Universität
Zürich UZH

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

PhD dissertation

Timothée Bonnet

Contents

1 General introduction

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

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

*Corresponding author: timothee.bonnet@ieu.uzh.ch

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 et al., 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 et al., 2009; Steiner et al., 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 et al., 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 et al. (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 et al., 2010; Orzack et al., 2011), as well as to a compilation of 22 vertebrate populations (Tuljapurkar et al., 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) com-

monly report evidence for fixed heterogeneity (e.g. Cam and Monnat, 2000; Royle, 2008; Chambert et al., 2013; Guillemain et al., 2013; Chambert et al., 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 et al., 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 et al. (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. de Villemereuil et al., 2013; Brooks et al., 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 num-

ber 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 pseudorandom number generator `Mersenne Twister` (Matsumoto and Nishimura, 1998) and a command file procedure following that of `IBDsim` (Leblois et al., 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 et al. (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 et al., 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 et al., 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:

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

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 et al., 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 et al. (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:

$$\rho_{i,t} \sim \mathcal{P}(\mu_\rho); \text{ for second year individuals,} \tag{2.3a}$$

$$\rho_{i,t} \sim \mathcal{P}(\mu_\rho + m(\rho_{i,t-1} - \mu_\rho)); \text{ for older individuals,} \tag{2.3b}$$

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:

$$\phi_{i,t} \sim \mathcal{B}(\mu_\phi + \beta_j); \text{ for juveniles} \tag{2.4a}$$

$$\phi_{i,t} \sim \mathcal{B}(\text{logit}^{-1}(\text{logit}(\mu_\phi) + c(\rho_{i,t-1} - \mu_\rho))); \text{ for adults,} \tag{2.4b}$$

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 (van Noordwijk and de 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_ϕ^2 and σ_ρ^2, 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_ϕ^2, σ_ρ^2, 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 et al. (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 et al., 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 et al., 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_1^2 and χ_0^2 (Self and Liang, 1987). For survival, first a logistic model not allowing for individual-level heterogeneity was fitted:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} \,, \tag{2.5}$$

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:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} + z_{\phi,i} \,; \text{with } z_\phi \sim \mathcal{N}(0, \hat{\sigma}_\phi^{\,2}) \,. \tag{2.6}$$

Model (2.6) estimated the individual-level heterogeneity in survival probability, $\hat{\sigma}_\phi^{\,2}$. Moreover, a LRT comparing model (2.6) to model (2.5) tested for the significance of $\hat{\sigma}_\phi^{\,2}$.

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

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} \,, \tag{2.7}$$

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:

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} + z_{\rho,i} \,; \text{with } z_\rho \sim \mathcal{N}(0, \hat{\sigma}_\rho^{\,2}) \,. \tag{2.8}$$

Model (2.8) estimated the individual-level heterogeneity in reproductive ability, $\hat{\sigma}_\rho^{\,2}$. Moreover, a LRT comparing model (2.7) to model (2.8) 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 (2.7) and (2.8) 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.
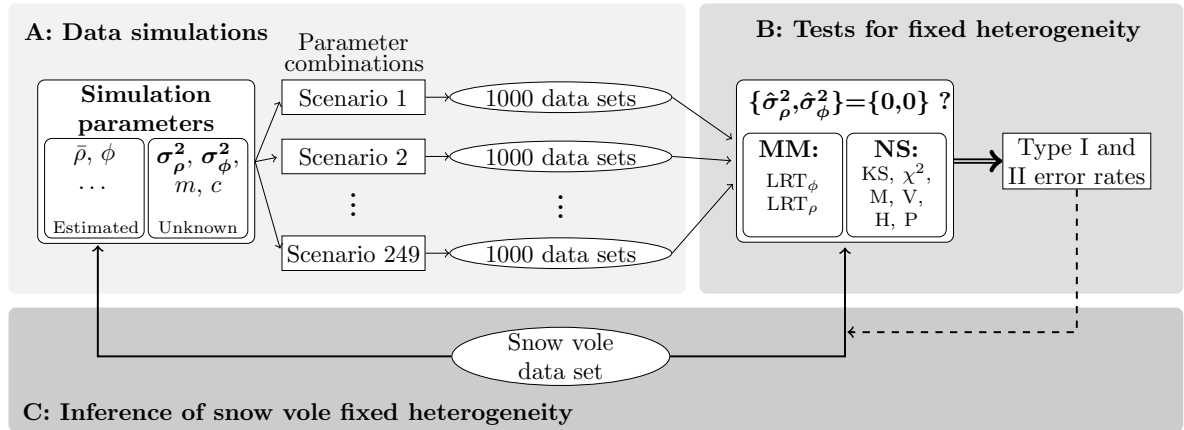

