287.

Vaupel, JW, KG Manton, and E Stallard. 1979. “The impact of heterogeneity in individual frailty on the dynamics of mortality.” *Demography* 16 (3): 439–454.

Villemereuil, Pierre de, Olivier Gimenez, and Blandine Doligez. 2013. "Comparing parent-offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits." *Methods in Ecology and Evolution* 4, no. 3 (March): 260–275.

Wandeler, Peter, and Glauco Camenisch. 2011. "Identifying Y-chromosomal diversity by long-template PCR." *Molecular Ecology Resources* 11:835–841.

Wandeler, Peter, R Ravaioli, and T. B. Bucher. 2008. "Microsatellite DNA markers for the snow vole (*Chionomys nivalis*)." *Molecular Ecology Resources* 8:637–639.

Wilberg, Michael J., and James R. Bence. 2008. "Performance of deviance information criterion model selection in statistical catch-at-age analysis." *Fisheries Research* 93, nos. 1-2 (September): 212–221.

Wolf, Jason B, and Michael J Wade. 2009. "What are maternal effects (and what are they not)?" *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364, no. 1520 (April): 1107–15.

Supplementary information

2.8 Checking the properties of the data sets

The following Generalized Linear Models were fitted to the simulated data sets in order to test whether the data set properties matched the parameters used to generate them:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} \text{ ; using a binomial error structure} \tag{2.7a}$$

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \rho_{i,t} \text{ ; using a binomial error structure} \tag{2.7b}$$

$$\log(\rho_{i,t}) = \mu_\rho + \rho_{i,t-1} \text{ ; using a quasi-Poisson error structure} \tag{2.7c}$$

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} \text{ ; using a quasi-Poisson error structure} \tag{2.7d}$$

These were used to check that survival depended on age (2.7a), that survival depended on annual reproductive success only when that was required (2.7b), that ARS depended on previous reproductive attempts only when fixed heterogeneity for reproductive success or Markovian reproduction was simulated (2.7c) and that ARS of adults was not age-dependent (2.7d). The simulated data had all the expected properties. Furthermore, we never found a significant association between reproduction and survival. This goes against the claim made in Steiner, Tuljapurkar, and Orzack 2010 that dynamic heterogeneity alone can generate a positive association between reproduction and survival.

Instead, we argue here that the findings on which they base their claim reflects their use of reproductive stage-specific survival in their NS, and reproduction and survival being positively correlated in the source data (Cam et al. 2002). Hence, it is not the random transitions themselves that are responsible for the positive association, but the positively associated stage-specific probabilities of survival and reproduction. The origin of the latter remains unexplained, but is consistent with variation in latent fitness among individuals.

2.9 Optimal number of neutral simulations per data set.

The neutral simulation approach (NS) is computationally intensive: as the focal population consists of 10 cohorts of 100 individuals, performing 1000 neutral simulations (i.e. simulating 1000 hypothetical populations), requires 1,000,000 individual trajectories to be simulated for every simulated data set (and 75,000,000,000 individual trajectories for the complete study). To minimize computational time, we determined the number of neutral simulations per simulated data set beyond which statistical power did not change. Out of the six tests mentioned above, only χ^2 tests on LRS distributions are sensitive to the number of neutral simulations; while χ^2 tests based on 1000 neutral simulations differ from those based on 100 neutral simulations ($\Delta\text{power}_{1000-100}$=-0.067, se=0.033), the tests based on 100,000 neutral simulations do not have more statistical power than those based on 1000 neutral simulations ($\Delta\text{power}_{100,000-1000}$=-0.031, se=0.033), and the correlation of the statistical power across scenarios is high ($R^2 = 0.92$). Accordingly, each simulated data set was analyzed using 1000 neutral simulations. Note that the fact that in this case statistical power plateaus already above 1000 neutral simulations is the result of the relatively short lifespan of the simulated animals, which allows for a quick exploration of all the possible individual trajectories.

2.10 Selection for latent quality

As outlined in the main document, we simulated fixed heterogeneity by attributing to each individual i a fixed quality for annual reproductive success ($q_{\rho,i}$) and a fixed quality as survivor ($q_{\phi,i}$). These two kind of individual qualities are normally distributed, with mean zero and variance σ^2_ρ and σ^2_ϕ, respectively. The selection acting on, or due to, this variation in latent individual qualities for reproduction and for survival was measured as the individual-level covariance between the qualities and a proxy for fitness (ω): relative lifetime reproductive success (Robertson 1966).

The selection coefficients increase with increasing variance in individual latent qualities, both for reproduction (figure 2.102.10) and for survival (figure 2.102.10). This confirms that the heterogeneity simulated is non-neutral.

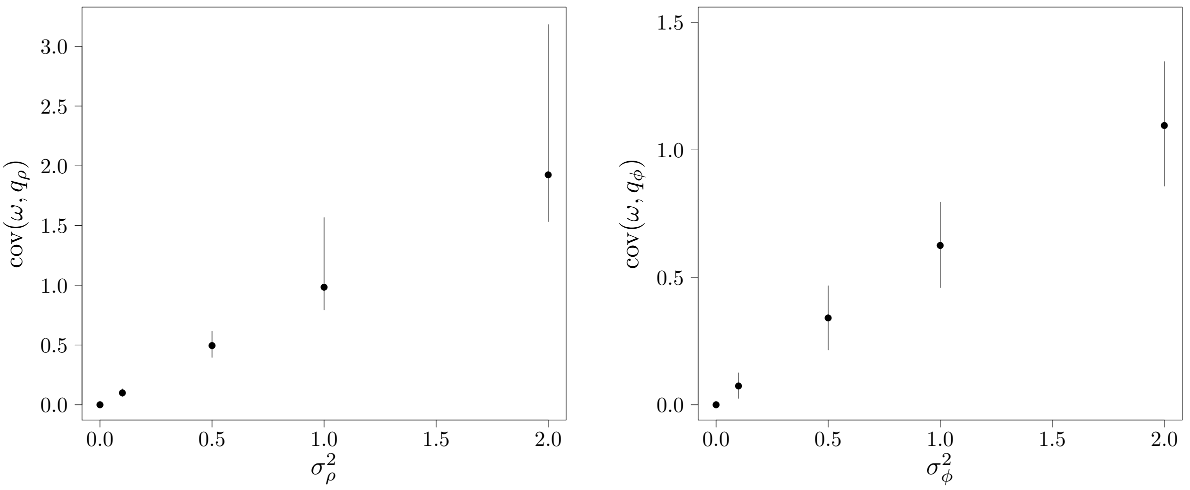

Figure 2.5: Appendix C Strength of selection on individual fixed qualities for survival and reproduction, as a function of the expected variance in these qualities. Strength of selection was measured as the individual-level covariance between the qualities and a proxy for fitness (ω): relative lifetime reproductive success; for reproduction quality and for survival quality. Vertical bars show the 95% interval of the estimate distributions.

2.11 Simulating fixed heterogeneity on the original scale

It could be argued that the superior statistical power of the LRT$_\rho$ is the result of the simulation process used to introduce fixed heterogeneity has the same structure as the MM estimating it. To address this, additional simulations were performed in which individual reproductive success and survival probability depended on their qualities on the original scale rather than on a transformed scale. Otherwise simulations were similar to those where fixed heterogeneity was introduced on the transformed scale. To this end, the reproductive success and survival of an individual i, at time t, are drawn from

$$\rho_{i,t} \sim \mathcal{P}\left(\mu_\rho + q_{\rho,i}\right) \tag{2.8a}$$

$$\text{and } \phi_{i,t} \sim \mathcal{B}\left(\mu_\phi + \beta_{age} + q_{\phi,i}\right). \tag{2.8b}$$

Although when the variance in quality for reproduction is included on the original, non-transformed, scale, mean reproductive success ($\overline{\text{ARS}}$) has a dramatic negative influence on the power of the different tests, the hierarchy in the performance of the different tests does not change across the values of mean reproductive success. Therefore, we chose to present the results with pooled $\overline{\text{ARS}}$ only (figure 2.6) Furthermore, it should be noted that although the σ_ρ^2 parameter values are the same in this section as in the previous one (0,0.1,0.5,1 and 2), they correspond to much smaller realized variances, as the variance is introduced on the original scale and not on a log-scale as previously. For correspondence between the variances on the two scales, see table 2.3.

Table 2.3: Realized variance on the log scale as a function of variance introduced on the original scale (σ_ρ^2) and mean reproductive success ($\overline{\rho}$)

$\overline{\text{ARS}}$	σ_ρ^2 on original scale				
	0	0.1	0.5	1	2
3	0	0.01143	0.06649	0.16947	0.39091
10	0	0.00100	0.00506	0.01027	0.02108
50	0	0.00004	0.00020	0.00040	0.00079

Note: Each realized variance was estimated from the variance of the log of 1,000,000 draws from a normal distribution of mean $\overline{\rho}$ and variance σ_ρ^2.

