IMPROVING RISK ASSESSMENT OF TRANSGENE INVASION: ASSESSING THE ROLE OF UNCERTAINTY IN PREDICTIONS

By

Ashok Ragavendran

A DISSERTATION

Submitted to
Michigan State University
in partial fullfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Fisheries and Wildlife

2009

Genetic
ronment
the asso
genic or
high dis
environm
gression
models o
uncertain
tions and
mitigating

Chapte
graphic st
existing m
approach.
of the tran
mographic
allows for i
predicts, ev

In Chap
in the estima
iments, usin

ABSTRACT

IMPROVING RISK ASSESSMENT OF TRANSGENE INVASION: ASSESSING THE ROLE OF UNCERTAINTY IN PREDICTIONS

By

Ashok Ragavendran

Genetically Engineered organisms (GEOs) have the potential to pose a threat to the environment and it is imperative to develop scientifically sound methodologies for assessing the associated ecological risks. The main emphasis of ecological assessment is for transgenic organisms that have wild relatives and are relatively difficult to contain or have high dispersal capabilities (e.g., plants, insects and aquatic organisms). Key to managing environmental threats of GEOs lies in the evaluation of the possibility of transgene introgression into natural populations and this is currently accomplished by using deterministic models of varying degrees of statistical and computational complexity. Accounting for uncertainty in these assessments is vital for quantifying risks to make realistic predictions and also to gain insight for guiding further research and effective policy towards mitigating potential hazards.

Chapter 1 of this dissertation addresses the issue of quantifying the effects of demographic stochasticity on predictions for transgene introgression. This work generalizes existing methodology by incorporating demographic stochasticity using a Monte Carlo approach. Results show that considerable variation arises in the prediction distributions of the transgene frequency and extinction probabilities, and also that the effects of demographic stochasticity vary among fitness components. Most importantly, stochasticity allows for increased persistence of the transgene beyond what the deterministic model predicts, even for cases where the transgene would have been lost early.

In Chapter 2, another source of variability is addressed, which arises from uncertainty in the estimated parameters on predictions. We estimate fitness components from experiments, using genetically modified Zebrafish (*Danio rerio*), to be used as inputs into the

rat fitne

tness d

tion b

Results

can gen

among

in mes

assumpt

In Chap

fitness

tions. R

overly c

differ fr

termina

Bas

formali

burden.

of para

was em

computa

BGP m

of a Bay

model e

will faci

net fitness component model. Applying bootstrap approaches we show that estimates of fitness components for the wildtype and transgenic genotypes present considerable variation, both in absolute and relative values, leading to extra variation in model predictions. Results also show that even moderate variation in estimates of individual components can generate large effects at the population level due to non-linearity and the interactions among genotypes. The model predictions showed good agreement with results observed in mesocosm experiments, indicating that realistic prediction in the case of GEO risk assessment can be made using carefully constructed experiments and modeling approaches. In Chapter 2 we also present a Copula methodology to incorporate dependencies among fitness components due to life-history trade-offs, to assess their effects on model predictions. Results show that assuming independence among fitness components can produce overly conservative predictions. Further, correlations among absolute fitness components differ from that among the relative values of fitness components, which are the final determinants of transgene fitness.

Based on the simulations for the first two research chapters, there arose a clear need to formalize approaches for exploring the parameter space while reducing the computational burden. In Chapter 3, a global sensitivity analysis was considered to examine the effect of parameters on the model. Meta-modeling using a Bayesian Gaussian Process (BGP) was employed to improve the efficiency of sensitivity analysis and thus reduce the overall computational burden without sacrificing model complexity. The predictions from the BGP model are shown to provide satisfactory performance as an emulator. The choice of a Bayesian approach is deliberate as this framework is flexible enough to incorporate model extensions as well combine outputs across multiple models. Moreover, this choice will facilitate integration with other areas in future.

This th

This thesis is dedicated to my mother Bhavani Raghavendran and my sister Asha

First and
brother, wh
am in this p

My ter
Animal Sc
would like
level.

First, I
for his supp
thank my c
Dr. Rob T
it. I can't
encouragem
thank Dr T
just for the
through thi
by providin

In the F
support and
the current
for providin
wonderful s

In the A
and encoura
especially th

ACKNOWLEDGEMENTS

First and foremost I would like to acknowledge my family, my mother, sister and brother, who have made a lot of sacrifices during my entire education because of which I am in this position today. I thank you for your faith and belief in me.

My term at MSU involved being part of two departments, Fisheries and Wildlife and Animal Science. I have had many memorable interactions in both departments and I would like to thank everybody who made a difference, both, on a academic and personal level.

First, I want to thank my advisor, Dr. Guilherme J.M. Rosa. I am deeply grateful for his support, encouragement, motivation and his commitment to me throughout. I also thank my committee members Dr Jim Bence, Dr. Bryan Epperson, Dr. Kim Scribner, Dr. Rob Tempelman and Dr. Scott Winterstein, a long list and I have never regretted it. I can't thank them enough for coming through time and time again with support, encouragement and their willingness to discuss ideas anytime. I would like to especially thank Dr Tempelman for having given me the wonderful opportunity for consulting, not just for the financial support it afforded, but for letting me strengthen my statistical skills through this avanue. Dr Winterstein also contributed substantially to my financial support by providing teaching assistanships and summer funds and I thank him for his support.

In the FW department, I want to especially thank Dr. Bill Taylor, the past chair, for support and encouragement when Guilherme moved to Wisconsin and Dr. Mike Jones, the current chair, for support in the last year of my program. I also thank the FW dept for providing a wonderful environment during my term here and a special thanks to the wonderful support staff, especially Jill Cruth and Mary Witchell.

In the Animal Science department, I want to thank Dr. Cathy Ernst for her support and encouragement in the last year and everybody from the PRP project. I would like to especially thank Dr Juan Pedro Steibel who has been a constant figure in discussion and help, first a

from him.

Drs Ron B

I also g

ment Proje

funded the

Finally

my life wh

and Sudee

and for tha

to Chada R

want to ac

especially

to thank m

for the ma

help, first as my lab mate and later as a mentor in the PRP project, and I have learnt a lot from him. I would like to thank people from the functional genomics group, especially Drs Ron Bates,Ted Ferris and Paul Coussens.

I also gratefully acknowledge the support from my collaborators on the Risk Assessment Project, Dr. Bill Muir and Dr. Rick Howard, and the USDA-BRAG for having funded the project.

Finally, last but not least, I want to acknowledge on a personal note all the people in my life who have contributed to this endeavour. I would like to thank Punidan Jeyasingh and Sudeep Perumbakkam, my friends who have always been there for me throughout and for the many stimulating intellectual discussions we have had. I am deeply grateful to Chida Ramanathan and his family who have been a second family for me here. I also want to acknowledge my friends from Chennai, the “ Beach ” gang and the “Pondy” gang, especially Nagarajan, Arun, Srikrishna and Kate, my love, for her support. Lastly, I want to thank my friends/collaborators at MSU, Jeff Evans, Pat Forsythe and Sean MacEachern, for the many good times here, academic and otherwise.

TABLE OF CONTENTS

1 Sta
qu
the
fr

2 Sc
st
lev

.3 Pa
to
wa
m
p
an

.1 Sa
ra
T
fo

.2 S
cr
na
th
d
o
re

.3 S
da
in
sc

LIST OF TABLES

LIST OF FIGURES

2.1

2.2

2.3

2.4

2.5 C

g

t

i

c

s

T

er

2.6

3.1

3.2

3.3

3.4 S
T

3.5 S
I

3.6 F
c

A.22 C

for the

scenar

A.23 C

for the

scenar

Introduction

Current rates of glob

and habitats have re

2003), leading to rap

quent ecological imp

these impacts ecolog

especially in the co

systems. Ecologica

stressors and their a

sion making and pr

Recently, there has

sideration in ecolo

importantly integra

(Nacci and Hoffma

fundamental comp

house and Sorenso

use in ecological

complexity (Barnt

technology have a

tool for ecologica

system where a co

Introduction

Current rates of global environmental change due to anthropogenic modification of natural habitats have resulted in accelerated extinction of natural populations (Gaggiotti, 2003), leading to rapid changes in biotic interactions (Kareiva et al., 1993) with consequent ecological impacts. To formulate effective policy and regulations that can mitigate these impacts ecological risk assessments have risen to prominence in the last decade, especially in the context of new technologies that can potentially further stress natural systems. Ecological risk assessment focuses on the relationships among one or more stressors and their associated ecological effects and serves as a guide for regulatory decision making and provides insights into the possible exposures and concomitant hazards. Recently, there has been an emphasis on making population-level risk an explicit consideration in ecological risk assessments (Barnthouse and Sorenson, 2007), and more importantly integrating genetics and population dynamics into model based approaches (Nacci and Hoffman, 2008). Modern ecological risk assessments include modeling as a fundamental component to predict population trajectories and to quantify risks (Barnthouse and Sorenson, 2007). Many types of population models have been developed for use in ecological risk assessment with varying degrees of statistical and computational complexity (Barnthouse and Sorenson, 2007). In the past decade, advances in computing technology have allowed for the emergence of complex simulation models as a primary tool for ecological assessments. Simulations are recognized as a kind of experimental system where a complex model formulation mimics the behavior of natural systems. This can help re

sible otherv

Advanc

worldwide

ally friend

than 30 spe

nology is a

tion. Pilson

as transgen

primarily d

is for trans

plants, inse

tect biolog

for hybridi

Research C

neered org

towards po

ies under a

without the

2006; Tied

crops recog

wherein po

Nacci and

Key to

of transgen

Hedrick, 2

on the fitne

can help researchers investigate system behavior by manipulations that would be impossible otherwise, either logistically or ethically (Peck, 2004).