Figure 2.1: Illustration of the simulation and testing process. (A) Data simulation: The simulation model is parametrized using the life cycle and vital rates of a snow vole population, along with additional, unknown, parameters introducing fixed heterogeneity (σ_ϕ^2 and σ_ρ^2) and dynamic heterogeneity (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_ϕ^2 and σ_ρ^2 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 $\{\sigma_\rho^2 = 1, \sigma_\phi^2 = 2, \bar{\rho} = 50\}$) 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_ρ^2 and σ_ϕ^2 (for $\{\sigma_\rho^2 = 0, \sigma_\phi^2 = 0, \bar{\rho}.\}$ $\overline{p\text{-value}} = 0.998$, SE $= 0.001$; for $\{\sigma_\rho^2 = 2, \sigma_\phi^2 = 2, \bar{\rho}.\}$ $\overline{p\text{-value}} = 0.776$, 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_ϕ^2, while the power peaks at intermediate values of simulated σ_ρ^2 and decreases again for higher σ_ρ^2 (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 et al., 2013, chapter 9), as is the case in our short-lived simulated animals.
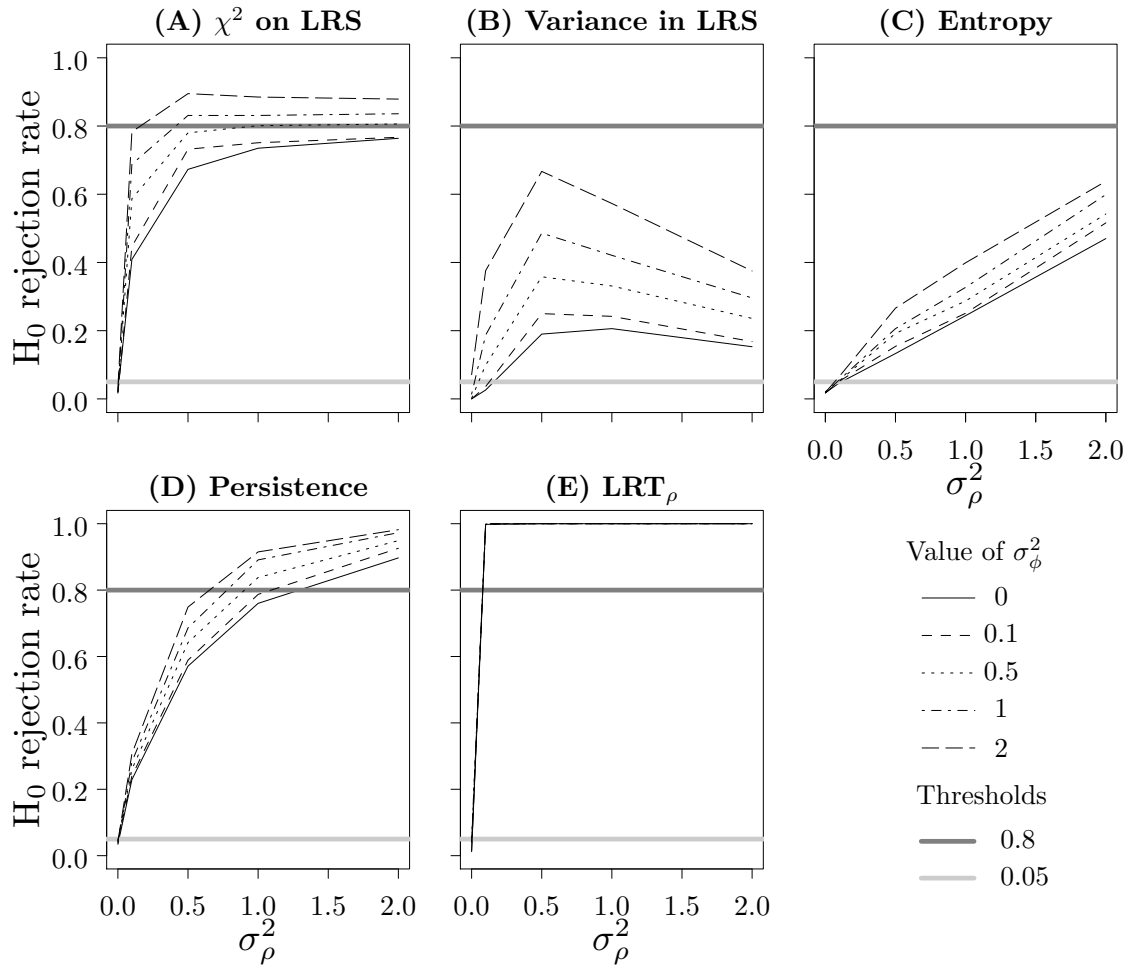