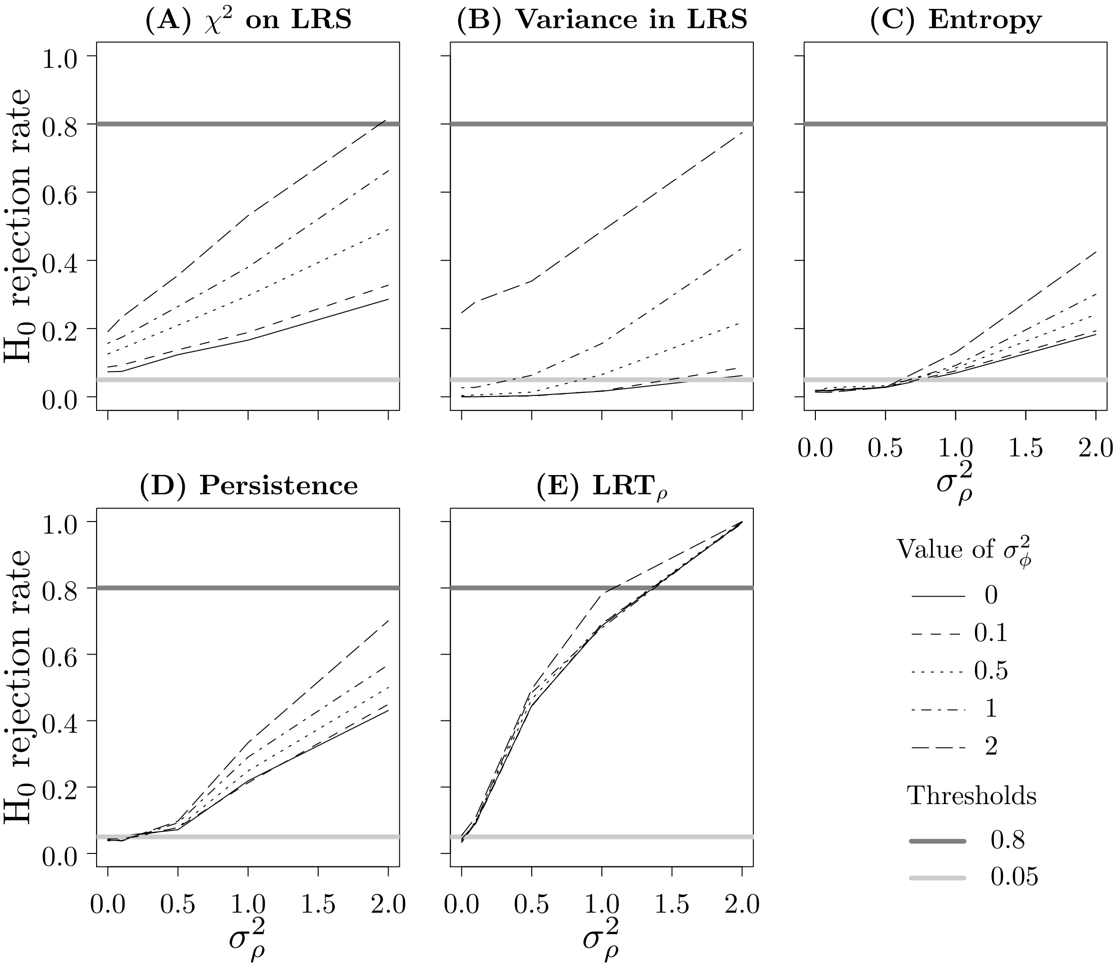

Figure 2.6: Appendix D Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, depending on the variance in reproductive propensity, σ^2_ρ, and on the variance in survival propensity, σ^2_ϕ, when these variances are introduced on the original scales. The methods are: (A) a χ^2 test comparing the Lifetime Reproductive Success (LRS) distribution in a focal data set to the distribution of LRS distributions obtained through the neutral simulation approach (NS); tests based on proportion of values obtained by NS greater or equal to the value in the focal data set for (B) the variance in LRS, (C) the entropy of the transition matrix between successive annual reproductive success and (D) the persistence of this matrix; (E) a Likelihood Ratio Test for the significance of the individual random intercept in reproductive success. When $\sigma^2_\rho = \sigma^2_\phi = 0$, the null-hypothesis rejection rates are equal to the type I error rates, which is expected to be 0.05 (light gray line). When $\sigma^2_\rho \neq 0$ or $\sigma^2_\phi \neq 0$, the null-hypothesis rejection rates give (1-type II error rate), i.e. statistical power, which should be above 0.8 (dark gray line). (A)-(D) are related to NS, (E) is related to MM.

2.12 The snow vole population

A snow vole population, located in the central eastern Alps near Churwalden, Switzerland (46°48′ N, 9°34′ E) at 2000m above sea level, has been monitored continuously since 2006. Analyses presented here are based on data collected until 2013. The study site consists of scree, which is the favourite habitat of the species, interspersed by patches of alpine meadows and surrounded by forest and larger meadows, which are not suitable habitats (Janeau and Aulagnier 1997). Four trapping nights are necessary for sampling the complete area. Trapping throughout the whole study area took place two (in one year), three (in three years) or five times (in four years), between late May and mid-October.

Unknown individuals were marked with a subcutaneous passive transponder (PIT, ISO transponder, Tierchip Dasmann, Tecklenburg) and an ear tissue sample was taken (maximum 2mm diameter, Thumb Type Punch, Harvard Apparatus) and stored in 90% ethanol at $-20°$C. DNA extracted from the tissue samples was genotyped for 18 specific autosomal microsatellites developed for this population (Wandeler, Ravaioli, and Bucher 2008), and the *Sry* locus was genotyped in order to confirm the sex of all individuals. To identify cases of PIT loss as well as recaptures of juveniles initially too light for PIT injection, an identity analysis in CERVUS v.3.0 (Marshall et al. 1998) was carried out to detect resampled individuals. Parentage was assigned to all juveniles and all first-time captured adults by simultaneously reconstructing parentage and sibship using the R package MasterBayes (Hadfield, Richardson, and Burke 2006). Analyses were performed for each year separately assuming polygamy for males and females and a uniform genotyping error rate of 0.5% for all 18 loci. Parentage was assigned using a parental pool of all adults present in the examined year and the previous year. Because some rare first year individuals reproduce at the end of the season, as evidenced by the observation of pregnant and lactating first year individuals, the “juveniles” were also included in the parental pool of a second analysis excluding parent-offspring mating. Thereby eight additional parentage links could be identified. There were no inconsistencies between the reconstructed pedigree and the transmission of two sex-specific markers: a polymorphic Y-chromosome locus developed for this population (Wandeler and Camenisch 2011) and a fragment of the mitochondrial DNA control region, amplified using vole specific primers (Haring, Herzig-Straschil, and Spitzenberger 2000). This pedigree was used to measure annual and lifetime reproductive success.

Apparent year-to-year survival could be obtained without mark-recapture modeling as the recapture probability on a given year was virtually 1: no animal was not captured in a year but captured later, and no animal was ever found to be a parent of a juvenile in a year when it had not been captured. This is not surprising since mark-recapture modeling within years estimated a between-occasion recapture probability of 0.924 (SE 0.012) for adults and of 0.814 (SE 0.030) for juveniles.

2.13 Univariate models of survival and reproduction in the snow vole population

The following two tables (2.1 and 2.2) present all the estimates from the univariate models used to estimate the individual-level variance in survival and reproductive propensities for the snow vole population.

Table 2.1: Estimates of coefficient of the mixed model for survival in the real snow vole data set

	Estimate	SE	p-value	Bootstrap 95% CI
Random effects:				
σ^2_{id}	0.000	-	0.500	[0;0.248]
σ^2_{year}	5.622	-	$< 10^{-16}$	[1.133;13.158]
Fixed effects:				
intercept	-1.754	0.830	0.035	[-3.393;-0.111]
age (Juvenile)	1.841	0.230	0.000	[1.369;2.411]
sex (Male)	0.306	0.295	0.300	[-0.389;0.93]
age:sex	-0.705	0.333	0.034	[-1.449;0.091]

Note: σ^2_{id} and σ^2_{year} refer to the variance between individuals and between years, respectively. All estimates are shown on the latent scale. The p-values for the significance of the two random effects are computed through a one-sided LRT. No standard errors (SEs) are provided for random effects. Instead, confidence intervals are computed using 1000 parametric bootstraps. The significance of the fixed effects is computed through the default Gaussian approximation provided by the package `lme4`.

Table 2.2: Estimates of coefficients of the mixed model for annual reproductive success in the real snow vole data set

	Estimate	SE	p-value	Bootstrap 95% CI
Random effects:				
σ^2_{obs}	3.3×10^{-10}	-	0.499	[0 ; 0.194]
σ^2_{id}	0.371	-	$< 10^{-16}$	[0.151 ; 0.475]
σ^2_{year}	0.101	-	$< 10^{-16}$	[0.026 ; 0.452]
Fixed effects:				
intercept	0.724	0.131	0.000	[-0.254 ; 0.266]
age (Juvenile)	-5.703	0.369	$< 10^{-16}$	[-7.425 ; -5.125]
sex (Male)	0.046	0.101	0.645	[-0.118 ; 0.200]

Note: σ^2_{id} and σ^2_{year} refers to the variance between individuals and between years, respectively. σ^2_{obs} is a dummy random effect having one level per observation and used to account for potential over-dispersion in Poisson GLMMs. The p-value testing for the significance of these three random effects is computed through a one-sided likelihood ratio test. The significance of the fixed-effects is computed through the default normal approximation provided by the package lme4. Confidence intervals are computed using 1000 parametric bootstraps. The interaction between sex and age was not estimable by lme4: its inclusion produced convergence warnings and its SE was above 10^4, without affecting other parameter estimates, and therefore it was removed from the model.

2.14 Estimation of the latent correlation between survival and reproduction

Here we provide additional details on the bivariate models to test for the latent correlation between the propensity to reproduce and the propensity to survive. See main text for more details on the univariate analyses.

In univariate models for reproduction fitted using lme4, neither the sex by age interaction, nor the dummy random effect controlling for overdispersion was significant. With MCMCglmm, the non-significance was confirmed by bivariate models using the deviance information criterion (DIC) and Bayesian credibility intervals for these two parameters. Moreover, by default MCMCglmm takes into account any overdispersion in a distribution assumed to be Poisson. Therefore we did not include these two explanatory variables in the final model. Posterior predictive checks revealed that the bivariate model correctly predicted the number of zeros for ARS (observed 820, predicted 807 ± 23). Moreover, the year-level covariance between survival and reproduction was estimated close to zero, and fixing it to zero improved DIC, so it was fixed to zero in the final model. Finally, the package MCMCglmm always includes a residual variance component for binary variables, although this variance is not estimable. We fixed this residual variance to 1, as suggested in the package course notes (http://www.cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf). This model can be written as:

$$\begin{pmatrix}\rho_{i,t}\\ \phi_{i,t}\end{pmatrix} \sim \begin{pmatrix}f_\rho\\ f_\phi\end{pmatrix} + \begin{pmatrix}\sigma^2_{\rho(year)} & 0\\ 0 & \sigma^2_{\phi(year)}\end{pmatrix} + \begin{pmatrix}\sigma^2_{\rho(ind)} & \sigma_{\rho\phi(ind)}\\ \sigma_{\rho\phi(ind)} & \sigma^2_{\phi(ind)}\end{pmatrix} + \begin{pmatrix}\sigma^2_{\rho(res)} & \sigma_{\rho\phi(res)}\\ \sigma_{\rho\phi(res)} & \sigma^2_{\phi(res)}\end{pmatrix}$$

where f_ρ and f_ϕ denote the fixed part of the model and both include an intercept, sex, age and their interaction. The σ^2 terms refer to variances and the $\sigma_{\rho\phi}$ terms refer to the covariances between ARS and survival, either at the level of years $_{(year)}$, of individuals $_{(ind)}$ or of the residuals $_{(res)}$.