Advances in the biotechnology realm combined with escalating demand for food worldwide has fostered the growth of transgenic solutions for generating environmentally friendly produce (Aerni, 2004) and is especially true in aquaculture wherein more than 30 species are awaiting commercial release (Devlin et al., 2006). Transgenic technology is also gaining ground in livestock (Clark and Whitelaw, 2003) and crop production (Pilson and Prendeville, 2004), but has not been widely received by consumers and no transgenic animal has been approved for food consumption in the USA or elsewhere primarily due to environmental concerns. The main emphasis of ecological assessment is for transgenic organisms that have wild relatives and high dispersal capabilities (e.g., plants, insects and aquatic organisms; National Research Council, 2004). In this context biological containment strategies will have to be employed if there exists potential for hybridization with wild populations leading to introgression or speciation (National Research Council, 2004). Clearly, potential environmental impacts of genetically engineered organisms (GEOs) should be evaluated and regulated using a scientific approach towards policy, wherefore research and assessment of impacts should capture the dynamics under a spatio-temporal context relative to a comparator, usually the same organism without the transgene (see reviews in Andow and Hilbeck, 2004; Andow and Zwahlen, 2006; Tiedje et al., 1989)). Scientific approaches for assessing the risks of transgenic crops recognize the need to utilize an ecological basis (e.g. Andow and Hilbeck, 2004) wherein population-level strategies employ the integration of genetics and demography (Nacci and Hoffman, 2008).

Key to managing environmental threats of GEOs lies in evaluating the possibility of transgene introgression into natural populations and its resultant ecological impact (Hedrick, 2001). The likelihood of invasion and the persistence of a transgene depends on the fitness of the transgenic type relative to that of wildtype genotypes (Tiedje et al., 1999). The problem

the transgene into w

the population, and

sion and therefore e

the environmental

organization: for e

altering biotic inte

ecosystem dynami

Current meth

organisms are ba

els (Knibb, 199

tion models (Cla

2002), and spati

2003; Richter a

are mainly dev

bridization with

The last decade

predicting the

introgression for

incorporating

sal (Meagher,

tion dynamics

population m

variability (C

cently the A

population st

sequence of

1989). The problem consists of two aspects: (a) prediction regarding the introgression of the transgene into wild populations and the associated changes in genetic composition of the population, and (b) predicting changes in ecological interactions following introgression and therefore ecosystem effects from changes in the wildtype population. Assessing the environmental effects of GEOs requires information at multiple levels of biological organization; for example, transgenic individuals may induce population-level effects by altering biotic interactions (Devlin et al., 2004) which is shown to impact community and ecosystem dynamics (Kareiva et al., 1993).

Current methodology for assessing the ecological risks and impacts of transgenic organisms are based on a plethora of modeling frameworks ranging from verbal models (Knibb, 1997), reaction-diffusion methods (Soboleva et al., 2003), matrix population models (Claessen et al., 2005; Garnier and Lecomte, 2006; Muir and Howard, 2001, 2002), and spatial transfer models coupling population dynamics and genetics (Meagher, 2003; Richter and Seppelt, 2004). A recurrent theme amongst these models is that they are mainly developed for assessing the risk of transgenic crops and focus more on hybridization with wild relatives of the modified crop species (Hails and Morley, 2005). The last decade has also seen increased complexity in the modeling frameworks towards predicting the fate of the transgene introgression into natural populations. Models of introgression for transgenic crops evaluate gene flow based on pollen dispersal, for example, incorporating exponential distance and directional effects of wind mediated pollen dispersal (Meagher, 2003) or transport equations of atmospheric physics coupled with population dynamics and genetic models (Richter and Seppelt, 2004). Additionally, structured population models combining dispersal (Garnier and Lecomte, 2006) and environmental variability (Claessen et al., 2005) have also been employed in this context. More recently the AMELIE modeling framework proposed for transgenic crops combines both population structure, stochasticity and dispersal (Kuparinen and Schurr, 2007). The consequence of increased model complexity are twofold, the number of parameters in the model increase

in the proposed

mated model y

Aquacultur

awaiting field

frameworks in

the wild (e.g.

create real pot

the communit

2002 propose

Caswell, 198

age-structured

ability, adult

male fertility,

in linking ind

evolutionary

widely empl

conservation

work to evalu

advocated as

A charac

from chance

survival and

ation within

in vital rates

studies of na

both environ

model increases and the relationships among variables need not be linear. For example, in the proposed AMELIE framework there are 29 parameters and this is the most sophisticated model yet for risk assessment of transgenic crops.

Aquaculture is another area wherein there is a large and increasing diversity of GEOs awaiting field release (Devlin et al., 2006) and there is a relative paucity of modeling frameworks in this area. Most species used in aquaculture have natural populations in the wild (e.g. salmon) and therefore, in conjunction with their dispersal characteristics, create real potential for the transgenes to invade natural systems and cause perturbation at the community and ecosystem levels. For the case of GE fish, Muir and Howard (2001; 2002) proposed a generic modeling framework based on matrix population models (sensu. Caswell, 1989) that incorporates genetics and demography. Their model is based on an age-structured population model characterized by six net fitness components (juvenile viability, adult viability, mating advantage, age at sexual maturity, female fecundity and male fertility, and reproductive longevity). The utility of this modeling framework lies in linking individual life history traits to population dynamics to provide insight into the evolutionary fate of populations (Grant and Benton, 2000). Matrix population models are widely employed in applied ecology to predict the fate of populations of economic and conservation concern (Tuljapurkar, 1997). This approach provides a reasonable framework to evaluate the probability of gene introgression in natural populations and has been advocated as a tool for assessing risks of GE crops (Hails and Morley, 2005).

A characteristic feature of biological populations is the inherent variability arising from chance events that individuals within the population experience influencing their survival and reproduction. Sources of demographic variation can be divided into variation within individuals with the same vital rates (sensu. Pollard (1966)) and variation in vital rates among individuals (sensu. Fox and Kendall (2002)). Recent demographic studies of natural populations have underscored the importance of the role of variation, both environmental and demographic, and is shown to be imperative in understanding the influence of a

example, it ha

demographic

rate Tuljapu

limited m de

population m

uncertainty i

One involves

forecast and

their individ

2006). Many

the projectio

(Grant and B

conservation

this context

uations with

impossible

It is als

tant, especia

focused pri

1990; acco

to decrease

has long re

evolutionar

have show

can modify

Population

influence of anthropogenic modifications on natural populations (Boyce et al., 2006). For example, it has been shown that variation in vital rates can signicantly affect outcomes of demographic models with important ramifications, such as estimates of population growth rate (Tuljapurkar et al., 2003) and persistence (Engen et al., 2005b). Uncertainty in estimated model parameters is another source of variation for model predictions in using population models to predict population fate. In the context of matrix population models, uncertainty in vital rates has been implemented using two broad schemes of resampling. One involves resampling of the projection matrices over time for a single trajectory of the forecast and the other involves sampling a random set of projection matrices, each having their individual trajectories (e.g. Life-stage Simulation Analysis (LSA), Wisdom et al., 2000). Many methods have been developed to elicit the effect of individual parameters of the projection matrix on the overall growth rate of populations, such as elasticity analysis (Grant and Benton, 2000; Caswell, 1989), and utilized in forecasting for management and conservation (Morris and Doak, 2002). Monte Carlo approaches are specifically used in this context to evaluate the effect of uncertainty in parameter estimates, especially in situations with limited field data or when analytical derivations of the model dynamics are impossible or cumbersome (Tuljapurkar et al., 2003; Caswell, 1989).

It is also widely recognized that dependencies among the vital rates are also important, especially for forecasting. Long standing considerations of these dependencies have focused primarily on temporal correlations among the projection matrices (Tuljapurkar, 1990) accounting for environmentally driven correlation in vital rates, which is shown to decrease extinction times (Inchausti and Halley, 2003). However, life history theory has long recognized the importance of associations among vital rates in determining the evolutionary fates of different genotypes in populations (Stearns, 1992). Recent studies have shown that incorporating associations in the vital rates based on life-history theory can modify evolutionary stable states (Koons et al., 2008) as well as modify estimates of population growth rates (Ramula and Lehtila, 2005). Most approaches utilize an ad-hoc

method of el
2005; Wisd

Using a m
in extracting
model output
the contributi
tance, also kn
analysis), both
review see Sa
of population
and thus are n
for population
stage-structure
is linear and n
variance-based
recently becon
complex mode
niques can be f
approach to eco

Previously,
sion have utiliz
largely ignoring
in parameter est
ity. Claessen et a
persistence of fe
transgene introgr
fitness of the tran

method of eliciting the role of correlations among the vital rates (e.g. Ramula and Lehtila, 2005; Wisdom et al., 2000), based on linear correlations to measure dependencies.