Figure 2.2: Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, as a function of the variance in reproductive propensity, σ_ρ^2, and survival propensity, σ_ϕ^2, when these variances are introduced on the transformed scales. The methods are: (A) a χ^2 test 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 $\sigma_\rho^2 = \sigma_\phi^2 = 0$, the null-hypothesis rejection rates are equal to the type I error rates, which is expected to be 0.05 (light gray line). When $\sigma_\rho^2 \neq 0$ or $\sigma_\phi^2 \neq 0$, 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_ρ, 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.
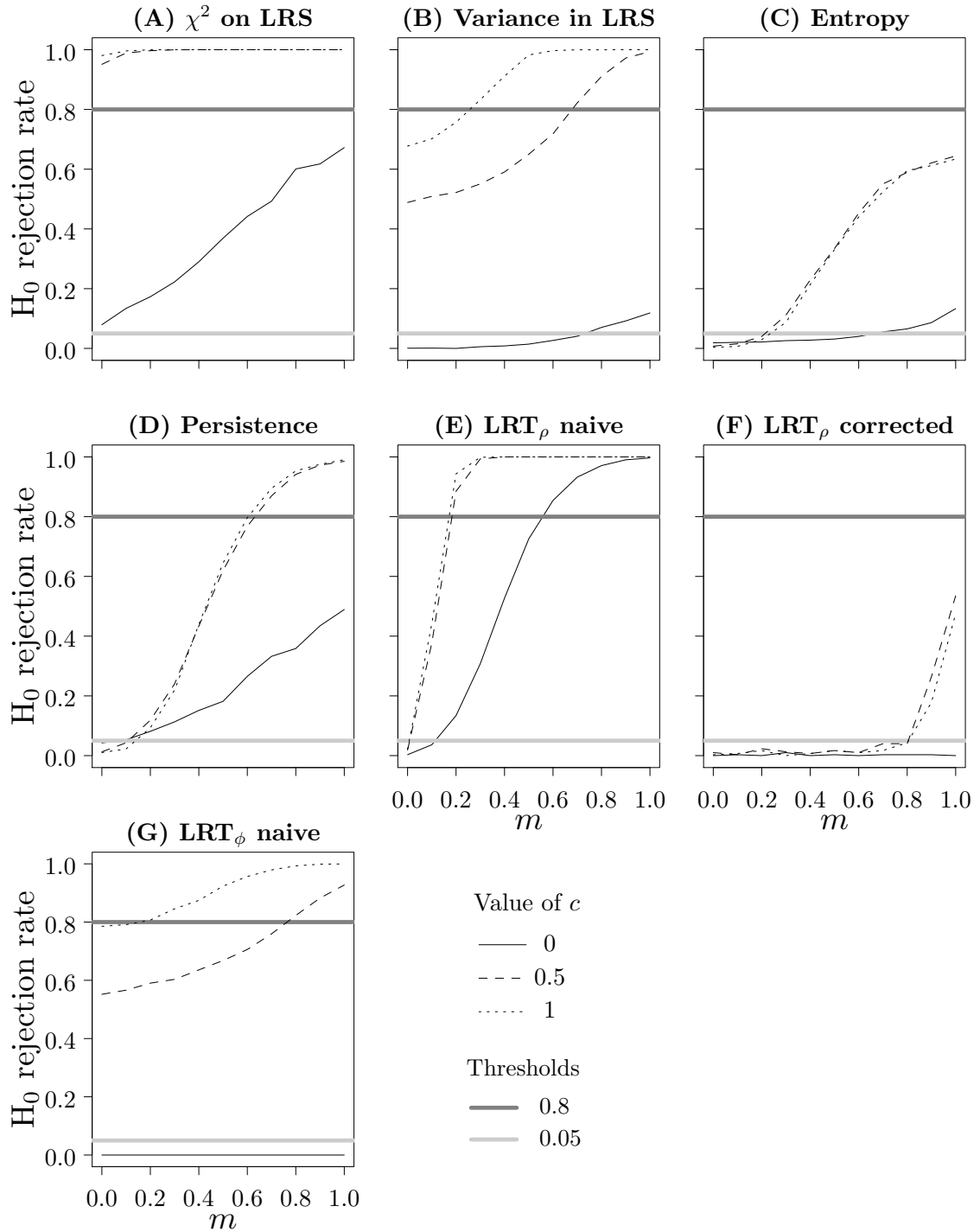
Figure 2.3: Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, when none is explicitly simulated, depending on the parameter m, controlling the structure of transitions between successive annual reproductive successes, and on the parameter c, controlling 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.
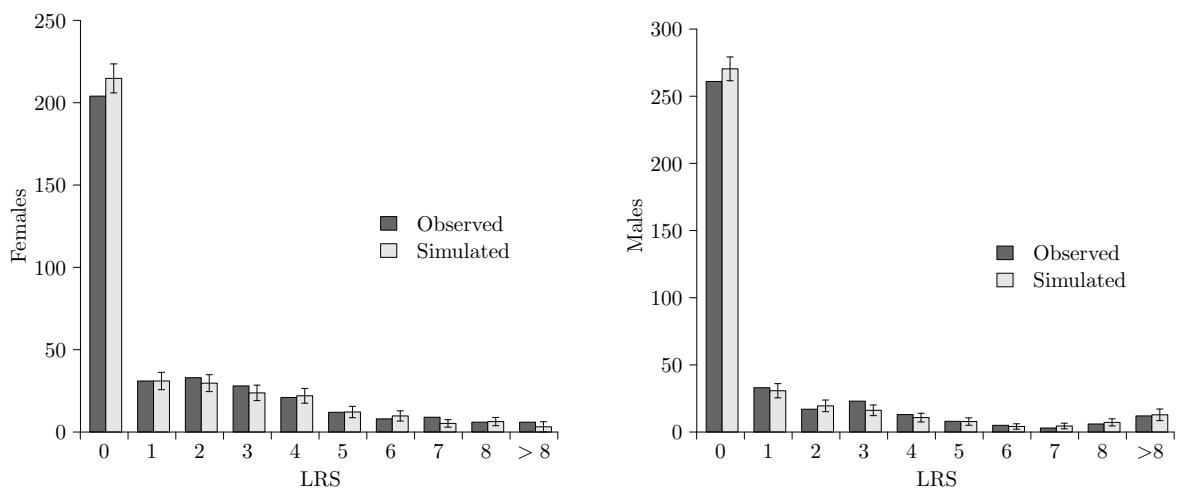