The correlation between the individual propensity to survive and to reproduce was then calculated as $\sigma_{\rho\phi(ind)}/\sigma_{\rho(ind)}\sigma_{\phi(ind)}$. We used 1000 MCMC samples from 1,100,000 iterations with a thinning of 1000 and a burn-in of 100000. We used a non-informative parameter expanded prior. The residual variance of survival was fixed to 1, as this variance is not identifiable in binomial models. We then refitted the same model while fixing $\sigma_{\rho\phi(ind)}$, $\sigma^2_{\rho(ind)}$ or $\sigma^2_{\phi(ind)}$ to zero, in order to compare the DIC of the two models. Although model selection on the variance-covariance random components is an active area of research (e.g. Burnham and Anderson 2002, chapter 6), the use of DIC has been shown to be robust, at least under some conditions (Wilberg and Bence 2008; Barnett et al. 2010). All models were checked by graphically assessing convergence and good mixing, and using Heidelberg stationarity tests. Moreover, thinning was sufficient to keep all auto-correlations between successive samples below 0.05.

The Bayesian bivariate model identifies variance in the ability to reproduce, $\sigma^2_{\rho(id)}$. Although it is smaller than in the univariate model (table 2.4), it was still different from zero, as 97% of the posterior sample is above 0.01 and removing the random effect from the model substantially increases the DIC (table 2.3). Similar to the univariate model, the estimate of the variance in the ability to survive is small, with a large uncertainty. Including this effect in a model improves (i.e. decreases) DIC in one instance (model 4 versus model 5) but not in another instance (model 2 versus model 3), see table 2.3. However, this effect appears in the best model. There is thus a large uncertainty in the estimation of variance in the ability to survive and mixed evidence for its existence. Similarly, the correlation between the two individual random effects is estimated with a large credibility interval overlapping 0 (table 2.4), and the inclusion of this parameter improves only marginally the DIC of the models (table 2.3). Nevertheless, the mode of the posterior distribution is positive and the effect is present in the best model. Altogether, these results provide limited support for the biological significance of the latent correlation between survival and reproduction.

Table 2.3: Deviance information criterion (DIC) and difference to the best model (ΔDIC), for five bivariate models of ARS and survival with different individual random effect structures

model	$\sigma^2_{\rho(ind)}$	$\sigma^2_{\phi(ind)}$	$\sigma_{\rho,\phi(ind)}$	DIC	ΔDIC
1	Yes	Yes	Yes	2554.587	0.000
2	Yes	Yes	No	2556.793	2.206
3	Yes	No	No	2556.100	1.513
4	No	Yes	No	2560.945	6.358
5	No	No	No	2564.187	9.600

Note: A "Yes" indicates that the parameter was included in the model, a "No", that it was not. The parameters are $\sigma^2_{\rho(ind)}$, the individual-level variance in ARS; $\sigma^2_{\phi(ind)}$ the individual-level variance in survival; $\sigma_{\rho,\phi(ind)}$ the individual-level covariance between reproduction and survival. Note that it is possible to include $\sigma_{\rho,\phi}$ only when both $\sigma^2_{\rho(ind)}$ and $\sigma^2_{\phi(ind)}$ are also included in the model.

Table 2.4: Variance and correlation components for a bivariate model of survival and reproduction

	Posterior mode	95% CI
$\sigma^2_{\rho(ind)}$	0.167	$[1.4 \times 10^{-4}; 0.342]$
$\sigma^2_{\phi(ind)}$	8.9×10^{-3}	$[9.4 \times 10^{-7}; 1.048]$
$\sigma_{\rho\phi(ind)}$	0.322	$[-0.682; 0.974]$
$\sigma^2_{\rho(year)}$	0.122	$[0.030; 0.917]$
$\sigma^2_{\phi(year)}$	7.585	$[2.074; 73.123]$
$\sigma^2_{\rho(res)}$	0.230	$[1.4 \times 10^{-4}; 0.342]$
$\sigma^2_{\phi(res)}$	1	fixed
$\sigma_{\rho\phi(res)}$	0.180	$[-0.313; 0.576]$

Note: 95% CI shows 95% highest posterior density intervals.

3

Disentangling evolutionary, plastic and demographic processes underlying trait dynamics: A review of four frameworks

> *"Then why do you want to know?"*
> *"Because learning does not consist only of knowing what we must or we can do, but also of knowing what we could do and perhaps should not do."*
> — Umberto Eco, *The Name of the Rose* (1954)

Koen J. van Benthem*, Marjolein Bruijning*, **Timothée Bonnet***, Eelke Jongejans†, Erik Postma†, Arpat Ozgul†

* co-first authors

† co-last authors

3.1 abstract

1. Biologists are increasingly interested in decomposing trait dynamics into underlying processes, such as evolution, plasticity and demography. Four important frameworks that allow for such a decomposition are the quantitative genetic animal model (AM), the 'Geber' method (GM), the age-structured Price equation (APE), and the integral projection model (IPM). However, as these frameworks have largely been developed independently, they differ in the assumptions they make, the data they require, as well as their outcomes and interpretation.
2. Here we evaluate the way each of these frameworks decompose trait dynamics into underlying processes. To do so, we apply them to simulated data for a hypothetical animal population. We simulated scenarios with and without selection on body size, and with high and low heritability. In all four scenarios, body size also contained a plastic component.
3. The APE and IPM provided similar results, as did the AM and GM, with important differences between the former and the latter. All frameworks detected positive contributions of selection in the high but not in the low selection scenario. However, the APE and IPM did not distinguish between the high and low heritability scenarios, while the AM and GM did. Both the AM and GM revealed a high contribution of plasticity. In all scenarios, the APE and IPM attributed most of the change in body size to ontogenetic growth and inheritance, which include the effects of plasticity, maternal effects and heritability. We show how these apparent discrepancies are mostly due to the aims and definitions of the different frameworks. For example, the APE and IPM capture selection, whereas the AM and GM focus on the response to selection. Furthermore, the frameworks differ in the processes that are ascribed to plasticity and to their method for taking into account demography.
4. We conclude that because of the inherent differences among frameworks, no single framework provides the 'true' contributions of evolution, plasticity and demography. However, with a thorough

understanding of any of these frameworks, they will provide valuable biological insight. This work thus supports both future analysis as well as the careful interpretation of existing work.

3.2 Introduction

Understanding trait and population dynamics and how the two are intertwined is crucial for predicting population resilience and viability (**merila2014**). Hence, which processes shape population-level trait dynamics (i.e., changes in trait distributions over time) is a fundamental question in ecology and evolution, and one which is gaining in urgency given mounting concern regarding the consequences of anthropogenic environmental change for natural populations (**parmesan2006**).

Phenotypic trait distributions may be altered across generations by genetic (i.e. evolutionary) processes, as well as by non-genetic processes, such as phenotypic plasticity. Since the realisation that evolutionary and ecological processes may act on the same time scale, distinguishing between the role of evolution and plasticity has been the subject of a substantial body of research (**Hairston2005**; **Gienapp2008**; **Post2009**). To complicate matters further, changes in the demographic structure of a population may additionally shape trait distributions (**Coulson2008**). Hence, understanding and predicting trait dynamics ideally requires simultaneously taking into account all three processes (**Pelletier2007**; **Schoener2011**).

To date, four major frameworks aiming at distinguishing between the role of evolution, phenotypic plasticity and demography have been developed: 1) The quantitative genetic framework, particularly the animal model (**Henderson1950**), 2) the 'Geber method' (**Hairston2005**), 3) the age-structured Price equation (**Coulson2008**), and 4) the application of the APE in conjunction with an integral projection model (**easterling2000**; **Ellner2006**; **Coulson2010**). As of yet, several studies have tried to explicitly estimate the relative importance of evolution, plasticity and/or demography using one of these approaches (**Reale2003**; **Ezard2009**; **Ozgul2009**; **Rebke2010**; **Becks2012**; **Morrissey2012b**). Nevertheless, fully disentangling and quantifying evolutionary, ecological and demographic processes and thereby predicting the consequential trait dynamics has proven to be problematic (**Gienapp2008**; **Schoener2011**; **merila2014**). At least some of these difficulties can be attributed to the large amounts of (individual-based) long-term data required, which are often unavailable for natural populations (**Clutton-brock2010**). However, even if sufficient data are available, synthesis of the results from the four frameworks is hampered by the fact that they have been developed largely independently of each other. As a consequence, they differ in their focus and aims, and as we show here, they define biological processes in non-equivalent ways.

Here we provide a comparison of the differences, similarities and complementarity of each of these four decomposition frameworks by applying them to the same simulated datasets and comparing their outcomes. Thereby, we evaluate how they quantify the role of different ecological and evolutionary mechanisms in shaping trait dynamics under a range of biological scenarios. Together with a critical review of the theory underlying each of the frameworks, we provide comprehensive insight into their underlying assumptions, as well as the conceptual differences and similarities. This provides a much needed overview of the suitability of each framework with respect to both research questions and data availability.

3.3 Applying the four frameworks

3.3.1 Data simulation

Although it comes with the loss of some biological realism, using simulated rather than empirical data enables us to evaluate the frameworks under different scenarios and allows for replication. Furthermore, it ensures perfect knowledge of the processes that shape trait dynamics, acting as a reference against which to compare the results of each framework. Finally, simulated data do not suffer from the complications introduced by missing data.

Data were simulated using a two-sex individual-based model of a closed population of a hypothetical animal species, implemented in R (**R**). Here, we provide a brief overview, while a more complete description can be found in SI **??**. We also provide the R code on `https://github.com/koenvanbenthem/Disentangling_Dynamics_IBM`. We simulated a single trait, body size z. Size at birth is determined by an individual's genotype (10 loci, with 10 alleles each and mendelian inheritance, more details in SI **??**), the body size of its mother, and a stochastic component (drawn from a Gaussian distribution; SI **??**). Ontogenetic growth results in an increase of body size with age. Growth rate is proportional to body size and decreases with age, and is further influenced by per-capita food availability (SI **??**). Males were randomly assigned to females, who have a 50% chance of becoming reproductive after one year and whose reproductive probability increases with age. The litter size that a female produces depends on per-capita food availability, a stochastic component, and body size (SI **??**). Survival probability first increases with age, but starts decreasing after year five, reflecting senescence, and is further influenced by per-capita food availability and body size. Maximum age is 30 years. Furthermore, a trade-off exists between female reproduction and survival, i.e. reproducing at time t decreases survival probability to time $t+1$ (SI **??**).