Using a modeling approach, the first step towards understanding system behavior lies in extracting the relationship between variation in parameters and the resultant variation in model output, known as uncertainty analysis (Saltelli, 2000). The next step is to examine the contributions of the parameters to the variation in the output and their relative importance, also known as sensitivity analysis. A wide array of methods exist for sensitivity analysis , both local and global, covered by an extensive literature (for a comprehensive review see Saltelli, 2000). Local sensitivities have been widely used for simple models of population dynamics as they are derivative based, usually can be analytically derived and thus are not computationally intensive. For example, elasticity analysis performed for population viability analysis (Caswell, 1989; Morris and Doak, 2002, PVA; see) using stage-structured models are local methods, however, this is simple only if the response is linear and monotonic. Global sensitivity analysis (GSA) for ecological models, using variance-based techniques (an ANOVA like decomposition of the output variance), has recently become prominent and provide insight on the relative importance of factors in complex model settings. An accessible introduction to GSA using variance based techniques can be found in Saltelli et al. (2008) as well as two recent reviews introducing this approach to ecological applications (Cariboni et al., 2007; Fieberg and Jenkins, 2005).

Previously, models based evaluation of the risks associated with transgene introgression have utilized a deterministic framework for the the dynamics of the population, largely ignoring stochastic variation intrinsic to the populations as well as uncertainty in parameter estimates. Using matrix projection models with environmental stochasticity, Claessen et al. (2005) showed that extrinsic variation impacts the invasion risk and persistence of feral crops once transgene introgression has taken place. Prediction of transgene introgression into natural populations is contingent on evaluating the relative fitness of the transgenic type and the selection dynamics in natural populations. Fitness of the transgene is associated with pleiotropic effects on fitness components which can impact risk (Muir and Howard, 2001, 2002). Genotype × Environment ($G \times E$) effects (Devlin et al., 2006), when environments vary, can further compound the associated risk and thus necessitates multiple model evaluations. Furthermore, accuracy of assessments is compounded by the fact that transgenic phenotypic effect can only be deduced from laboratory experiments, therefore creating uncertainties in the estimated parameters. Thus constructing assessment models for transgene introgression within a predictive framework necessitates research integrating the above facets of uncertainty in conjunction to provide effective results for regulatory decision-making.

This dissertation research focuses on filling the lacunae in the existing methodology for assessing risk of transgene introgression into natural populations. Specifically, this research address the effects of demographic stochasticity and uncertainty in parameters on predictions for transgene introgression using a modeling approach. Furthermore, this research also provides some methodology improvements to existing modeling frameworks so they can implement biological mechanisms in a realistic manner. Specifically, the following objectives are addressed : 1) Determine the role of demographic stochasticity in modifying predictions of transgene introgression and the relative contributions of individual fitness components to the resultant variation, 2) Determine the role of uncertainty in estimating parameters on model predictions of transgene introgression and 3) Assess the global effect of parameters on model behavior using sensitivity analysis and the efficiency of sensitivity analysis through meta-modeling using a Bayesian Gaussian Process, to reduce the computational burden. Each Chapter in the thesis addresses a specific objective. Chapter 1 and Chapter 2 developed from the void in the existing literature on the impact of stochasticity for making predictions. While working on Chapter 2, it was clear that it was important to account for correlations among fitness components to make predictions realistic and biologically meaningful. However, current approaches are primarily ad-hoc

in their implementation of correlations and so we alternatively implemented a Copula approach (For an introduction to the theory of copulas see Nelsen, 2006) as formal methodology for incorporating correlations. Finally, based on conducting simulations for the first two research chapters, there arose a clear need to formalize approaches for running them and thus reduce the overall computational burden without sacrificing model complexity. This led to the adaptation of meta-modeling using a Bayesian approach in Chapter 3. The Bayesian framework is flexible enough to incorporate model extensions as well combine outputs across multiple models, e.g. using hierarchical modeling for population ecology (Clark, 2003) or model averaging (BMA; Gelman, 2004). With increased computational power Bayesian approaches are becoming popular for assessments in conservation (Wade, 2000), population modeling (Millar and Meyer, 2000), population genetics (Beaumont and Rannala, 2004) and evolutionary biology (OHara et al., 2008) and this choice will facilitate integration with those areas in future. Finally, as part of the research program a stochastic re-implementation of the net fitness component model(Muir and Howard, 2001) was also created, which will be deployed in the near future as software tool for decision-makers.

Cha

Den

prec

1.1 I

Escala

ing envir

Transgen

crop proc

sumers a

or elsewh

hazards a

(e.g., inse

tial envir

approach

the dynam

organism

Chapter 1

Demographic stochasticity alters predicted risk after a GEO invasion

1.1 Introduction

Escalating demand for food worldwide has fostered transgenic solutions for generating environmentally friendly produce, this is especially true in aquaculture (Aerni, 2004). Transgenic technology is gaining ground in livestock (Clark and Whitelaw, 2003) and crop production (Pilson and Prendeville, 2004), but has not been widely received by consumers and no transgenic animal has been approved for food consumption in the USA or elsewhere primarily due to environmental concerns. The potential for environmental hazards are higher with organisms that have wild relatives and high dispersal capabilities (e.g., insects and aquatic organisms National Research Council, 2004). Clearly, potential environmental impacts of GEOs should be evaluated and regulated using a scientific approach towards policy, wherefore research and assessment of impacts should capture the dynamics under a spatio-temporal context relative to a comparator, usually the same organism without the transgene (see reviews in Andow and Hilbeck, 2004; Andow and

Zwahlen

quires in

individua

as Devl

Key

of transg

Hedrick

the fitnes

The prob

transgene

populatio

and there

shown th

impacts n

Previo

transgene

from verb

matrix po

et al., 200

(Meagher,

2001; 200

pleiotropic

an age-stru

viability, ad

male fertili

Zwahlen, 2006; Tiedje et al., 1989). Assessing the environmental effects of GEOs requires information at multiple levels of biological organization; for example, transgenic individuals may induce population-level effects by altering interactions among individuals (Devlin et al., 2004) that can impact community and ecosystem dynamics.

Key to managing environmental threats of GEOs lies in evaluating the possibility of transgene introgression into natural populations and its resultant ecological impact (Hedrick, 2001). The likelihood of invasion and the persistence of a transgene depends on the fitness of the transgenic type relative to that of wildtype genotypes (Tiedje et al., 1989). The problem consists of two aspects: (a) prediction regarding the introgression of the transgene into wild populations and the associated changes in genetic composition of the population, and (b) predicting changes in ecological interactions following introgression and therefore ecosystem effects from changes in the wildtype population. If it can be shown that the probability of transgene introgression is null or very low then ecological impacts need not be considered (Muir and Howard, 2004).

Previous studies have used a plethora of modeling approaches to evaluate the fate of a transgene in populations (Hails and Morley, 2005). Models range in varying complexity from verbal models (Knibb, 1997), reaction-diffusion methods (Soboleva et al., 2003), matrix population models (Claessen et al., 2005; Garnier and Lecomte, 2006; Garnier et al., 2006), and spatial transfer models coupled with population dynamics and genetics (Meagher, 2003; Richter and Seppelt, 2004). For the case of GE fish, Muir and Howard 2001; 2002 found that the risk associated with transgene introgression depends on the pleiotropic effects of a transgene on fitness components. Their analysis was based on an age-structured population model characterized by six net fitness components (juvenile viability, adult viability, mating advantage, age at sexual maturity, female fecundity and male fertility, and reproductive longevity). They show that population extinction could resul

ing

vent

ind

ty

Ber

thes

ec

sna

Ha

M

of m

lato

whic

Earl

appr

stoch

risk

of de

vital

and p

the in

(Tulj

nam

ulatio

incor

is larg

result when the transgene was influenced by both sexual selection through increased mating advantage of male transgenics and natural selection through reduced viability of juvenile transgenic offspring, the "Trojan Gene" effect (Muir and Howard, 1999). Their model falls under the aegis of matrix population models (Caswell, 1989*sensu.*), the utility of which lies in linking individual life history traits to population dynamics (Grant and Benton, 2000) and provides insight into the evolutionary fate of populations. Currently these models are widely employed in applied ecology to predict the fate of populations of economic and conservation concern (Tuljapurkar, 1997). This approach provides a reasonable framework to evaluate the probability of gene introgression in natural populations (Hails and Morley, 2005).

Muir and Howard's model (Muir and Howard, 2001) can be considered an extension of matrix population models to multiple genotypes in a deterministic setting. Real populations, however, consist of a discrete number of individuals that experience chance events which influence their survival and reproduction, resulting in demographic stochasticity. Earlier models of transgene introgression have been based primarily on a deterministic approach. A notable exception is the use of matrix projection models with environmental stochasticity (Claessen et al., 2005), which show that stochasticity impacts the invasion risk and persistence of feral crops once transgene introgression has taken place. Sources of demographic variation can be divided into variation within individuals with the same vital rates (sensu. Pollard, 1966) and variation in vital rates among individuals (sensu. Fox and Kendall, 2002). Most work on stochasticity in age-structured populations evaluates the impacts of stochastic variation in vital rates (the parameters of the projection matrix) (Tuljapurkar, 1997). However, almost all considerations of stochasticity on population dynamics largely ignore genetic variation within populations. Contrastingly, classical population genetics focuses mainly on the dynamics of allele frequencies and only implicitly incorporates the effect of population size and age structure. Stochasticity in these models is largely a function of random sampling of gametes from an finite pool i.e random genetic drift

growth

can h

affect

for st

Shpa

gene

mutat

that tr

five c

et al.

stocha

and de

In

individ

gene u

How d

purely

stochas

probab

1.2

1.2.1

Our m

and How

drift and gloss over the effect of population dynamics (Ewens, 2004). Emerging studies combining aspects of both, genetics and population dynamics, suggest that the interplay can have significant impacts. For example, demographic stochasticity can influence the effective population size in age-structured populations (Engen et al., 2005a) and selection for stable population growth rates can favor genotypes with lower variance in vital rates (Shpak, 2007). Prediction of transgene introgression into natural populations is contingent on evaluating the relative fitness of the transgenic type and the selection dynamics in natural populations. Furthermore, accuracy of assessments is compounded by the fact that transgenic phenotypic effect can only be deduced from laboratory experiments, therefore creating uncertainties in relation to Genotype × Environment ($G \times E$) effects (Devlin et al., 2006). Thus constructing assessment models for transgene introgression within a stochastic framework necessitates understanding the role of stochasticity in both genetics and demography.