Figure 2.4: Distribution of lifetime reproductive success in the real snow vole data set, observed (dark bars) and simulated through 1000 neutral simulations (light bars with black error bars showing $\pm$ standard deviation), for 2.4.4 females and 2.4.4 males.WRONG

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 (ΔAIC = 87; LRT: p-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 et al., 2009). This finding directly contradicts the claim "[...] that random effect models will always detect unobservable fixed effects" Steiner et al. (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 et al. (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 et al., 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 et al., 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 et al., 2010; Orzack et al., 2011; Tuljapurkar et al., 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 et al., 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 et al., 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$_\phi$ has no statistical power for simulated data sets with simulated $\sigma_\phi^2 \leq 2$, while confidence and credibility intervals indicate that the possible values of σ_ϕ^2 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_ρ^2, 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 et al., 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

Bibliography

Albert, A., and J. Anderson. 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1–10.

Atkins, D. C., S. A. Baldwin, C. Zheng, R. J. Gallop, and C. Neighbors. 2013. A tutorial on count regression and zero-altered count models for longitudinal substance use data. Psychology of Addictive Behaviors 27:166–77.

Barnett, A. G., N. Koper, A. J. Dobson, F. Schmiegelow, and M. Manseau. 2010. Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology. Methods in Ecology and Evolution 1:15–24.