Fifty time steps were simulated, of which the first ten were discarded from further analyses to allow the age structure to stabilize (Fig. **??**). After ten years, total food availability started to decline. As food is divided over all individuals alive in a given year, with some individuals randomly obtaining more than others, the decline affects population and trait dynamics through individual survival, growth and (female) reproductive success (SI **??**). The remaining data spanned 40 years (i.e. approximately 13 generations), which is comparable to the length of some of the field studies these frameworks have been applied to (**Clutton-brock2010**).

To evaluate the behaviour of the frameworks under different circumstances, we simulated four different scenarios. First, survival and fertility selection on body size was either present (s_+) or absent (s_0). Under the s_+ scenarios, there was a positive effect of body mass on survival and on litter size for mothers. Second, the relative importance of genetic variation in shaping body size, commonly measured as heritability, was either high (h_+) or low (h_-). This was done by using either of two pre-defined genotype-phenotype maps: one with big and one with small variation in the effects of alleles. Furthermore, the plastic component in birth size was varied antagonistically. The parameter values for each of the four scenarios (s_0h_-, s_0h_+, s_+h_- and s_+h_+) can be found in SI **??**. To evaluate the effect of stochasticity, each scenario was replicated 100 times.

Figure **??** provides an illustration of some of the key characteristics of the datasets simulated under each scenario. Despite a substantial amount of stochastic variation across replicates within each scenario, clear differences in trait and population dynamics are apparent. As expected, the s_+ scenarios show a positive relation between body size and annual fitness, calculated as the sum of the survival and litter size to $t+1$, whereas the s_0 scenarios do not (Fig. **??**). Furthermore, the proportion of the phenotypic variance attributable to variance in the simulated genotypic values (i.e. broad-sense heritability

H^2) was ca. 0.50 in the h_+ and 0.08 in the h_- scenario.

Although in all scenarios population size first increased (until year 20) and then decreased (Fig. **??**), the population size averaged across replicates reached up to 322 and 334 individuals in scenarios s_+h_- and s_+h_+, whereas in s_0h_- and s_0h_+ the maximum average population size was 245 and 252 individuals, respectively. Mean body size first increased rapidly, but started to decrease in all scenarios after the tenth year (Fig. **??**): in s_0h_- with -0.47 $[-1.45; 0.63$ 95% range among replicates], in s_0h_+ with -0.46 $[-1.59; 0.0.68]$, in s_+h_- with -0.75 $[-1.87; 0.08]$, and in s_+h_+ with -0.16 $[-1.12; 0.83]$.

Genotypic values for birth size, however, continued to increase only in scenario s_+h_+, whereby the change was 0.62 $[0.23; 1.04]$ (Fig. **??**). In s_+h_- a smaller increase was observed 0.08 $[-0.074; 0.24]$, whereas s_0h_- and s_0h_+ show on average no change in genotypic values. Correspondingly, average birth size increased only in the s_+h_+ scenario, with 0.58 $[0.092; 1.11]$, between year 11 and year 50 (Fig. **??**).

3.3.2 Decomposing simulated trait dynamics

Rather than providing an exhaustive overview of all methods allowing for the decomposition of trait dynamics, we have chosen to focus on four, commonly-used frameworks. The four frameworks have different data requirements (see Fig. **??** for a schematic overview) and do not yield identical results. This is illustrated in the following section, in which we analyse the simulated data using each of the frameworks.

Animal Model The animal model (AM) is a quantitative genetic method that was developed for commercial breeding (**Henderson1950**; **Henderson1976**), where it has been used successfully for several decades (**Lynch1998**). Only recently has it been applied to wild animal populations (**reale2003**; **Postma2014**) and plant (**Stinchcombe2014**) populations. For extensive explanations of the AM as applied to natural populations, see **Kruuk2004** and **Wilson2009**

The AM is a linear mixed effects model that is fitted to individual-level data. The variance in a phenotypic trait (z) is partitioned into genetic and non-genetic sources of variation. Under the assumption that this partitioning is additive (i.e. in the absence of genotype-environment correlations and interactions), z can be written as the sum of a population mean (μ), an additive genetic effect (the breeding value, a) and a residual (environmental) value capturing plasticity (e), thus $z = \mu + a + e$. Information on the relatedness between individuals (estimated from a pedigree or genetic markers) is used as a constraint in the fit, allowing for the estimation of a. If the data allow for it, other components contributing to variation in z, such as maternal, common, and permanent environmental effects can be accounted for explicitly. This variance decomposition can be used to predict response to selection on the long-term, assuming a quantitative genetic model, where ideally an infinity of genes contributes to phenotypic variation so that the distribution of effects is Gaussian. In practice, four or five bi-allelic genes are sufficient (**Roff2007**).

There are several ways to estimate evolution within the AM framework (see discussion), but here we illustrate only one of these. We fitted a univariate AM and quantified the change in the best linear unbiased predictors (BLUPs) for the breeding values over time (**Postma2006**; **Hadfield2010**). We used body size as the sole response variable, and intercepts for breeding values, maternal effects, permanent environment effects, and year effects were included as random effect. Maternal and permanent environment effects were modelled by fitting maternal and individual identity, respectively. Age was included as a continuous fixed effect (both as linear and quadratic terms). All fits were performed using the R-package `MCMCglmm` (**Hadfield2010**). The posterior distributions were estimated based on 1,000 MCMC samples, from 50,000 iterations with a thinning interval of 40 and a burn-in of 10,000, thus maintaining the correlation between successive samples of all parameters below 10%.

We estimated the temporal trend in the BLUPs for all random effects. We accounted for their uncertainty following **Hadfield2010b** by performing a regression of the BLUPs on time for each MCMC sample of the model. This provided a posterior distribution of linear slope coefficients, estimating the change in additive genetic, maternal, and permanent environment effects per time step. More details on the fitted models are given in SI **??**.

As depicted in Figure **??**, in all scenarios the contributions of evolution and individual plasticity were largest, while the contributions of permanent environment and maternal effects were very small. On average, we found positive changes in breeding values in both scenario s_+h_- (0.0013 $[-0.0038; 0.0095]$ and scenario s_+h_+ 0.014 $[0.00021; 0.029]$. Note that genetic drift introduces a substantial amount of variation in the rate of evolutionary change among replicates in the h_+ scenario, which is reflected in the large error bars in Fig. **??**). Negative contributions of individual plasticity were found, particularly in the scenarios with selection -0.02 $[-0.049; 0.0018]$ and -0.019 $[-0.045; 0.0029]$ for h_- and h_+, respectively.

In the s_0 scenarios, although drift generates variation around the mean contribution of evolution, we would expect the latter to be 0. However, our model inferred a genetic decline for h_- and h_+ in the scenarios without selection -0.0057 $[-0.016; 0.0040]$ and -0.0073 $[-0.024; 0.0087]$, respectively. This pattern was confirmed by the genetic values we simulated. The AM therefore estimates evolution with a negative bias. The reason seems to be that we fit a model not matching the simulation process. As mean size decreases with time, the contributions of mothers to birth size decreased too. Maternal effects are modelled as maternal identity rather than maternal current size, however, and the AM incorporates this change in the evolution term.

Geber method The 'Geber method' (GM) (**Hairston2005**) is a very general method that quantifies how temporal changes in various factors influence the response variable of interest. The biological assumptions vary with the specific factors and response variable. It may for example estimate how temporal changes in mean breeding value $\overline{a}$ and in an environmental factor k such as food availability propagate to a population-level response variable X, such as mean trait value. Examples of its application can be found in **Ellner2011** and **Becks2012**

For the implementation of the GM, we followed the analysis of fledgling mass of great tits in **Ellner2011** Here, the average body size is taken as the population-level response variable. The change in mean trait ($\overline{z}$) in the simulated populations was decomposed into a contribution of the environment ($\overline{k}$) and a contribution of a phenotypic change in size at birth. The latter was decomposed further into an evolutionary ($\overline{a}$) and a plastic component ($\overline{p}$):

$$\frac{d\overline{z}}{dt} = \frac{\partial \overline{z}}{\partial \overline{k}} \frac{d\overline{k}}{dt} + \frac{\partial \overline{z}}{\partial \overline{a}} \frac{d\overline{a}}{dt} + \frac{\partial \overline{z}}{\partial \overline{p}} \frac{d\overline{p}}{dt} \tag{3.1}$$

For each year between years 11 and 50, we calculated the mean body size ($\overline{z}$), mean size at birth of newborns, the average food availability that alive individuals had access to during their lifetime up to that moment ($\overline{k}$), and the mean breeding value as estimated by the AM ($\overline{a}$) (see above). Hence, as breeding values are typically not observed directly, the application of the GM to empirical data relies on other methods such as the AM for their estimation. Finally, we calculated a plasticity term ($\overline{p}$), equal to the difference between the average size at birth and the average breeding value for size at birth. Thereby this term only captured plasticity in mass at birth. We fitted a linear model to estimate the effects of $\overline{a}$, $\overline{p}$ and $\overline{k}$ on $\overline{z}$. Using this model, together with separate linear models that describe how each of the three underlying factors changes over time, we evaluated their respective influence on $\overline{z}$. This procedure is described in more detail in SI **??**.

The results of the GM are shown in Fig. **??**). The results for the evolutionary component are, as expected, nearly identical to the results of the AM. This evolutionary component is counter-acted by a decrease in food availability, as is shown by the negative 'environmental' contributions. The latter is largest for the s_+ scenarios, under which population size is higher (Fig. **??**) and per capita food availability therefore lower.

The average contributions of plasticity are more equivocal. Here we would have expected a very minor negative contribution for most scenarios, to account for the slight decrease in maternal body size and therefore a lowering of the maternal effect. However, these effects do not show up and instead we see mainly positive contributions. This contribution mainly serves to counteract the biased estimates in breeding values from the AM. When the analysis was repeated with the genotypic values from the simulations instead of the estimated breeding values, all scenarios showed negative trends (SI **??**).

Age-structured Price Equation The age-structured Price equation (APE) (**Coulson2008**) is an extension of the Price equation (**Price1970**). The APE focuses on phenotypes only and makes no assumptions on underlying genetics. It assumes that the change in mean trait value can be decomposed additively in seven contributions. All these contributions are either averages of or covariances between several measured properties (e.g. individual survival and body size) among individuals at a given time.

The two selection terms describe how selective disappearance and selective reproduction alter the mean trait value (viability selection, VS, and fertility selection, FS, respectively). Here VS, is the covariance between z and survival, which scales with the difference in the average trait value of the whole population and the part of the population that survives to the next time step (**rebke2012**). This is referred to as the selection differential in evolutionary literature (**Robertson1966**; **Lande1983**).