In this study, we investigate the effect of demographic stochasticity, primarily within individual variation in age structured models on the probability of fixation of a transgene using a Monte Carlo implementation. To this end we pose the following questions: How does demographic stochasticity affect predictions of extinction times relative to a purely deterministic model? What are the relative contributions of fitness components to stochastic variation, both individually and jointly? And finally, what is the impact on the probability of fixation and the dynamics of the transgene?

1.2 Materials and Methods

1.2.1 Stochastic Net Fitness Component Model

Our model is an extension of Muir and Howards Net Fitness component approach (Muir and Howard, 2001). A brief description of the deterministic model follows, for more details see Muir and Howard 2001 and see Appendix A.

We assume a diallelic model conferring three genotypes in the population (transgenic homozygote *ww*, wildtype homozygote *WW*, and the heterozygote *Ww*). The dynamics of the population of each genotype, comprised of a vector of age classes, can be formulated as:

$$\text{Reproduction}$$

$$n_{0,\,WW}^{t+1} = F_t^{WW} : f\left(\overrightarrow{N^{WW}} \times \overrightarrow{N^{WW}} \text{ and } \overrightarrow{N^{WW}} \times \overrightarrow{N^{Ww}}\right)$$

$$\begin{aligned} n_{0,\,Ww}^{t+1} = F_t^{Ww} : f\Big(&\overrightarrow{N^{WW}} \times \overrightarrow{N^{Ww}}, \overrightarrow{N^{WW}} \times \overrightarrow{N^{ww}} \\ &\text{and } \overrightarrow{N^{Ww}} \times \overrightarrow{N^{Ww}}\Big) \end{aligned}$$

$$n_{0,\,ww}^{t+1} = F_t^{ww} : f\left(\overrightarrow{N^{Ww}} \times \overrightarrow{N^{ww}} \text{ and } \overrightarrow{N^{ww}} \times \overrightarrow{N^{ww}}\right)$$

$$\text{Survival}$$

$$\begin{bmatrix} \overrightarrow{n^{WW}} \\ \overrightarrow{n^{Ww}} \\ \overrightarrow{n^{ww}} \end{bmatrix}^{t+1} = \begin{bmatrix} \mathbf{S}^{\mathbf{WW}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{\mathbf{Ww}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}^{\mathbf{ww}} \end{bmatrix} \times \begin{bmatrix} \overrightarrow{n^{WW}} \\ \overrightarrow{n^{Ww}} \\ \overrightarrow{n^{ww}} \end{bmatrix}^{t} \qquad (1.1)$$

where the F_t^g represents reproductive output for each genotype g (see Appendix A equation 5) at time t and are nonlinear functions of the reproducing individuals in all age classes over all three genotypes; the $\mathbf{S}^{\mathbf{g}}$ submatrices represent the age-specific survival parameters for each genotype g and $\overrightarrow{n^g}$ represents the population vectors of each genotype. The model is not analytically tractable when there are multiple genotypes, except under very simplified assumptions (Bodmer, 1965). Details on an comprehensive review of the population genetic analysis of age-structured populations in the deterministic context can be found in Charlesworth (1994).

The implementation of demographic stochasticity follows Pollard (1966), wherein the number

ements

the distr

tions. F

is a rand

vival par

related t

for femal

process f

fertility a

were appl

1.2.2 M

We use

tion based

tion. A th

ble 1.2) an

threshold o

mately 250

guage and a

2003(http:

Two diff

of eleven di

mentation of

Success, A a

correspond to

number of individuals $n^g_{a,t}$, within an age class a at time t, are random variables. The elements of the projection Matrix **A** (i.e. age specific vital rates) are parameters underlying the distributions generating the $n^g_{a,t}$ over time and are therefore constant in all simulations. For example, the number of individuals surviving from age 1 (n_1) to the next age is a random draw from a binomial distribution, i.e. ($n_2 \sim Bin(n_1, s)$), where s is the survival parameter (probability of survival) for all age 1 individuals. Similarly, stochasticity related to the other components in the model are implemented using a Poisson process for female fecundity, binomial processes adult and juvenile survival, and a multinomial process for mating success (also see Appendix B for a more detailed explanation). Male fertility and age of sexual maturity components were the only fitness components that were applied in a deterministic manner.

1.2.2 Monte Carlo Implementation

We used a Monte Carlo approach to understand the dynamics in the total population based on demographic stochasticity applied to the three genotypes in the population. A thousand runs were performed for each scenario (see paragraph below; Table 1.2) and each run stopped only when the total population reached the extinction threshold or reached a maximum of 20,000 time steps which corresponds to approximately 250 generations. The model was implemented in the C++ programming language and all random number generators were used from the GSL libraries (Galassi et al., 2003)(http://www.gnu.org/software/gsl/).

Two different sets of runs (hereafter Run1 and Run2) were performed. Run1 consist of eleven different scenarios of which the first four scenarios comprise stochastic implementation of each particular component singly (Scenarios F for Fecundity, M for Mating Success, A and J for adult and juvenile viability respectively). The next five scenarios correspond to stochastic implementations of combinations of components, to evaluate the

joint con
Juvenile
an imple
size of th
a reduce
in the n
juvenile
tial pop
were bas
numeric
genotype
the...t[W]
into the
sex ratio
populati
of the W
a thresho
stochasti

The s
eter valu
speed up
rameters
success,

joint contributions of each fitness component to the overall variation (e.g. Fecundity and Juvenile viability: Scenario F+J; detailed notation in Table 2). Scenario 10 (3Mill) is an implementation of stochasticity in all components with a larger starting population size of three million individuals for the WW genotype and Scenario 11 (AS2) was with a reduced age structure (weekly rates instead of daily). We used the same parameters for the fitness components for all scenarios based on a single combination of transgenic juvenile viability and transgenic mating advantage (2-fold) that leads to extinction of the total population (see Table 1.3 for paramter values used). Parameters for the simulations were based on the values estimated for transgenic Medaka (Muir and Howard, 2001). A numerical search algorithm was used to initially adjust the juvenile survival of the WW genotype to start the initial population (N^{WW} only) at a stationary stable age distribution (i.e., $\lambda^{WW} \sim 1$). Adult male and female transgenic individuals (N^{ww}) were introduced into the population at 0.001% of the WW population size (60,000 individuals in a 1:1 sex ratio). Generation time was calculated as the average age of females to replace the population (see Caswell, 1989, pg. 110) and all generation times reported are in terms of the WW population only. We used quasi-extinction times defined as the time to reach a threshold population size to alleviate confounding between the effects of demographic stochasticity and error propagation due to rounding.

The second set of runs (Run2) was performed over a two-dimensional grid of parameter values for juvenile survival and mating success of the ww genotype (Table 1.3). To speed up the simulations we used a weekly instead of daily age structure and the other parameters were adjusted accordingly. For each combination of juvenile survival and mating success, we ran 200 simulations to obtain the distributions at that particular combination.

1.3 Re

We prese

fecundity (F

in the main

they illustrat

appendix (se

At a thre

times was sl

time to quas

on the quas

involving m

variation ar

spectively. T

(=5% Table

tion of quas

the μ_{det} co

ponent cont

and any con

Other s

statistics of

or M and lik

(See Fig.

M~A+J, F~

Fig. A.1.1 s

1.3 Results

We present population trajectories and time to extinction for only four of the scenarios: fecundity (F), mating success (M), juvenile viability (J), and F+M+A(adult viability)+J, in the main text (see Table 1.2 for scenario descriptions). We focus on these scenarios as they illustrate the main patterns in the results, but do present plots for all scenarios in the appendix (see Fig. A.1.1), and highlight some specific results for other scenarios.

At a threshold population size of ~ 100 individuals, the distribution of quasi-extinction times was skewed to the right (Fig. 1.1; see also the quantiles in Table 1.1). The average time to quasi-extinction (hereafter μ_{QE}, calculated over all the 1000 runs) was centered on the quasi-extinction time from the deterministic model (hereafter μ_{det}) for scenarios involving mating success and fecundity only (scenarios F & M, Fig. 1.1). Additionally, variation around μ_{QE} for these components was low (%CV 1.3%, 0.6% and 1.26% respectively, Table 1). In all other scenarios, the μ_{QE} was lower than, but close to, the μ_{det} (±5% Table 1.1). However, μ_{det} was in the upper tail ($\geq 75^{th}$percentile) of the distribution of quasi extinction times from the stochastic simulations, except for scenario J where the μ_{det} coincides with μ_{QE} (Fig. 1.1, Table 1.1, μ_{det} percentile). The viability component contributed most variation in the time to extinction, specifically juvenile viability and any combination including it.

Other scenarios have similar trends to those used above, e.g., the distribution and statistics of (quasi) extinction times for scenario F+M were similar to those of scenario F or M and likewise scenarios AS2 and 3mill are qualitatively similar to scenario F+M+A+J (See Fig. A.1.1). Combinations with juvenile viability, i.e., scenarios with +J: A+J, M+A+J, F+A+J, had very similar distributions and moments to those of scenario J (See Fig. A.1.1 scenarios J, A+J, M+A+J, F+A+J, M+F+A+J).