Bates, D., M. Maechler, B. Bolker, and S. 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 27:330–6.

Brandon, R., and J. Beatty. 1984. The propensity interpretation of ' fitness '. No interpretation is no substitute. Philosophy of Science 51:342–347.

Brooks, M. E., M. W. McCoy, and B. M. Bolker. 2013. A method for detecting positive growth autocorrelation without marking individuals. PloS One 8: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, E., O. Gimenez, R. Alpizar-Jara, L. M. Aubry, M. Authier, E. G. Cooch, D. N. Koons, W. A. Link, J.-Y. Monnat, J. D. Nichols, J. J. Rotella, J. A. Royle, and R. Pradel. 2013. Looking for a needle in a haystack: inference about individual fitness components in a heterogeneous population. Oikos 122:739–753.

Cam, E., W. a. Link, E. G. Cooch, J.-Y. Monnat, and E. Danchin. 2002. Individual covariation in life-history traits: seeing the trees despite the forest. The American Naturalist 159:96–105.

Cam, E., and J. Y. Monnat. 2000. Stratification based on reproductive state reveals contrasting patterns of age-related variation in demographic parameters in the kittiwake. Oikos 90:560–574.

Caswell, H. 2011. Beyond r_0: Demographic models for variability of lifetime reproductive output. PLoS One 6:e20809.

Chambert, T., J. J. Rotella, and M. D. Higgs. 2014. Use of posterior predictive checks as an inferential tool for investigating individual heterogeneity in animal population vital rates. Ecology and Evolution 4:1389–97.

Chambert, T., J. J. Rotella, M. D. Higgs, and R. A. Garrott. 2013. Individual heterogeneity in reproductive rates and cost of reproduction in a long-lived vertebrate. Ecology and Evolution 3:2047–60.

Charlesworth, B. 2015. Causes of natural variation in fitness : Evidence from studies of Drosophila populations. Proceedings of the National Academy of Sciences 112:1662–1669.

Clutton-Brock, T. H. 1988. Reproductive success in male and female red deer. Pages 325–343 *in* T. H. Clutton-Brock, ed. Reproductive success. Studies of individual variation in contrasting breeding systems. University of Chicago Press, Chicago and London.

Clutton-Brock, T. H., and B. 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 25:562–573.

Crainiceanu, C. M., and D. Ruppert. 2004. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66:165–185.

de Villemereuil, P., O. Gimenez, and B. 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 4:260–275.

Ellegren, H., and B. C. Sheldon. 2008. Genetic basis of fitness differences in natural populations. Nature 452:169–175.

Fisher, R. 1958. The genetical theory of natural selection. 2nd ed. Dover Publications, New York.

Fox, G. A., and B. E. Kendall. 2002. Demographic stochasiticy and the variance reduction effect. Ecology 83:1928–1934.

Guillemain, M., A. J. Green, G. Simon, and M. Gauthier-Clerc. 2013. Individual quality persists between years: individuals retain body condition from one winter to the next in Teal. Journal of Ornithology 154:1007–1018.

Hadfield, J., D. Richardson, and T. Burke. 2006. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a bayesian framework. Molecular Ecology 15:3684.

Hadfield, J. D. 2010. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software 33:1–22.

Hadfield, J. D., A. J. Wilson, and L. E. B. Kruuk. 2011. Cryptic evolution: does environmental deterioration have a genetic basis? Genetics 187:1099–113.

Hansen, T. F., C. Pélabon, and D. 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, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. Applied logistic regression, third edition. Wiley, Hoboken, New Jersey.

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

Keller, L., and D. Waller. 2002. Inbreeding effects in wild populations. Trends in Ecology & Evolution 17:19–23.

Kendall, B. E., G. A. Fox, M. Fujiwara, and T. M. Nogeire. 2011. Demographic heterogeneity, cohort selection, and population growth. Ecology 92:1985–1993.

Krimbas, C. B. 2004. On fitness. Biology & Philosophy 19:185–203.

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, T., J. Slate, L. E. Kruuk, and J. 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 8:3–30.

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

Mousseau, T. A., and D. A. Roff. 1987. Natural selection and the heritability of fitness components. Heredity 59:181–197.