The change in mean trait value due to ontogenetic development of surviving individuals is captured by the growth term. The two demography contributions, here combined into one, describe change resulting from the age structure (i.e. the covariance between average body size of a cohort, and the change in proportional size of a cohort) (SI **??**). Finally, the two inheritance related contributions, also combined into one (SI **??**), represent the contribution of differences between offspring and parental trait values.

The five remaining contributions are calculated per age class, thereby taking into account that they depend not only on the trait value of an individual, but also on its age. The total contribution is obtained by a weighted sum of the age specific contributions. The APE thereby allows for an exact decomposition of $\Delta\overline{z}$ in discrete time into components of viability selection, fertility selection, ontogenetic growth, inheritance, and demography in populations with overlapping generations. See SI **??** for the full equation and an explanation of the terms. The APE has been applied to a range of mammals species (**Coulson2008**; **Ozgul2009**; **Ozgul2010**; **Canale2016**). Note that a stage-structured version of the Price equation has also been developed (**Barfield2011**).

As is commonly done in demographic analyses, we applied the APE to the female part of the population. Under the s_0 scenarios, we find that the average VS and FS are both indistinguishable from zero (Fig. **??**). For the s_+ scenarios, the contribution of selection is positive, but there is no difference between the s_+h_+ and s_+h_- scenarios (VS: 0.08 and 0.09 respectively, FS: 0.05 and 0.06 respectively). Finally, the demographic contribution differs between the s_0 and s_+ scenarios, but is largely unaffected by the heritability. This combined demography term scales with the between-age class covariance between fitness and body size (SI **??**). In our simulations, this is strong and positive, as older age classes are larger, and larger individuals have higher fitness in the s_+ scenarios. The negative contribution in the s_0 scenarios is the result of a negative effect of age on survival, which in the absence of positive selection will dominate the between-age class covariance. The biggest contribution to changes in average body size, comes from ontogenetic growth. This component is slightly lower in the s_+ scenarios, due to smaller per capita food availability.

The inheritance term measures changes in average body size due to the difference in the mother's body size (at time of giving birth) and the offspring's size at birth. This leads to the inheritance contribution being negative. Again, differences are most prevalent between s_+ and s_0. Due to selection, larger mothers produce more offspring. For these mothers, the difference in size between mother and offspring will be bigger (the maternal trait value is higher, the offspring's trait value will not increase by the same amount), leading to a more negative contribution. Finally, we see that the h_+ scenarios have slightly smaller (less negative) contributions from inheritance than the h_- scenarios. If heritability increases, this will decrease the mother-offspring difference, leading to this small effect.

Integral Projection Model The integral projection model (IPM) describes population and trait dynamics by considering four life history processes: survival, reproduction, growth and inheritance (**Ellner2006**). The dependencies of these processes on a continuous phenotypic trait z are estimated using regression models. No assumptions concerning the underlying genetics are made. Based on these regressions, the trait distribution at time $t+1$ can be predicted from the trait distribution at time t (**Adler2010**; **merow2014**)). Over the past years, IPMs have been used to address a range of eco-evolutionary questions (**metcalf2008evolution**; **Smallegange2013**; **traill2014demography**).

The specific decomposition we use, involves applying the APE on the IPM, as proposed by **Coulson2010** Note that different approaches exist (**Coulson2011modeling**; **traill2014demography**), which instead use a sensitivity analysis.

An IPM was parametrized for each simulated dataset, and as was the case for the APE, we only considered females. Models describing individual growth, survival, reproduction probabilities and number of offspring were fitted using generalized linear mixed models with appropriate link functions (logit for survival and reproduction probability, log for number of offspring). Inheritance was estimated as a linear regression of offspring size at birth on the size of the mother at the time of giving birth, as done in **traill2014demography** The slope of this regression differs from the heritability (h^2) of body mass in the AM, which relates offspring size to the mother's size, both at the same fixed developmental stage (e.g. birth) (**chevin2015**). For all life history processes, we tested five different models: a full model containing age, size and their interaction, as well as all models nested within this full model. Furthermore, each model included a random effect for year. The model with the lowest AIC was selected and used for the IPM.

Using all models a 3100×3100 matrix was parametrized (i.e. 31 age classes, 100 size classes per age class, ranging between 1 and 50 kg) for each replicate. See SI **??** and **??** for more details on model fitting and the construction of the IPMs.

For each IPM, we used the observed population vector at each time step (excluding the first ten years) to project the population vector to the next time step ($t+1$). Changes in population structure, and thereby changes in $\bar{z}$, are decomposed into contributions from different life history processes (Fig. **??**). Very similar patterns as in the APE were found. The largest contribution to $\Delta\bar{z}$ was attributed to inheritance in all scenarios, and was more negative in the s_+ scenarios. Furthermore, there was a consistently positive contribution of ontogenetic growth, with weaker effects in the s_+ scenarios, again due to lower per capita food availability. As with the APE, we considered both demographic terms together, these showed positive contributions that were stronger in the s_+ scenarios.

Both viability and fertility selection were strongest in the s_+ scenarios (VS was 0.045 and 0.041; FS was 0.012 and 0.012, for h_- and h_+). In contrast, in the s_0h_- and s_0h_+ scenarios, average viability selection was -0.024 and -0.019, respectively, and fertility selection was -0.00069 and -0.00068.

To allow for a better comparison with other frameworks, here we focus on the average value of $\Delta\bar{z}$, and how much various processes contribute to this. When quantifying how much of the *year-to-year variation* in $\Delta\bar{z}$ is explained by each process (**Ozgul2009**), the IPM and APE provide more divergent results (SI **??**).

3.4 Discussion

We have decomposed changes in mean body size into underlying processes by applying four major frameworks to simulated data. Thereby we have shown that these frameworks differ substantially in their data requirements and in their actual decompositions (Fig. **??**). In the following sections we will discuss and compare the theory underlying the four frameworks, illustrated by our simulations. We will discuss the inherent differences among frameworks regarding evolution, plasticity, demography, and measures of uncertainty. We finish by addressing the applicability of each framework with respect to data availability and the research question at hand.

We have simulated scenarios with and without selection on body size, and with low and high heritability. As multiple processes influence and interact with body size, these scenarios resulted in divergent and relatively complex population and trait dynamics (Figure **??**). For example, in addition to additive genetic effects, size at birth was influenced by maternal effects and stochasticity. Moreover, ontogenetic growth was subject to both stochastic variation and a decrease in per-capita food availability. We also included a trade-off between viability and fertility. It is exactly this complexity that highlights the need for a robust framework that allows disentangling the underlying processes and quantifying their importance.

Selection and evolution

All four frameworks infer positive selection on body size in the s_+ scenarios, but not in the s_0 scenarios (Fig. **??**). The APE and IPM detect positive viability and fertility selection in both the s_+h_+ and the s_+h_- scenarios. The AM and GM detect a strong increase in mean breeding values in the s_+h_+ scenario and a small yet positive contribution in the s_+h_- scenario. Importantly, the contribution of evolution, as obtained in the AM and GM, is not equivalent conceptually nor numerically to the contribution of selection, as obtained in the IPM and APE.

This is highlighted by the fact that the AM and GM estimate a much larger contribution of evolution in the s_+h_+ compared to the s_+h_- scenario. This contrasts with the IPM and APE, where the contribution of selection in the inferred contribution of selection is independent of the heritability. This is the result of the fact that the AM directly estimates evolution by evaluating the change in mean breeding value, which is close to zero in the h_- scenarios. The same applies to the the GM, since it relies on the AM. The APE and IPM frameworks on the other hand estimate the contribution of phenotypic selection, irrespective of an evolutionary response to this selection.

Furthermore, due to a model misspecification for the maternal effects in the AM, we find a negative contribution of evolution to the s_0 scenarios. This mismatch highlights that the model structure should be adapted to the study system. Only then reliable conclusions can be drawn from the AM

(**Hadfield2011**). Contributions are closer to the simulation process when using a more appropriate specification of the maternal effects (SI **??**).

Here we have chosen to quantify the contribution of evolutionary change to trait dynamics by measuring the temporal change in BLUPs for breeding value in a univariate animal model. Within a quantitative genetic framework, we could also have used the heritability estimated by the AM to apply the breeder's equation and estimate the expected response to selection. This approach has proven its effectiveness under breeding conditions, although nonlinearities in the parent-offspring regression or the trait value-fitness relationship may bias predictions (**Heywood2005**). More serious difficulties arise in natural populations, where the prediction of evolution can be biased when selection acts on genetically correlated traits or when an environmental variable dominates the covariation between traits and fitness (**Rausher1992**; **Morrissey2010**).

A third approach relies on a bivariate AM that estimates genetic and environmental (co)variances between a trait and a proxy for relative fitness (**Lande1979**; **Lynch2014**). The additive genetic covariance is of particular interest, as it provides a direct estimate of the expected evolutionary change, i.e. the Robertson-Price identity (**Robertson1966**; **Price1970**; **Lynch1998**). Although more data demanding, this approach does not require the assumptions of the breeder's equation (**Morrissey2012**), and avoids potential biases in trends in breeding values (**Postma2006**).

Unlike the AM and GM, which quantify the change in breeding values, the APE and IPM estimate the contribution of selection, irrespective of whether this yields a genetic response. The overall contribution of selection is obtained by summing over all age-specific selection contributions. This is an attempt to remove the between-age covariation between traits and fitness (**engen2014b**), which is instead captured by the demography term. However, the age correction is not continuous, and therefore the choice of age classes determines how this total contribution of demography and selection is partitioned (see SI **??** for an example).

Most studies that have applied the APE or IPM framework to natural vertebrate populations have found a relatively small role for selection in shaping trait dynamics (**Ozgul2009**; **traill2014demography**). This is in line with our application, as even in the s_+ scenarios, the contribution of the other processes was estimated to be many times larger. In the IPM, the interpretation of selection in terms of evolutionary potential critically depends on the heritability. Heritability is, however, not necessarily reflected by the inheritance function. Indeed, the latter relates juvenile to adult (maternal) trait values, and thereby ignores the fact that individual growth trajectory may be heritable (**chevin2015**). This definition of the inheritance function can thus result in an underestimation of the heritability of adult trait value across generations. An alternative way to incorporate trait inheritance in the IPM, is by implementing size at birth as a fixed trait influencing offspring size (**vindenes2015**). A more rigorous method to model the transmission of additive genetic variance has recently been proposed by **Coulson2015** and **Childs2016** who formulated a framework that incorporates quantitative genetic inheritance into IPMs.