1.3.1

When

scenario

frequenci

variances

Stocha

tions only

typically, it

only when t

is supportec

F-M).

In the ca

contribution

than adult vi

variances; for

1000 simulati

Variance i

over time incr

& Fig. 1.2b).

population size

of juvenile viab

in population si

almost exponent

Importantly,

variances of the

and thereafter it i

1.3.1 Fitness Components and Variance

When one fitness component was stochastic at a time, stochasticity in juvenile viability (scenario J) led to the highest variance in trajectories of total population size and gene frequencies, and stochasticity in fecundity (F) and mating success (M) led to the lowest variances (Fig. 1.2a & Fig. 1.2b, See Fig. A.1.2 & Fig. A.1.3 as well).

Stochasticity in fecundity and mating success components make significant contributions only when the population size is very small (Fig. 1.2a, Scenarios F & M). Analytically, it can be shown that stochasticity in these components contributes to variation only when the number of reproducing individuals are very low (see Appendix C) and this is supported by our simulations, singly (scenarios F & M) or in combination (scenario F+M).

In the case of viability, as expected (see Appendix C), juvenile viability has a larger contribution to variance in population size (Fig. 1.2a) and transgene frequency (Fig. 1.2b) than adult viability. In some circumstances, adult age classes contributed little to the variances; for example, in scenario A in which only two populations went extinct out of 1000 simulations (see Fig. A.1.2.

Variance in total population size and transgene frequency follows a convex trajectory over time increasing to a peak initially and then declining in all scenarios (Fig. 1.2a & Fig. 1.2b). Declining population size results in concurrent increase of the %CV in population size for all scenarios; however, in scenarios with a stochastic implementation of juvenile viability, either jointly or singly (scenarios J, F+A+J, M+A+J etc), the %CV in population size was $\geq$ 50% at half the time to extinction, and the rate of increase was almost exponential (See Fig. A.1.4a & Fig. A.1.4b scenario J and scenario F+M+A+J).

Importantly, the variance in total population size is initially lower than the sum of variances of the individual genotype populations (around 15-20 generations, Fig. 1.2a) and thereafter it is greater than the sum of the variances. This implies that the covariance

of the WW
time when t

1.3.2 Al

In the det
gene freque
ticity, it is e
to be center
erage transg
variation in
than for see
viability (see

An imp
the former, t
and the prob
Furthermore
tion (Pr(p^{WW}
M+A+J, F+
dity or matin
persistence o
gene frequen
p^{WW} = 0.00
i.e. the transg
in the stochas

of the WW genotype with the other genotypes changes both in sign and magnitude over time when the allelic diversity is highest at the transgenic locus.

1.3.2 Allelic Frequency of the Transgene

In the deterministic model, there was no fixation of the transgene; the maximum transgene frequency equalled 0.75, before the total population became extinct. Under stochasticity, it is expected that the distribution of maximum transgene frequency ($\max(p^{WW})$) to be centered around the deterministic value. For all scenarios, the evolution of the average transgene frequency ($\bar{p}^{WW}$) closely follows the deterministic trajectory. However, variation in the transgene frequency over time was lower for scenarios F, M, and F+M, than for scenarios that included stochasticity in those components, plus stochasticity in viability (scenarios F+J, M+J, and F+M+J; Fig 1.2b).

An important distinction between the stochastic and deterministic models is that, in the former, transgene frequencies can go to fixation in individual populations ($p^{WW} = 1$), and the probability of fixation depends on the components that are stochastic (Fig. 1.4). Furthermore, an important consequence of stochasticity is that the probability of fixation $\left(\Pr(p^{WW} = 1)\right)$ was high when viability components were stochastic (scenarios J, M+A+J, F+M+A +J; Fig. 1.4). There was no fixation of the transgene when only fecundity or mating success components were stochastic (scenarios F, M, F+M; Fig. 1.4). The persistence of the transgene was assessed as the proportion of simulations where the transgene frequency (p^{WW}) is ≥ 0.01 in 49 generations (Fig. 1.4). In the deterministic case $p^{WW} = 0.00176$ (at ww juvenile viability=0.65, ww mating success=5) in 49 generations, i.e. the transgene frequency is 10-fold lower than the persistence threshold. Contrastingly, in the stochastic case 1.5% of the 200 simulated populations had $p^{WW} \geq 0.01$.

1.

T

cu

wh

of g

diff

org

e.g.

nec

for c

C

unde

grap

stoch

It ha

unce

2005

wher

perce

T

tions

frequ

1.4 Discussion

The environmental risk associated with transgene introgression into natural populations can be characterized using a simple probabilistic approach (Muir and Howard, 2002) as

$$Risk \propto \Pr(Harm|Hazard) \times \Pr(Hazard) \tag{1.2}$$

where *Harm* denotes the possible community impacts and *Hazard* denotes the possibility of gene introgression and the symbol | represents the conditional operator. *Risk* can be difficult to characterize given the potential for impacts at multiple levels of biological organization, from populations to communities, except under straightforward situations, e.g. the Trojan Gene hypothesis (Muir and Howard, 1999), wherein population extinction occurs. Alternatively, if the probability of transgene introgression (Pr(*Hazard*)) is zero (or close to zero) then the *Risk* is also very small.

Our results show that that predictions from a purely deterministic model can severely underestimate the possible risks of extinction (Fig. 1.1; Scenario F+M+A+J). Demographic stochasticity is usually considered to be more important than other forms of stochastic variation for small population sizes, usually ≤ 100 individuals (Lande, 1993). It has been shown that the variance due to demographic stochasticity can contribute to uncertainty in estimates of extinction times in age-structured populations (Engen et al., 2005b). This also applies to the multi-genotype case and is supported in our simulations wherein the (quasi) extinction times from the deterministic model are usually in the upper percentiles of the distribution (Table 1.1).

Transgene introgression acts as a perturbation and modifies the variances in predictions of population size (Fig. 1.2a) which depends on the covariances among genotypic frequencies. Although we are unable to derive any explicit analytical explanation for

this phenom
structure of
of demogra
when age
taken into
size. Saethe
duce stocha
natural popu
brown bear
be inversely
2005a). The
has the pote
role in this p
invasion and
duce harvest
(Muir, 2004

We see in
tion probabil
offspring (ran
adult transge
This is a con
example, it is
from net-pen
derscores the
demography e

An import
promote highe

this phenomenon, it occurs under all scenarios and is independent of the form of age-structure of the population or population size. Recent studies have reconsidered the role of demographic variation in the context of predicting population dynamics, especially when age structure (Engen et al., 2005b) or mating system (Legendre et al., 1999) are taken into account. Demographic variances can range from 20-40% of the population size (Saether et al., 2004) and can slow down biological invasions (Snyder, 2003), reduce stochastic growth rates (Engen et al., 2005a) and impact demographic parameters in natural populations (e.g. avian populations (Saether and Engen, 2002) and Scandinavian brown bear populations (Saether et al., 1998)). Effective population size (N_e) is shown to be inversely related the demographic variance in age-structured populations (Engen et al., 2005a). Therefore, our results suggest that transgene invasion into wildtype populations has the potential to impact N_e and that demographic stochasticity can play an important role in this process. This needs to be accounted for in risk prediction models of transgene invasion and is especially valid within the context of using transgenic technology to reduce harvesting pressure on declining natural populations, such as the fisheries industry (Muir, 2004).

We see in our simulations that demographic stochasticity can impact predicted fixation probabilities (Fig. 1.4), even in the absence of random sampling of alleles to form offspring (random genetic drift). The simulations are based on a one time introduction of adult transgenic individuals and a low starting transgene allele frequency ($p^{ww} = 0.001$). This is a conservative approach, both in the number and frequency of introduction. For example, it is estimated that approximately 2 million farmed salmon escape every year from net-pen aquaculture into north Atlantic populations (Naylor et al., 2005). This underscores the importance for utilizing a stochastic framework to account for the role of demography even when assessing baseline risks of transgene introgression.

An important insight derived from our results is that demographic stochasticity can promote higher persistence of the transgene and make a significant contribution to the

variance in P
persistence o
the wildtype
an additional
a potential f
based on dyn
due to huma
et al., 1993;
estimated ris
the stability
Jones et al.,
pleiotropic e
populations t

Given the
to GEOs, co
effective app
1996). In c
using Monte
it possible t
2004). The n
genetics and
tributes diffe
are strong as
avian popula
offs can be l
1992) and th
ties underlyi

variance in predicted allele frequencies (Fig. 1.4). Demographic stochasticity elevates persistence of the transgene even when the overall fitness of the transgene is lower than the wildtype and predicted to be lost quite early in the deterministic context. This poses an additional risk at two levels; First, under the context of $G \times E$ effects this creates a potential for selection pressures on the transgenic type to vary over space and time based on dynamic environmental regimes. Current rates of global environmental change due to human populations can bring about rapid changes in biotic interactions (Kareiva et al., 1993) and elevated persistence of the transgene may necessitate modification of the estimated risk of impacts. Secondly, recent microevolutionary studies have questioned the stability of the genetic variance-covariance matrix (G-matrix (Ferriere et al., 2004; Jones et al., 2003)) and therefore, higher persistence of a transgene combined with the pleiotropic effects has the potential to transform the adaptive landscape of introgressed populations through the G-matrix.