Orzack, S. H., U. K. Steiner, S. Tuljapurkar, and P. Thompson. 2011. Static and dynamic expression of life history traits in the northern fulmar *Fulmarus glacialis*. Oikos 120:369–380.

Pinheiro, J. C., and D. 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. Gaillard. 2012. Modeling reproductive trajectories of roe deer females: Fixed or dynamic heterogeneity? Theoretical Population Biology 82:317–328.

Postma, E. 2014. Four decades of estimating heritabilities in wild vertebrate populations: improved methods, more data, better estimates? *In* A. Charmentier, D. Garant, and L. E. B. Kruuk, eds., Quantitative genetics in the wild. Oxford University Press, Oxford.

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

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

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

Rotella, J. J. 2008. Estimating reproductive costs with multi-state mark-recapture models, multiple observable states, and temporary emigration. Pages 157–172 *in* D. L. Thomson, E. G. Cooch, and M. J. Conroy, eds. Modeling Demographic Processes In Marked Populations. Series Environmental and Ecological Statistics, Vol 3. Springer.

Royle, J. A. 2008. Modeling individual effects in the Cormack-Jolly-Seber model: a state-space formulation. Biometrics 64:364–70.

Schauber, E. M., B. J. Goodwin, C. G. Jones, and R. S. Ostfeld. 2007. Spatial selection and inheritance: applying evolutionary concepts to population dynamics in heterogeneous space. Ecology 88:1112–1118.

Self, S. 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 82:605–610.

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

Steiner, U. K., and S. Tuljapurkar. 2012. Neutral theory for life histories and individual variability in fitness components. Proceedings of the National Academy of Sciences 109:4684–4689.

Steiner, U. K., S. Tuljapurkar, and S. H. Orzack. 2010. Dynamic heterogeneity and life history variability in the kittiwake. Journal of Animal Ecology 79:436–44.

Tuljapurkar, S., U. K. Steiner, and S. H. Orzack. 2009. Dynamic heterogeneity in life histories. Ecology Letters 12:93–106.

Turner, B. M. 2009. Epigenetic responses to environmental change and their evolutionary implications. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364:3403–18.

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

Vaupel, J., K. Manton, and E. Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16:439–454.

Vaupel, J. W. 1988. Inherited frailty and longevity. Demography 25:277–287.

Wandeler, P., and G. Camenisch. 2011. Identifying Y-chromosomal diversity by long-template PCR. Molecular Ecology Resources 11:835–841.

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

Wilberg, M. J., and J. R. Bence. 2008. Performance of deviance information criterion model selection in statistical catch-at-age analysis. Fisheries Research 93:212–221.

Wolf, J. B., and M. J. Wade. 2009. What are maternal effects (and what are they not)? Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364:1107–15.

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:

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \text{Age}_{i,t} \text{ ; using a binomial error structure} \tag{2.9a}$$

$$\text{logit}(\phi_{i,t}) = \mu_\phi + \rho_{i,t} \text{ ; using a binomial error structure} \tag{2.9b}$$

$$\log(\rho_{i,t}) = \mu_\rho + \rho_{i,t-1} \text{ ; using a quasi-Poisson error structure} \tag{2.9c}$$

$$\log(\rho_{i,t}) = \mu_\rho + \text{Age}_{i,t} \text{ ; using a quasi-Poisson error structure} \tag{2.9d}$$