Plasticity

Plasticity includes all phenotypic changes that are not attributable to genetic changes. While all four frameworks estimate a large contribution of plasticity, across all scenarios, they differ in the biological processes that this is attributed to. This makes it difficult to directly compare the importance of plasticity across frameworks and may potentially lead to confusion. In this section we will focus on plasticity of birth size.

We used the AM to separately estimate plasticity due to maternal effects and permanent environment

(Fig. **??**). The contribution of maternal effects was very small. This may seem at odds with the effect of maternal adult size on offspring size at birth in our simulations, but as explained above, this was due to a mismatch between the model structure and the data generating process. The contribution of permanent environment was low, because there was no trend in the stochastic component of birth size.

The GM captures plasticity in size at birth due to both maternal effects and stochasticity in one single term (Fig. **??**). Here, plasticity at birth is defined as the difference between actual birth weight and the breeding value for birth weight of an individual. As mentioned in the results, the plasticity mainly serves to compensate for the bias in breeding value estimation.

In the APE and IPM frameworks, plasticity at birth and growth are intrinsically entangled. In the results, ontogenetic growth forms the main plastic contribution to $\Delta\bar{z}$ (Figs. **??** and **??**). However, the body size that is attained through ontogenetic growth is only partially (through maternal effect) transmitted to the offspring. Most of the ontogenetic growth will thus be reset in the offspring: this is reflected in the strong negative contribution from inheritance (for a more detailed explanation of the inheritance terms, see SI **??**). Because we applied the APE only on the female part of the population, changes in offspring body size due to selection on males (and thus fathers) will show up in the inheritance term.

The role of the environment

Of the focal frameworks, only the GM defines an explicit environmental factor. In the other frameworks, the environment influences trait dynamics only indirectly through selection, plasticity and/or demography. For example, high food availability may lead to an increase in average body size through plasticity. At the same time, increased food availability may decrease competition, and thereby affect selection.

In our implementation of the GM, we defined the environment as the total food intake of an individual. Hence, the environment mainly acts through within-individual plasticity, due to its effect on ontogenetic growth. Importantly, the outcome of the GM depends completely on the chosen definitions of evolution, plasticity and environment. When applying the GM to field data, where not all processes are known, it is thus crucial to first identify and define the processes of interest.

Although in the APE and IPM effects of the environment are implicitly present in all terms, in our implementation there is no explicit quantification of this environmental effect. However, in alternative applications of the IPM, this can be achieved by incorporating environmental factors (such as food availability) as explanatory variables in the functions that underlie the IPM (**vindenes2011**). Alternatively, one can parametrize different IPMs for different environments (**Ozgul2010**) and use comparison methods such as life table response experiments to see how population and trait dynamics differ between these environments (**rees2009**).

In our version of the AM, all contributions of changes in the environment, such as decreasing food availability, are captured within the residual individual plasticity term. Environmental contributions can be estimated more explicitly by including additional fixed or random effects (**Charmantier2014**). One possibility is the inclusion of a fixed effect of food availability. Furthermore, it is possible to model interaction between the environmental variable and one or more random effects, for example to account for genotype-environment, -age or -sex interactions.

Demography

We showed how the combined demography terms in the APE scale with a measure of between age class selection (i.e. the covariance of age class specific fitness and age specific average body size). It does thus not reflect changes in the age structure, but rather differences in fitness between age classes due to differences in average body size among age classes. As such it provides a demographic correction on selection, similar to the one proposed by **engen2014b**

In the AM we have quantified demographic contributions by multiplying the slope of body size with respect to age with the predicted change in average age. This contribution is most negative in the s_+ scenarios, meaning that here a change (decrease) in the average age in the populations over time led to a decrease in the average body size in these scenarios.

Unexplained variation and uncertainty

Making conclusive statements regarding which factor has the largest influence on $\Delta\bar{z}$ requires a measure of the uncertainty in the estimate of the contributions. So far we have only considered the range of point estimates over the replicates, generally showing smaller ranges for APE and IPM. However, APE and IPM were estimating processes that were constant throughout replicates (e.g. selection), whereas the AM and GM were estimating quantities subject to stochasticity (e.g. genetic drift). Differences in range are thus due to the stochasticity in the simulations rather than the uncertainty in the point estimates.

While the AM allows the estimation of confidence intervals for each estimated contribution, in the current use of the IPM, APE and GM there is no direct measure of uncertainty. However, for the GM, confidence intervals can be obtained using bootstrapping methods (**Ellner2011**). Although measures of uncertainty accompanying parameter estimates could be propagated to the decomposition in the IPM, as of yet, the lack of uncertainty quantification is a major drawback of the application of the IPM and APE. Uncertainty in the final estimates could, however, be estimated by using bootstrapping, or MCMC sampling in the case of the IPM.

Residual variance is explicitly quantified in the AM, the GM does evaluate the residuals of the underlying regressions, but does not include these in the final results (**Ellner2011**). In contrast, the APE is an exact framework and hence the residual variance is zero. However, it is still subject to sampling variance. Although the IPM uses the APE, it is constructed by fitting statistical models to the data, each with their own residual term.

The AM can also account explicitly for additional sources of variation, by including the corresponding random effects (here, we for example incorporated individual identity as a random effect to account for individual heterogeneity that could not be explained by additive genetic variation). The IPM framework that we discuss here, does include a random individual effect in all fitted functions underlying the IPM. However, although this inclusion accounts for individual heterogeneity when estimating vital rates, this individual heterogeneity does not propagate to the actual IPM. Instead, the IPM is often parametrized with the random effect set to zero and does not take into account all individual heterogeneity. Furthermore, this might bias the prediction because of Jensen's inequality (**Fox2002**). Individual heterogeneity may be incorporated by defining a "static trait", in addition to the continuous state variable. This static trait does not change during development, and reflects fixed individual heterogeneity caused by e.g. differences in size at birth, genetics or experienced environment (**Ellner2006**; **vindenes2015**). The role of individual heterogeneity is not captured in the GM and APE. However, in case of the GM, the effects of individual heterogeneity, as estimated by the AM can be propagated to the response variable.

Conclusions and future directions

The urge for a better understanding of eco-evolutionary dynamics is reflected in the range of frameworks that have been developed over the last few years aiming at quantifying the underlying processes (**Pelletier2009**; **Schoener2011**), especially within the light of the consequences of climate change (**Gienapp2008**; **Lavergne2010**). Yet, a general, predictive framework is lacking, and applications to field data remain scarce. We have shown that the animal model (AM), 'Geber' method (GM), age-structured Price equation (APE) and integral projection model (IPM) frameworks differ in generality and data requirements. Importantly, key processes are defined and interpreted differently in the different approaches. We emphasize that one should be careful when applying one of the frameworks and interpreting the outcomes as being the "true" contributions of different processes. Indeed, we have shown that depending on the used framework, one could come to different conclusions with respect to the importance of evolution, plasticity, environment and demography in influencing trait dynamics.

All four frameworks have only recently been proposed in their current form, and are only starting to be applied to conservation-related questions. In this review we have explored the frameworks and their assumptions and limitations. Our findings are summarized in Table **??**, where we provide an overview of which framework seems most suitable for which research question. The AM enables estimation of quantitative genetic parameters, and genetic change in particular, that cannot be estimated by the other frameworks. However, the AM, and the estimation on quantitative genetic parameters in general, is data demanding and it can be difficult to isolate confounding sources of variation when data sets are small. When individual data on reproduction, survival and growth are available, and one is interested in explicitly quantifying the contribution of within age class selection, IPM and APE are logical choices. The AM can explicitly evaluate the effect of individual heterogeneity. Although the IPM can take this information into account as well by fitting mixed effects models, it does not evaluate its effect on trait dynamics. The GM, in contrast to the other frameworks, focuses on population-level parameters, but knowledge (or assumptions) on processes is required beforehand, i.e. it must be known what processes are shaped by evolution (or plasticity) and which by the environment.

Overall we conclude that there is no single best framework, but each framework answers slightly different questions and has different data requirements. By highlighting both the similarities and the differences, we hope to have aid the interpretation of existing work. Furthermore, we hope this work will help researchers interested in eco-evolutionary questions in making an informed choice regarding the most suitable framework for their particular question.

Acknowledgements

We thank Robert O'Hara, Jarrod Hadfield, Loeske Kruuk and one anonymous reviewer for their help in considerably improving this manuscript. We also thank Tim Coulson for comments on an earlier version of the manuscript. Finally we are grateful to Luis-Miguel Chevin for providing us with constructive feedback throughout the process. This work was funded by ... the Swiss National Science Foundation project grants (31003A_141110) awarded to EP and (XXX) awarded to AO.

Table 3.1: A selection of research questions and to what extent frameworks may be used to answer them, ranging from impossible without major modifications ($--$) to being answered by the standard formulation of the framework already ($++$).

Question	AM	GM	APE	IPM
Does the change in trait value have a genetic basis?	$++$	$+$	$--$	$--$
Is selection acting on the trait?	$+$	$+$	$+$	$+$
Is the trait heritable?	$++$	$\pm$	$-$	$-$
Is a change in age structure responsible for the change in mean trait value?	$++$	$\pm$	$++$	$++$
How does individual heterogeneity affect trait value z?	$+$	$\pm$	$--$	$-$
How do eco-evolutionary dynamics affect population dynamics?	$-$	$+$	$-$	$++$

Supplementary information

4

The stasis that wasn’t: Adaptive body mass evolution is opposite to phenotypic selection in a wild rodent population

5

Fluctuating selection

6

General discussion

It is difficult to understand the universe if you only study one planet.
— Miyamoto Musashi, *A Book of Five Rings* (circa 1645)

7

Acknowledgments

I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve.
— J.R.R. Tolkien, *The Fellowship of the Ring* (1954)

Je feins l'adulte, mais, secrètement, je guette toujours le scarabée d'or, et j'attends qu'un oiseau se pose sur mon épaule, pour me parler d'une voix humaine et me révéler enfin le pourquoi du comment.
— Roman Kacew, dit Romain Gary, *La Promesse de l'aube* (1960)

Looking back to one's life, it is amazing to notice how much the path it took has been influenced by a myriad of people. These acknowledgments are about a PhD thesis, but they are bound to look back in time far beyond the PhD onset.