Given the uncertainties associated with the evaluation of environmental risks attributed to GEOs, concern has been expressed that experimentation and modeling may not be an effective approach to make quantitative predictions and sound decisions (Kareiva et al., 1996). In complex systems where experiments are impossible to conduct, simulations using Monte Carlo techniques can provide an avenue of experimental evaluation making it possible to derive insight into the plausible sets among competing hypothesis (Peck, 2004). The net fitness component approach provides an useful framework to link both the genetics and population dynamics and our results show that each fitness component contributes differently to demographic variances. Furthermore, it has been shown that there are strong associations between demographic variance and life history characteristics (e.g avian populations (Saether et al., 2004)). Extensions incorporating life history trade-offs can be linked to the vital rates through the fecundity and survival functions (Roff, 1992) and therefore our modeling framework can also be useful in evaluating uncertainties underlying the risk of transgene introgression based on life history characteristics.

In conclusion, we can say that demographic stochasticity makes a significant contribution to uncertainty in predicting evolutionary trajectories in transgene introgression. This certainly needs to be accounted for in evaluating the associated risks and the Net Fitness component model provides a flexible framework not only to incorporate various sources of stochasticity but also additional life history information.

Scenario
F
M
J
A+J
F+M
F+A+J
F+M+A+
M+A+J
3mill
AS2

Table 1.1: St

extinction thr

distribution of

Scenario	Mean	StdDev	%CV	Variance	Qu. 5%	Qu. 95%	μ_{det}	μ_{det} percentile
F	62.504	0.671	1.073	0.450	61.422	63.603	62.060	0.257
M	62.088	0.397	0.640	0.158	61.470	62.748	62.060	0.490
J	65.487	20.414	31.172	416.714	53.277	79.471	62.060	0.500
A+J	58.698	13.799	23.509	190.415	48.334	72.062	62.060	0.779
F+M	62.506	0.787	1.260	0.620	61.277	63.904	62.060	0.300
F+A+J	58.859	17.552	29.821	308.071	48.081	70.560	62.060	0.760
F+M+A+J	58.144	10.810	18.592	116.862	48.380	70.084	62.060	0.781
M+A+J	58.273	14.307	24.552	204.687	47.780	71.217	62.060	0.775
3mill	83.473	9.087	10.886	82.573	75.373	93.139	87.349	0.822
AS2	65.333	18.971	29.037	359.889	57.004	74.641	63.662	0.501

Table 1.1: Statistics and quantiles for the distribution of extinction times using a quasi-extinction threshold of 100 individuals. The last column represents the percentile of the distribution of extinction times wherein the value from the deterministic model falls under.

Table 1.2: Scen.
stochastic. The

Stochastic Components	Scenarios
Fecundity	F
Mating Success	M
Juvenile Viability	J
Adult Viability	A
Fecundity and Mating Success	F+M
Adult and Juvenile Viability	A+J
Fecundity, Adult and Juvenile Viability	F+A+J
Mating Success, Adult and Juvenile Viability	M+A+J
Fecundity, Mating Success Adult and Juvenile Viability	F+M+A+J
F+M+A+J:Initial population 3 million individuals	3Mill
F+M+A+J: Weekly Age-Structure	AS2

Table 1.2: Scenario names and the associated components of fitness that were considered stochastic. The scenario names will be used for reference in the text.

Simulat set
Run
Run2

Table 1.3:
Run1 inclu
in an 1:1 s
genotype w
for simulat

		Parameters						
Simulation	Genotype	Juvenile	Adult	Female	Male	Age of	Mating	final
set		viability	viability	Fecundity	Fertility	maturity	Success	Age
Run1	*WW*	0.9324	0.95	8.8	1	64 d.	1	280 d.
	Ww	0.9292	0.95	8.8	1	64 d.	2	280 d.
	ww	0.9292	0.95	8.8	1	64 d.	2	280 d.
	WW	0.7432	0.698	8.8	1	9 wks	1	31 wks
Run2	*Ww*	0.61-0.75 by 0.01	0.698	8.8	1	9 wks	1-5 by 0.2	31 wks
	ww	0.61-0.75 by 0.01	0.698	8.8	1	9 wks	1-5 by 0.2	31 wks

Table 1.3: Parameter values used in both sets of runs. 83 generations for all Scenarios in Run 1 including AS2. Initial population size of the WW genotype was 60,000 individuals in an 1:1 sex ratio (30,000 males and 30,000 females). 60 adult individuals of the ww genotype were introduced into the population with a 1:1 sex ratio in the 65 days age class for simulation 1 and at 10 weeks of age for Run2.

Figure 1.1:
quasi-extin
sian kernel
the determ
distributior

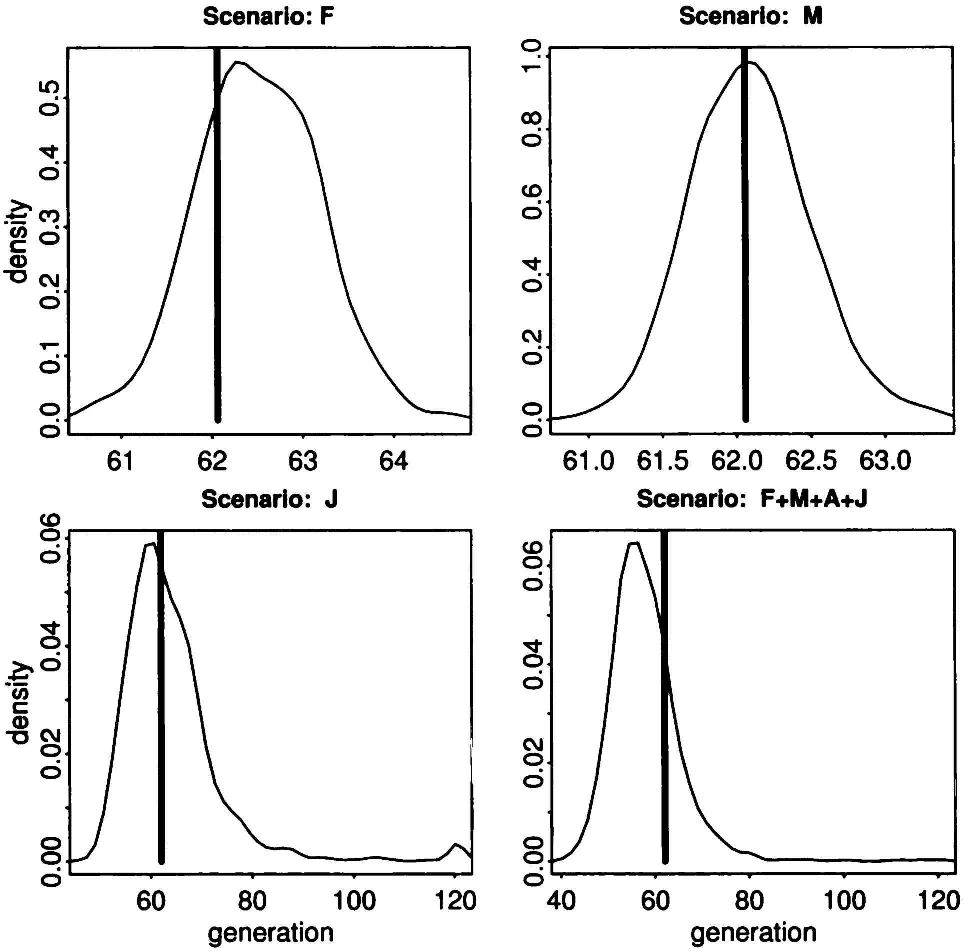

Figure 1.1: Distribution of extinction times for scenarios F, M, J and F+M+A+J at the quasi-extinction threshold of 100 individuals. A kernel density estimate using a Gaussian kernel is used for smoothing. The vertical bar represents the time to extinction in the deterministic model at the quasi-extinction threshold (μ_{det} Table 1). Plots showing distributions for all scenarios are available in the appendix (see Fig. A.1.4 in Appendix

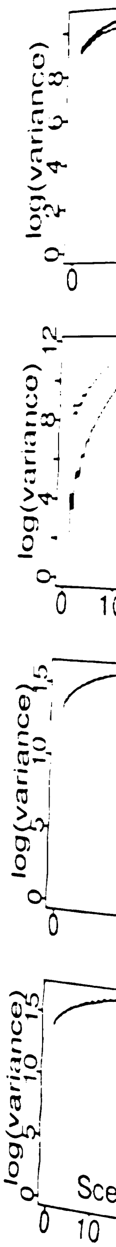

Figure 1.2: The
the total popula
Fig. 2b). Limit
were not extinct
variance of the
of the variances
transformed varia