These were used to check that survival depended on age (2.9a), that survival depended on annual reproductive success only when that was required (2.9b), that ARS depended on previous reproductive attempts only when fixed heterogeneity for reproductive success or Markovian reproduction was simulated (2.9c) and that ARS of adults was not age-dependent (2.9d). 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 et al. (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 (Δpower$_{1000-100}$=-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 (Δpower$_{100,000-1000}$=-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_ρ^2 and σ_ϕ^2, 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.
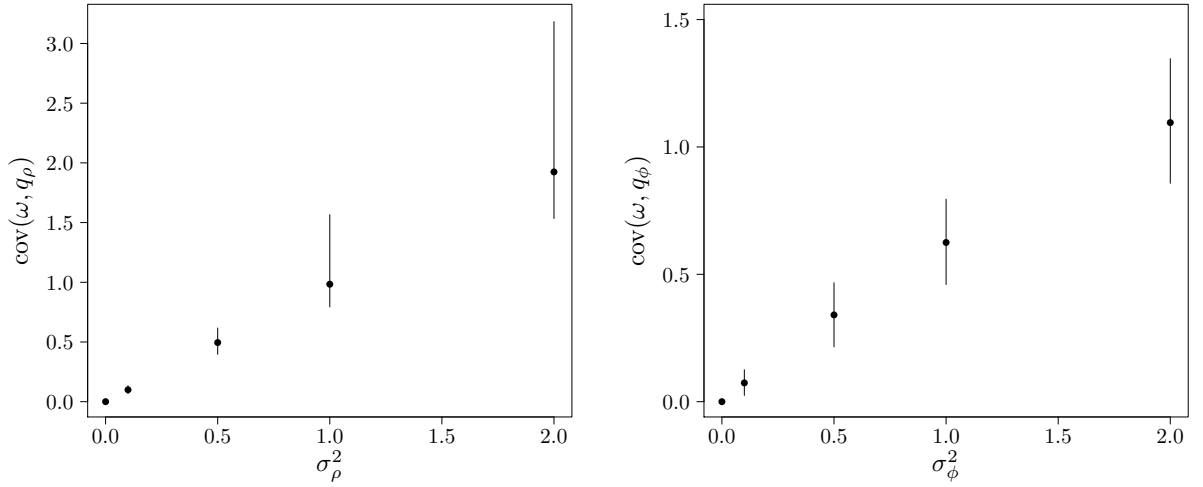
Figure 2.5: Appendix C Strength of selection on individual fixed qualities for survival and reproduction, as a function of the expected variance in these qualities. Strength of selection was measured as the individual-level covariance between the qualities and a proxy for fitness (ω): relative lifetime 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

$$\rho_{i,t} \sim \mathcal{P}\left(\mu_\rho + q_{\rho,i}\right) \tag{2.10a}$$

$$\text{and } \phi_{i,t} \sim \mathcal{B}\left(\mu_\phi + \beta_{age} + q_{\phi,i}\right). \tag{2.10b}$$

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 ($\bar{\rho}$)

$\overline{\text{ARS}}$	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 $\bar{\rho}$ and variance σ_ρ^2.
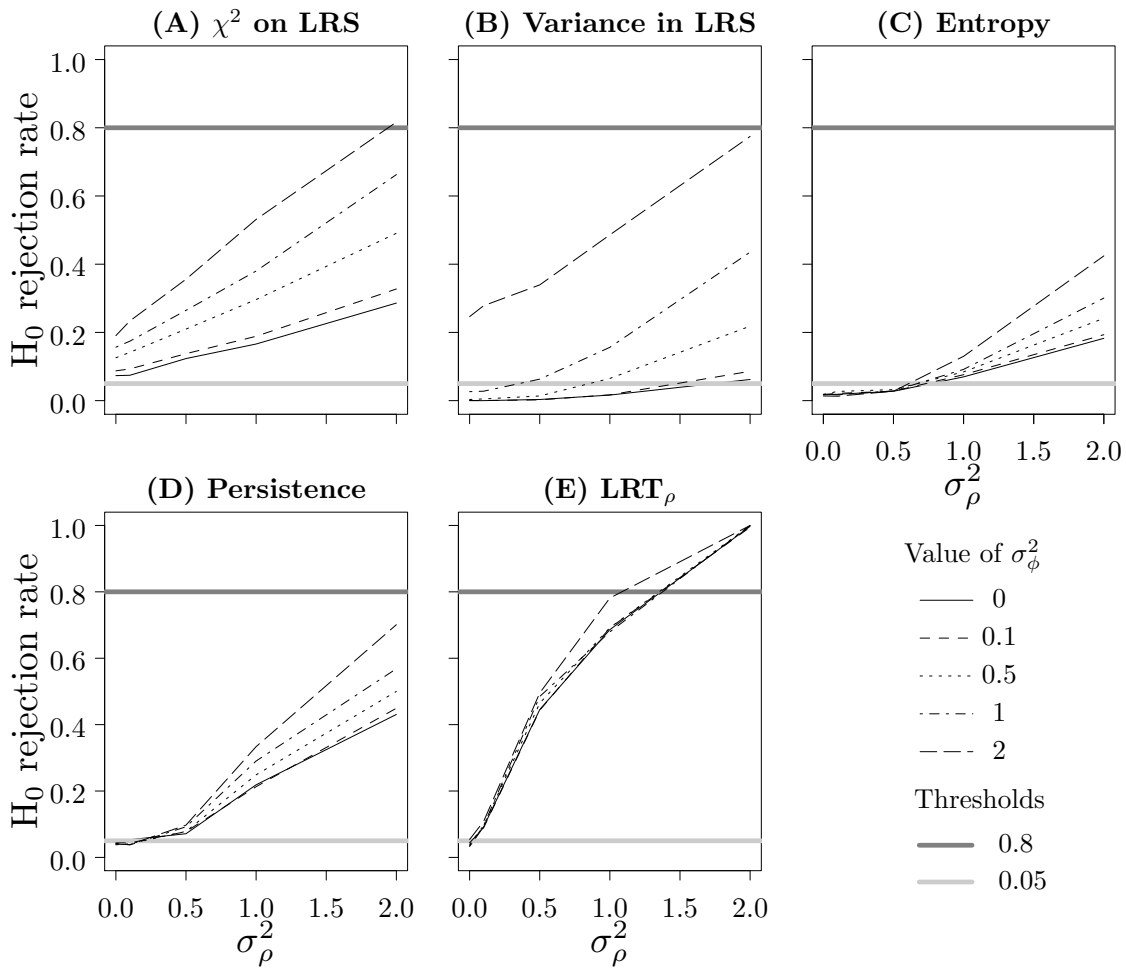