First and foremost, many thanks to Erik Postma for being such a great supervisor. Erik set the track for a fascinating and successful PhD project when he realized the potential of the snow vole monitoring and wrote a thoughtful NSF proposal full of surprisingly correct guesses about the biology of the population. Erik taught me statistical techniques, ideas from quantitative genetics and presentation techniques. More importantly, though, he trained me to consider logical arguments critically, read scientific publications in depth (in particular during epics journal clubs that ended up "rejecting" more than half of what we read based on technical or logical flaws) and thus learn from the mistakes of others, to learn a bit less from mine. For four years, Erik's door was literally *always* open (I cannot recall a time when he did not have some time to discuss) to answer my "little questions" and review my work. Finally, it impossible not to mention that Erik introduced me to running, pushed me to run more often, longer and faster, and thus kept me fit, healthy and entertained.

Not being one of his students, and given his very tight schedule, I am especially grateful to Lukas Keller for long conversations, old paper mining and spontaneous suggestions. His seemingly encyclopaedic knowledge, and his almost as rich library, not only solved a few crucial issues relevant to my PhD, but also kept acute my appetite for the wonders of population genetics, organismic biology and statistics.

Thanks to Arpat Ozgul for his constant enthusiasm about the project and for supporting its continuation during the postdoc to come. Also, thanks for being a bridge to the ideas and slang of population ecology and demography, thus clarifying "translation" issues.

Thanks to Barbara Tschirren for her precious suggestions when trying to understand the voles in a more mechanistic way, molecular and physiological. Thanks for letting me take your brand new respirometer in the humidity, low pressures and shakiness of the field.

The most important analyzes of my PhD were run with the package `MCMCglmm`, and I was extraordinarily lucky to count his creator, Jarrod Hadfield, among my committee members. Besides technical support, Jarrod helped clarifying some confusing concepts and results from quantitative genetics. Also, thanks to him for giving me the opportunity to review for the journal Evolution.

Thanks to Marc Kéry for his help and encouragements with complicated JAGS models, but also for some sound life-planing advice.

Thanks to Peter Wandeler for initiating the snow vole monitoring in Churwalden, while I was still a high-school student, unaware of the scree little happy folks.

I could never thank enough Glauco Camenisch for the help he provided in the lab, on the field, in front of the computer and in the kitchen. Glauco saved me many weeks of work and made constant efforts to improve the quality of data set. He would deserve to be awarded a good share of this PhD. With Glauco, the other backbone of the research group is Ursina Tobler. Ursina always keeps the administrative duties minimal on our side, but makes sure everything works fast and smoothly in the background. Thanks to her for making the live of students so easy.

During the last three years, the majority of my scientific discussions were with Koen van Benthem. I am grateful for his many ridiculous questions and discoveries, most of them turned out to be puzzling or insightful. A particular thank for pushing me towards LaTeXand Zurich Tonhalle. I wish he will one day understand the pattern behind prime numbers, and bring more mathematical rigor into ecology and evolution. Thanks also to him for taking the lead (i.e. first first authorship or sixth last authorship) on our "Price's equation review" (from which Price's equation was eventually extirpated), a long and tedious, but incredibly educational, collaboration, for which I also have to thank Marjolein Bruijning and Eelke Jongejans.

Vicente García-Navas can do research faster than a snow vole can run into a trap, and he publishes it in good journals before the apple is all eaten. It was very stimulating and good for my CV to work with him.

Andres Hagmayer was an exemplar Master student, very serious, autonomous and efficient. Thanks to him for almost two years of a fruitful collaboration that taught me a lot about teaching.

Thanks to Cindy Canale for helping with respirometry, for kind encouragements and nice parties. Many thanks to Dominique Waldvogel and Martina Schenkel for precious and cheerful help on the scree.

For four years, the majority of my non-sleeping time was spent in the company of the "permanent residents" of the office Y13-J-34. I will remember this office as home (although I kept my word and never spent a night there!) thanks to all our passionate discussions, our ire at wrong papers, our debugging struggles and our coffees breaks. Special thanks to Pirmin Nietlisbach for hosting me on Mandarte island; to Philipp Becker for teaching me how to cross-country sky and how to catch dippers; and to Judith Bachmann for nice frog expeditions. This office was such a great place to work, live and laugh with the essential contribution of its intermittent occupants, Erica Ponzi, Vanja Michel, Rien van Wijk and Johann Hegelbach.

Thanks to the Kokonuts—Hanna Kokko, Isobel Booksmythe, Anaïs Tilquin, Nina Gerber, Susanne Schindler and Xiang-Yi Li—for inspiring journal clubs, where you can sit on an orange primitive salamander (definitely not a pony). Christine Grossen and Daniel Croll are among the kindest humans I have ever met. In particular, thanks for welcoming me during my first days in Zürich and for lending me a piano for two years. Merci to Chelsea J. Little for being a great adventure buddy, a cheering friend and an example of scientific brightness.

Many thanks (sadly impersonal so that the acknowledgments do not become a thesis chapter on their own right) to the Zürich folks for four very fun and enriching years. In a jumble, thanks to Alexandra Jansen van Rensburg , Alexander Nater , Irene Weinberger , Simon Evans , Jasmin Winkler , Joel Pick , Mélissa Lemoine , Dennis Hansen , Stefanie Muff , Rassim Khelifa , Debbie Leigh , Josh van Buskirk , Kasia Sluzek , Frédéric Guillaume , BÃ©atrice Nussberger , Benedikt Gehr , Nina Vasiljevic , Ulrich Reyer , Hedwig Ens , Wolf Blanckenhorn , Tina Cornioley , Sam Cruickshank , Mollie Brooks , Juan Pablo Busso , Franziska Lörcher , Jobran Chebib , *Cini* Gabriella Gall , Andreas Sutter , Aurélie Garnier

, Nino Magg , Inge Juszak , Gianalberto Losapio . . .

Special thanks to Ashley E. Latimer for forcing me to cook and go birdwatching during stressful periods, for teaching me some English and for inspiring conversation bridging gaps between population genetics and evolution over geological timescales.

I never had the opportunity to thank all the great people I have met since the beginning of my scientific life (because I didn't write a report for the two first internship, and acknowledgments were *prohibited* (sic) for my master thesis). Still, I owe them a lot, and many people were instrumental in setting the path toward a PhD. In 2011, while I was trying to ~~escape from~~ leave engineering for fundamental research, without having much clue of what it was about, Jean-François Martin gave me the opportunity to give it a try by tutoring a gap year entirely dedicated to research. Pierre-André Crochet provided determinant help to find my way in the cloud of biology and reach a first research experiment, in the form of an internship at University of Oslo CEES. There, Glenn-Peter Sætre was an amazing first supervisor, extremely patient with my childish English, my misadventures in the lab and my perfect ignorance in evolutionary genetics, and was very supportive to find housing and survive one semester in Norway without any income. Thanks to him and all his team, especially Tore Oldeide Elgvin and Anna Fijarczyk, I had a great time that convinced me that fundamental research in evolutionary biology would be a good way to spend my time. This might be where I learned the most about what the life of an evolutionary biologist was about: from sequencing DNA, to searching and understanding scientific literature; from the quasi-universal fascination of scientists for coffee and Friday beer, to the poetic insight of Kimura's neutral theory, and many more things. Eventually, thanks to them for making me publish my first research papers, although I did not do a lot and was probably not really understanding the little I did. I then went to Chizé CEBC where I worked with a crew of amazing people including Adrien Pinot, Vincent Bretagnolle, David Pinaud, Vincent Lecoustre, Edoardo Tedesco, thanks to them and all the other Chizéens for all they taught me. A particular thank to Mathieu Authier for his (partially-successful) attempt to convert me to Bayesianism, and to Laurent Crespin for forcing me to go deeper into Mark-Recapture modeling and maximum of likelihood. Last but not least, Bertrand Gauffre was crucial in my scientific development, thanks to him for our many conversations, for taking me seriously, for introducing me to Richard Dawkins' writings and for pushing me to do complete my Masters in Montpellier. The B2E Master in Montpellier was a time of hard studying, and great fun, during which some teachers opened my mind to new scientific horizons, thanks in particular to Patrice David, Isabelle Olivieri, Olivier Gimenez and Michel Raymond for that. Thanks to Raphaël Leblois for introducing me to the coalescent theory and its powerful thought experiments, as well as to `C++` programming. Coding `ForwardBackward` and its improved sequels was a crazy adventure, full of traps and wonders. Thanks to François Rousset for giving me a small sight at what it means to understand population genetics, statistics, logic and the world in general. Thanks again to PAC for his patient help explaining me again and again evolutionary concepts and writing tricks. Also, thanks to Nicolas Bierne for his PhD offer and our fascinating discussion on oceanic streams, gene flow and the elusive eel, I still think about this alternative scientific and life path with curiosity and envy.

Thanks to all the Montpellierains, in particular Émeric Figuet, Léo Grasset, Joane Elleouet, Paul Sanders, Pascal Milesi, Julie Landes. . . our conversations account for of what I know in evolutionary biology. Out of the academic world, I could always count on the support of Gaëlle Jeanne Duranthon during the last 12 years. Thanks also to my two adoptive families who kept me cheerful with meals, games, music, remote places, short nights and craziness during the last decade. These are the Goret family with an otter, a weasel, a canary, a baboon, a snail, a goat, a bee, a goose, a panda, a wolf, a bear, an owl, a lion and a chameleon; and the Margicon (©*une marque déposée par terre*) family: Nounou, Chonchon, Caca, Doudou, Mirabelle and their mates/cats/penguins.

My early interest for ecology and evolution was much reinforced by wandering across the forests, meadows and moorlands of Lacaune mountains, by the contemplation of glittering and diverse carabids, by the excitement of long days monitoring bird migration, by astonishing conversations on conservation and nature management and by the constant realization that while we still know so little, it takes only some sweat and patience to discover new things. This early encounter with life would have not been sustainable without the naturalists who taught me so much and kept me amazed. Thanks in particular to Amaury Calvet and other members of the "Ligue pour la Protection des Oiseaux du Tarn" as well as to the members of BeaOsea: Adrien Chaigne, Camille Denozière, Aurélien Salesse, Denis Guillaumin and Manon Ghislain.

Some high school teachers were particularly influential not so much by what the taught, but rather by how they taught to learn, to see things in another way or by opening unexpected opportunities. Many would deserve acknowledgments for doing that at diverse degrees, but here are a few whose classes remain very alive: Benoît Leviandier, Mdm Rolland, Jacky Cariou, Jean Leblanc.