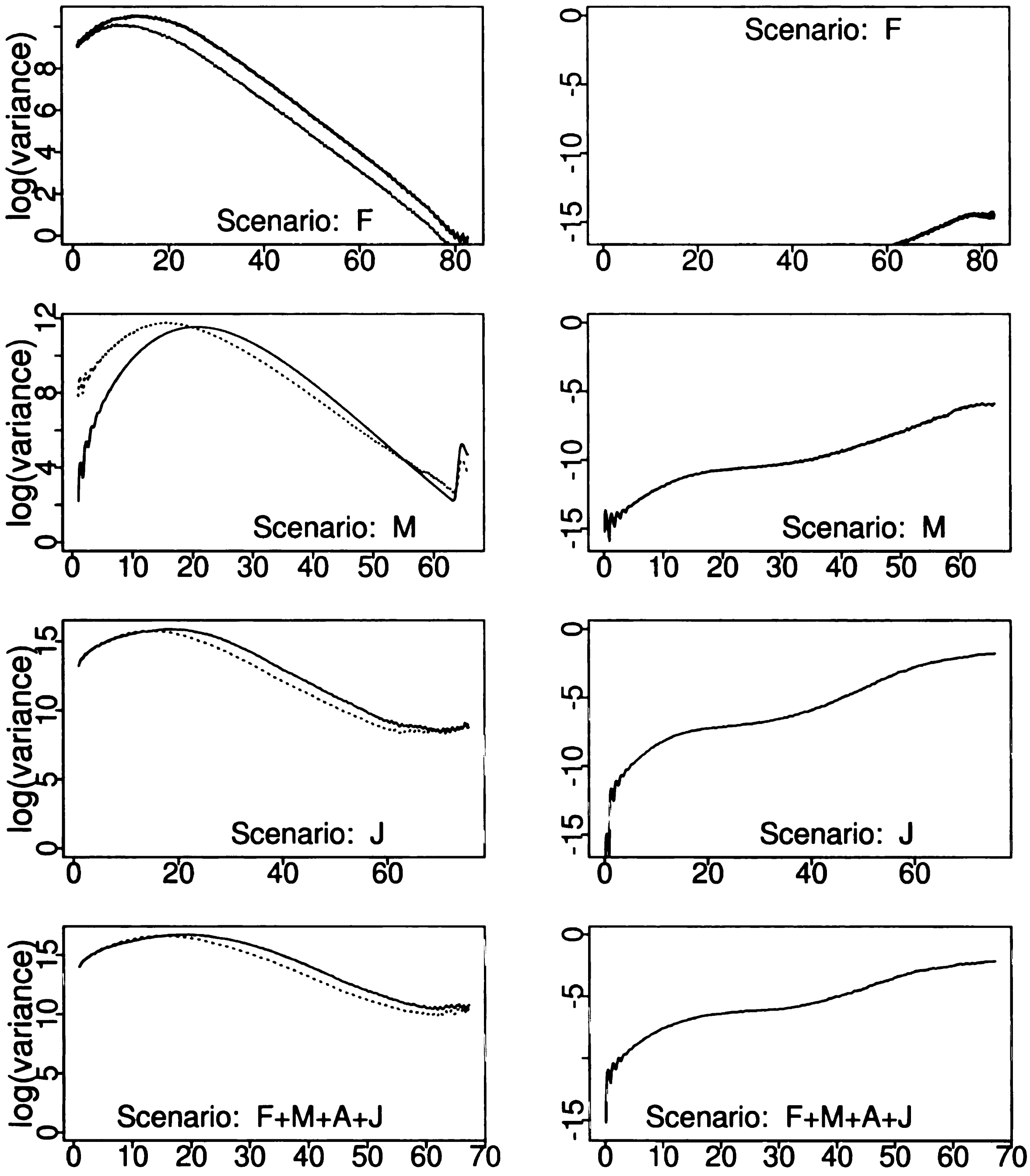

Figure 1.2: The evolution of variance over time for Scenarios F, M , J and F+M+A+J in the total population size (Left Column Fig. 2a) and the transgene frequency (right column Fig. 2b). Limit for the x-axis is chosen as the generation where at least 100 populations were not extinct over all runs for each scenario. a) The trajectory of the log transformed variance of the total population size (solid line) compared to the log transformed sum of the variances (dotted line) over the individual genotypes. b) The trajectory of the log transformed variance of the transgene frequency.

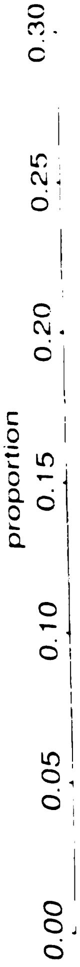

Figure 1.3:
tions in the
threshold

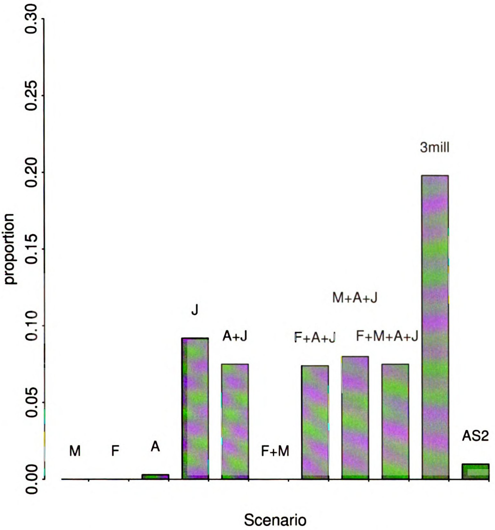

Figure 1.3: Probability of fixation of the transgene in all scenarios. Proportion of populations in the 1000 runs (Run1) where the transgene attains fixation at the quasi-extinction threshold of 100 individuals for all scenarios.

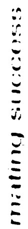

Fig.
whe
gen
eac

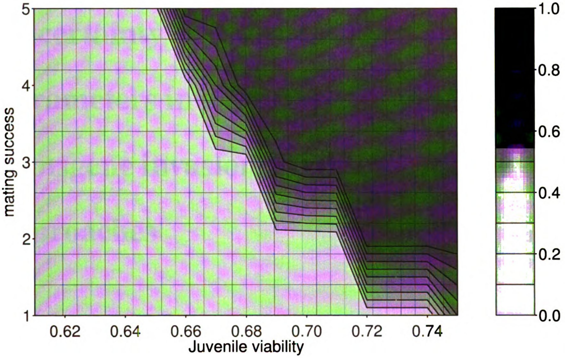

Figure 1.4: The persistence probability defined as the proportion out of 200 populations where $p^{ww} \geq 0.01$. Contours of the persistence probability of the transgene after 49 generations (Run2). The lines represent contours of the persistence probability across each combination of fitness components

Chapter 2

Uncertainty in estimates of model parameters in assessing the risk of transgene invasion: Monte Carlo approaches

2.1 Introduction

Ecological risk assessement focuses on the relationships among one or more stressors and their associated ecological effects. It serves as a guide for regulatory decision making and provides insights into the possible exposures and concomitant hazards. Recently, there has been an emphasis on making population-level risk an explicit consideration in ecological risk assessments (Barnthouse and Sorenson, 2007), and more importantly on integrating genetics and population dynamics into model based approaches (Nacci and Hoffman, 2008). Many types of population models have been developed for use in ecological risk assessment. For a recent review and classification of such models see, for example Barnthouse and Sorenson (2007). Risk assessment of the environmental hazards of genetically engineered organisms (GEOs) is one such area wherein population-level strategies need to be employed and this necessitates the integration of genetics and demography. Scientific approaches for assessing the risks of transgenic crops recognize the need to utilize an ecological basis (e.g. Andow and Hilbeck, 2004). The main emphasis is on transgenic organisms with dispersal strategies wherefore biological containment is imperative because they have wild relatives (e.g. plants, insects, aquatic organisms and microbes) with potential for hybridization leading to introgression or speciation (National Research Council, 2004).

Current methodology for assessing the ecological risks and impacts of transgenic organisms are based on a plethora of modeling frameworks ranging from verbal models (Knibb, 1997), reaction-diffusion methods (Soboleva et al., 2003), matrix population models (Claessen et al., 2005; Garnier and Lecomte, 2006; Muir and Howard, 2001, 2002), and spatial transfer models coupled with population dynamics and genetics (Richter and Seppelt, 2004; Meagher, 2003). A recurrent theme amongst these models is that they are mainly developed for assessing the risk of transgenic crops and focus on transgene introgression into wild relatives of the modified crop species (Hails and Morley, 2005). Aquaculture is another area wherein there is a large and increasing diversity of GEOs awaiting field release (Devlin et al., 2006) and there is a relative paucity of modeling frameworks in this area. Most species used in aquaculture have natural populations in the wild (e.g. salmon) and this, in conjunction with their dispersal characteristics, creates real potential for the transgenes to invade natural systems and cause perturbation at the community and ecosystem levels.

For the case of genetically engineered (GE) fish, Muir and Howard (2001; 2002) proposed a generic modeling framework based on matrix population models (*sensu.*Caswell, 1989) that incorporates genetics and demography. Their model is based on an age-structured population model characterized by six net fitness components (juvenile viability, adult viability, mating advantage, age at sexual maturity, female fecundity and male fertility, and reproductive longevity). The utility of this modeling framework lies in linking individual life history traits to population dynamics to provide insight into the evolutionary fate of populations (Grant and Benton, 2000). Muir and Howard (2001) found that the risk associated with transgene introgression depends on the pleiotropic effects of a transgene on fitness components. Additionally, they showed that population extinction results when the transgene is selected through increased mating advantage of transgenic males, even with lower relative viability of transgenic juveniles, the "Trojan Gene" effect (Muir and Howard, 1999). Currently, matrix population models are widely employed in applied ecology to predict the fate of populations of economic and conservation concern (Tuljapurkar, 1997) and thus are also advocated as a reasonable framework for assessing risks of GE crops (Hails and Morley, 2005).