Figure 2.6: Appendix D Null-hypothesis rejection rates for various methods testing for the presence of fixed heterogeneity, depending on the variance in reproductive propensity, σ_ρ^2, and on the variance in survival propensity, σ_ϕ^2, when these variances are introduced on the original scales. The methods 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 $\sigma_\rho^2 = \sigma_\phi^2 = 0$, the null-hypothesis rejection rates are equal to the type I error rates, which is expected to be 0.05 (light gray line). When $\sigma_\rho^2 \neq 0$ or $\sigma_\phi^2 \neq 0$, 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 et al., 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 et al., 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 et al., 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

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

5 Fluctuating selection

6 Acknowledgments

Looking back to one's life, it is amazing to noticed how much the path it took has been influenced by a myriad of people. These acknowledgments are about a PhD thesis, but it is not clear how far they should go backward in time.

Erik Postma

Glauco Camenisch

Lukas Keller

Arpat Ozgul

Barbara Tschirren

Jarrod Hadfield

Marc Kéry

Pirmin Nietlisbach PhilippXXX Becher Judith Bachmann

Vanja Michel, Rien van XXX, JohanXXX

Koen van Benthem Marjolein BXXX Cindy Canale

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 for my master thesis). Still, I owe them a lot, and the TRACE PERSISTANTE DE NOS RENCONTRE shape my PhD. In 2011, while I was trying to 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 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 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 serious, for introducing me to Richard Dawkins' writings and for pushing me toward 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++` low-level programming. Coding `ForwardBackward` and its improved sequels was a crazy adventure, full of EPREUVES (Oh gosh, these bloody probabilities of identity in small sexual populations!) but mainly wonders. Thanks to François Rousset for LAISSER ENTREVOIR UN MONDE DE RIGOUREUSE COMPREHENSION DU VIVANT. 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 scientific conversations account for half of my knowledge in evolutionary biology. Here start acknowledgments that go either beyond the scientific life. Muses who pushed my scientific interest and helped me working hard (often by actually preventing me from working) are Gaëlle Duranthon, Manon Ghislain, Léo Grasset, Joane Elleouet (yes you are both acknowledged twice), Chelsea J. Little ...

Thanks to my two adoptive families who kept me cheerful with meals, games, music, remote places, short nights and craziness during the last three years. 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 walks and rides across the forests, meadows and moorlandsXXX of Lacaune mountains, by the contemplation of glittering and diverse carabids, by the excitement of long days monitoring bird migration, by dazzlingXXX conversations on conservation and nature management and by the constant realization that we still know so little and that it takes only some sweat and patience to discover new things. This would have not been sustainable without the naturalists who I taught me so much and who 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 DenoziÃÍre, Aurélien Salesse, Denis Guillaumin and Manon Ghislain.

Some high school teachers are 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 acknowledgment for doing that at diverse degrees, but here are a few whose classes remain very alive: Benoît Leviandier, Mdm Rolland, XXX math terminale, 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.