Being born in such a family made life rather easy. Thanks especially to my parents Francine and Claude, to my siblings Kiyomi-Élodie and Cyrille, to my grand-parents Juliette, Louis, André and Simone and to my uncle Laurent. They kept supporting, sustaining and trusting me over long years of studies they could less and less understand.

8

CV Timothée Bonnet

8.1 Personal Data

Birth date:	November 1st 1988
Nationality:	French
Work address:	Department of Evolutionary Biology and Environmental Studies University of Zürich Winterthurerstrasse 190 CH-8057 Zürich Switzerland
Phone:	+41 (0)44 635 47 66
email:	timothee.bonnet@ieu.uzh.ch
Personal page	*http://www.ieu.uzh.ch/staff/phd/bonnet.html*
Languages:	French (mother tongue), English (fluent), Spanish (basic)

8.2 Education

Oct 2012 - *Current*	**PhD student in Zürich evolutionary biology PhD program.** *Individual-level causes and population-level consequences of variation in fitness in an alpine rodent.* Under the supervision of Dr Erik Postma.
Courses:	
May 27-28th 2013	NGS for Model and Non-Model Species, by K. Shimizu & al.
Jun 22-29th 2013	Evolutionary Biology Workshop in Guarda, by D. Ebert & S. Bonhoeffer
Oct 10-11th 2013	Workshop on Integral Projection Models, C. Merow & al.
Oct 14-18th 2013	Evolutionary Demography, by D. Levitis & al.
Jan 13-17th 2014	Bayesian Population Analysis using WinBUGS, by M. Kéry & M. Schaub
Nov 6-7th 2014	Advanced Software Carpentry, by M. D. Robinson & al.
May 18-19th 2015	Advanced NGS, by S. Wider & H. Lischer

Sep 2011 - Jun 2012	**M.Sc. in evolutionary biology and ecology. University Montpellier II, France.**
Research project:	*Neutral processes and biased mitochondrial introgression* at the center for population biology and management (CBGP). Supervised by Drs Raphaël Leblois, Pierre-André Crochet and François Rousset.

Sep 2008 - Sep 2011	**B.Sc Biology: National engineering school in biology and agronomy, Montpellier Supagro, France.** Specializations in biodiversity conservation, ecology, phylogeny, population genetics, GIS.
Research projects:	
Jan-Aug 2011	*Population dynamics of rodents, and agricultural practices* at Chizé Centre for Biological Studies (CEBC), France. Supervised by Drs Bertrand Gauffre and Vincent Bretagnolle.
Sep-Dec 2010	Centre for Ecological and Evolutionary Synthesis in Oslo, Norway: lab work, genetic data analysis and article redaction. Genetic identification, speciation, hybridization and role of gonosomes in flycatchers and sparrows. Supervised by Prof. Glenn-Peter Sætre.
May-Aug 2010	*Meadow birds phenology, conservation and agriculture* at LPO (Birdlife international), in Grenoble, France.

8.3 Seminars

8.3.1 Invited seminars

– Bern, Switzerland, March 9th **2016**.

– *Body mass selection in an alpine rodent: does it fluctuate? does it matter?* Radboud Universiteit Nijmegen, the Netherlands, December 4th **2014**.

- *Variation in fitness: proximal and ultimate causes.* CNRS Brunoy, France. October 28th **2014**.
- *Individual-level causes of variation in fitness in an alpine rodent.* IEU, Zurich, September 29th **2014**.

8.3.2 Contributed seminars

- *The stasis that wasn't: Adaptive evolution goes against phenotypic selection in a wild rodent population* Evolution 2016, Austin, Texas, USA, June 17th-21st **2016**.
- *The stasis that wasn't: Adaptive evolution goes against phenotypic selection in a wild rodent population* Biology16, the Swiss conference on organismic biology, Lausanne, Switzerland, February 11th-12th **2016**.
- *The stasis that wasn't: Adaptive evolution goes against phenotypic selection in a wild rodent population* 3rd Young Natural History scientists Meeting, Paris, France, February 2nd-6th **2016**.
- *Rapid adaptive evolution opposite to phenotypic selection. Or why snow voles get smaller despite selection for larger individuals* European Society for Evolutionary Biology (ESEB) 15th, Lausanne, Switzerland. August 10th-14th **2015**.
- *Evolution outreach through dirtiness* Poster at the ESEB Workshop on Teaching Evolution, Lausanne, Switzerland. August 9th **2015**.
- *Successful by chance? The power of mixed models and neutral simulations for the detection of individual fixed heterogeneity in fitness components* GDR Ecological Statistics meeting, Lyon, France, March 12th-13th **2015**.
- *Why voles do not become beavers: indirect relationships between traits and fitness counteract selection for larger individuals in a snow vole population* Poster at Biology15, the Swiss conference on organismic biology, Dübendorf, Switzerland, February 12th-13th **2015**.
- *Fluctuating selection and genetic gradients on snow vole mass* Wild Animal Models Biennial Meeting, University of St Andrews, U.K. July 21st-25th **2014**
- *Lord of the scree by chance or by merit? Dynamic vs. fixed heterogeneity in an alpine rodent population.* Evolutionary Demography Society (EvoDemoS) first meeting, in Odense, University of South Denmark. October 5th-10th **2013**.
- *Climatic variability, viability selection and demography in an alpine rodent.* European Meeting of PhD Students in Evolutionary Biology (EMPSEB) 19th, at university of Exeter, U.K. September 3rd-7th **2013**.
- *Neutral processes and cyto-nuclear discordant introgression.* Colloquium Petit Pois Déridé, Avignon, France. August 29th **2012**.

8.4 Skills

8.4.1 Scientific

Biology	Evolutionary biology, population and quantitative genetics, demography.
Statistics	Generalized Linear Mixed Models, Bayesian methods, Mark-Recapture analysis.
Mathematics	Linear algebra, analysis, probabilities.

8.4.2 IT

O.S	Microsoft Windows, Linux (Ubuntu), MacOS X
Scripting and programming	R/S4, C/C++, Bash shell, BUGS/JAGS, Matlab, LaTeX Some projects visible on GitHub: *https://github.com/timotheenivalis*
Software	Arlequin, Genemapper, Mega, Sequencher, CLC Workbench, BioEdit, Genepop, Mark, Ucare, ArcGIS,...

8.4.3 Miscelaneous

February 2013	Far from help first aid course (2 days)

8.5 Teaching

Aug 2016	Field course
Mar 2016	10 afternoons of practicals in Population Ecology
Mar 2015	One hour practical in quantitative genetics
Feb 2015	One day introductory course to LaTeX(self-organized)
Jul 2014 - *current*	One Master student
Dec 2013	3 weeks supervision of a Bachelor student project

8.6 Reviewing activity

8.6.1 Journals

Publons merit: 12 (*https://publons.com/a/822275/*)

2015-2016	Evolution
2015-2016	Molecular Ecology
2016	Heredity
2015	Proceedings of the Royal Society B: Biological Sciences
2015	Ecology and Evolution

8.6.2 Conferences

2016	Jury member for biodiversity and conservation at YNHM 16, Paris

8.7 External activities

8.7.1 Scientific outreach

Contributions to:
- Scientific "speed-dating" with the public at ESEB 15 and Biology 16 conferences
- Dans les testicules de Darwin. (7 articles, 2013 - 2015)
  http://danslestesticulesdedarwin.blogspot.ch
- Un pied dans le plat. (1 article, 2012)
  www.unpieddansleplat.fr/menu_gauche/alimentation_sante/laitage_et_cancer.php

8.7.2 Ornithology and naturalism

2012 - *current*	Member of the regional rare bird comity Tarn-Aveyron *http://www.faune-tarn-aveyron.org/index.php?m_id=20025*

8.8 Competitive funding

8.8.1

2016	Travel grant to attend the 3rd Young Natural History Scientists Meeting, Paris, France
2012	PhD fellowship at IEU UZH

8.9 Peer-reviewed publications

ISI citations: 30

- **Bonnet, T.** & Postma, E. **2016**. Successful by chance? The power of mixed models and neutral simulations for the detection of individual fixed heterogeneity in fitness components. *The American Naturalist* 187(1). Recommended by Faculty of 1000 F1000Prime RECOMMENDED
- García-Navas, V., **Bonnet, T.**, Waldvogel, D., Camenisch, G. & Postma, E. **2016**.Consequences of female philopatry for reproductive success and mate choice in an Alpine rodent. *Behavioral Ecology*.
- García-Navas, V., **Bonnet, T.**, Bonal, R. & Postma, E. **2016**. The role of fecundity and sexual selection in the evolution of size and sexual size dimorphism in New World and Old World voles (Rodentia: Arvicolinae). *Oikos* Early view.
- García-Navas, V., **Bonnet, T.**, Waldvogel, D., Wandeler, P., Camenisch, G. & Postma, E. **2015**. Gene flow counteracts the effect of drift in a Swiss population of snow voles fluctuating in size. *Biological Conservation* 191: 168–177.
- **Bonnet, T.**, Crespin, L., Pinot, A., Bruneteau, L., Bretagnolle, V. & Gauffre, B. **2013**. How the common vole copes with modern farming: Insights from a capture-mark-recapture experiment. *Agriculture, Ecosystems & Environment* 177: 21-?27.
- Elgvin, T.O., Hermansen, J.S., Fijarczyk, A., **Bonnet, T.**, Borge, T., Sæther, S. a, Voje, K.L. & Sætre, G.P. **2011**. Hybrid speciation in sparrows II: a role for sex chromosomes? *Molecular Ecology* 20: 3823-?3837.
- **Bonnet, T.**, Slagsvold, P.K. & Sætre, G.P. **2011**. Genetic species identification of a Collared Pied Flycatcher from Norway. *Journal of Ornithology* 152: 1069-?1073.

8.10 Completed manuscripts

- **Bonnet, T.**, Wandeler, P., Camenisch, G. & Postma, E.. Adaptive evolution goes against phenotypic selection in a wild rodent population. *In review in PLoS Biology*.
- van Benthem, K., Bruijning, M., **Bonnet, T.**, Jongejans, E., Postma, E. & Ozgul, A.. Disentangling evolutionary, plastic and demographic processes underlying trait dynamics: A review of four frameworks. *In review in Methods in Ecology and Evolution*.
- **Bonnet, T.**, Leblois, R., Rousset, F. & Crochet, P.A.. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *To be submitted to Evolution*.