Hitherto, most models for the evaluation of the risks associated with transgene introgression have utilized a deterministic framework largely ignoring any stochastic variation, both intrinsic and extrinsic, underlying the dynamics. However, recent advances in demographic theory have underscored the importance of the role of variation, both environmental and demographic, and is expounded to be imperative in understanding the influence of anthropogenic modifications on natural populations (Boyce et al., 2006). For example, it has been shown that variation in vital rates can signicantly affect outcomes of demographic models with important ramifications, such as estimates of population growth rate (Tuljapurkar et al., 2003) and persistence (Engen et al., 2005b). This is also relevant in the context of GEO risk assessment and it has been shown previously that stochasticity, both demographic (e.g.see Chapter 1) and environmental (e.g. Claessen et al., 2005), can influence persistence probabilities, gene introgression and extinction times.

Uncertainty of model parameters is another source of variation for model predictions of GEO risk and modeling approaches for transgene invasion have made feeble attempts to account for this in predictive assessments (e.g. Claessen et al., 2005; Kuparinen and Schurr, 2007). In the context of matrix population models, uncertainty in vital rates has

been implem
of the projec
involves sam
jectories (e.g
this incorpor
methods hav
tion matrix (
and Benton.
servation (M
context to ev
tions with li
impossible o

Along wi
nized that de
forecasting c
have focused
Japurkar, 199
for environm
times (Incha
ognized the i
fates of differ
lations model
based on life
as well modi
approaches u
rates (e.g. Ra

been implemented using two broad schemes of resampling. One involves resampling of the projection matrices over time for a single trajectory of the forecast and the other involves sampling a random set of projection matrices, each having their individual trajectories (e.g. Life-stage Simulation Analysis (LSA), Wisdom et al., 2000). However, this incorporates only the effect of environmental variation through the vital rates. Many methods have been developed to elicit the effect of individual parameters of the projection matrix on the overall growth rate of populations, such as elasticity analysis (Grant and Benton, 2000; Caswell, 1989), and utilized in forecasting for management and conservation (Morris and Doak, 2002). Monte Carlo approaches are specifically used in this context to evaluate the effect of uncertainty in parameter estimates, especially in situations with limited field data or when analytical derivations of the model dynamics are impossible or cumbersome (Tuljapurkar et al., 2003; Caswell, 1989).

Along with uncertainty in the parameter estimation process, it has also been recognized that dependencies among the vital rates are also important, especially under the forecasting context. On one hand, long standing considerations of these dependencies have focused primarily on temporal correlations among the projection matrices (Tuljapurkar, 1990) as opposed to among the individual vital rates themselves. This accounts for environmentally driven correlation in vital rates and is shown to decrease extinction times (Inchausti and Halley, 2003). On the other hand, life history theory has long recognized the importance of associations among vital rates in determining the evolutionary fates of different genotypes in populations (Stearns, 1992). In the context of matrix populations models, recent studies have shown that incorporating associations in the vital rates based on life-history theory can modify evolutionary stable states (Koons et al., 2008) as well modify estimates of population growth rates (Ramula and Lehtila, 2005). Most approaches utilize an ad-hoc method of eliciting the role of correlations among the vital rates (e.g. Ramula and Lehtila, 2005; Wisdom et al., 2000), based on linear correlations to measure dependencies. Linear correlation is widely used, with its popularity stemming from the ease of calculation. Although it is a natural scalar measure of dependence in elliptical distributions (e.g. the multivariate normal and t distributions), it also has the limitation that it is meaningful only on the linear scale. Therefore, using linear correlation when random variables are not jointly elliptically distributed might prove very misleading (Embrechts et al., 2003). While being largely unknown in the ecological literature, copulas have gained prominence for modeling dependence and are being extensively used in the financial literature (e.g in risk management Embrechts et al., 2003) and the hydrological sciences (Genest and Favre, 2007). Copulas provide a natural statistical framework for jointly modeling multivariate distributions that incorporate dependence without restrictions on the nature of the univariate marginal distributions. There are many excellent reviews in this area (e.g. Genest and MacKay, 1986) and we shall not attempt to do so here (For an introduction to the theory of copulas see Nelsen, 2006).

In this study, we evaluate the effect of uncertainty in parameter estimates on predictions of transgene invasion into natural populations using Monte Carlo approaches. We utilize data collected from laboratory experiments to estimate fitness components for GE zebrafish (*Danio rerio*). We use a non-parametric bootstrap approach (Efron and Tibshirani, 1998) to study effects of uncertainty in the estimated parameters on predictions of transgene introgression. Evaluations are restricted to the uncertainty in estimated fitness components among the genotypes in the population and the concurrent changes in gene frequencies over time. Since the experimental setup utilized allowed only independent estimation of fitness components, we propose an extension to the bootstrap methodology using a copula based approach to assess the effect of possible dependencies among the fitness components. We examine if the combination of the two approaches can provide a more realistic methodology for improving predictions in the risk assessment of GE organisms and therefore, more broadly be used as a general framework in applications of structured population models for conservation.

2.2 Model and Methods

2.2.1 Net Fitness Component Model

Our model is based on Muir and Howard's Net Fitness component approach (Muir and Howard, 2001), and a brief description of its deterministic version is given below. For further details see Muir and Howard (Muir and Howard, 2001).

We assume a diallelic model with three genotypes in the population (transgenic homozygote *ww*, wildtype homozygote *WW*, and the heterozygote *Ww*). The dynamics of the population of each genotype, comprising of the vector of age classes, can be formulated as:

Reproduction

$$n_{0,WW}^{t+1} = F_t^{WW} : f\left(\overrightarrow{N^{WW}} \times \overrightarrow{N^{WW}} \text{ and } \overrightarrow{N^{WW}} \times \overrightarrow{N^{Ww}}\right)$$

$$n_{0,Ww}^{t+1} = F_t^{Ww} : f\left(\vec{N}^{WW} \times \overrightarrow{N^{Ww}}, \overrightarrow{N^{WW}} \times \overrightarrow{N^{ww}} \text{ and } \overrightarrow{N^{Ww}} \times \overrightarrow{N^{Ww}}\right)$$

$$n_{0,ww}^{t+1} = F_t^{ww} : f\left(\overrightarrow{N^{Ww}} \times \overrightarrow{N^{ww}} \text{ and } \overrightarrow{N^{ww}} \times \overrightarrow{N^{ww}}\right)$$

Survival

$$\begin{bmatrix} \overrightarrow{n^{WW}} \\ \overrightarrow{n^{Ww}} \\ \overrightarrow{n^{ww}} \end{bmatrix}^{t+1} = \begin{bmatrix} \mathbf{S}^{\mathbf{WW}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{\mathbf{Ww}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}^{\mathbf{ww}} \end{bmatrix} \times \begin{bmatrix} \overrightarrow{n^{WW}} \\ \overrightarrow{n^{Ww}} \\ \overrightarrow{n^{ww}} \end{bmatrix}^{t} \tag{2.1}$$

where $g = WW, Ww$ or ww are the wildtype, heterozygote and transgenic genotypes respectively; the F_t^g ($g = WW,\ Ww$ or ww) represents reproductive output for each genotype g at time t and are nonlinear functions of the individuals of reprodutive age across all three genotypes; the $\mathbf{S}^{\mathbf{g}}$ submatrices represent the age-specific survival parameters for each genotype g; and $\overrightarrow{n^g}$ represents the population vectors of each genotype. The model is not analytically tractable when there are multiple genotypes, except under very simplified assumptions (Bodmer, 1965). Comprehensive review of the population genetic analysis of age-structured populations in the deterministic context can be found in Charlesworth (1994). Hereafter the WW genotype will be referred to as the WT genotype and the ww genotype as the TG genotype respectively.

2.2.2 Experimental Setup

The experimental setup is described in Muir et.al.(2006). Briefly, fitness components were measured for zebrafish individuals (*Danio rerio*) for the transgenic (TG) and the Wildtype (WT) genotype in the laboratory as follows. Age of maturity was measured for 10 females as the day on which they started laying eggs. Fertility of males and fecundity of females for the TG genotype was estimated from reciprocal crosses of WT males with TG females and vice versa. A total of 20 mating pairs were used for each cross. From an additional experiment, crosses of $WT \times WT$ matings (10 replicates) were available and these data were used to estimate the fecundity and fertility components for the WT genotypes. To estimate viability, 50 individuals (1:1 sex ratio) of each genotype (TG 2 reps; WT 4 reps) were followed from the fry stage and the number surviving in each replicate tank counted at approximately 30 day intervals. The replicate was assumed to have matured once females in each replicate started laying eggs. This experiment was terminated once the females stopped laying eggs after 501 days. To estimate mating success three different ratios of WT:TG (10:2, 5:5, 2:10) males were mated with 12 WT females with two replicates for each ratio. The total number of eggs laid were collected and the offspring genotypic proportions recorded after hatching.

Validation Experiment: Mesocosm: Two replicated mesocosms were set up with a starting population size of 500 adults (1:1 sex ratio) with 5:1 ratio of WT:TG genotypes.

Each mesocosm was followed for 10 generations and the time to loss of the transgene, and the frequency of WT to TG genotypes was recorded. The mesocosm protocol involved collecting the fry from each replicate and randomly selecting individuals from each replication to start the next generation with the same starting population size. Thus effectively we monitor the change in gene frequency over generations independent of population regulation.

2.2.3 Data Analysis

For each fitness component specific probability models underlying the data generating mechanism were assumed and a maximum likelihood approach was used to estimate the parameters. Standard distributions generally used for modeling applications in ecology were considered as follows: Fecundity $\sim$ *Poisson*(Λ); Fertility $\sim$ *Beta*(α,β); Age of Sexual Maturity $\sim$ *LogNormal*(μ, σ^2); Viability, both juvenile and adult, were assumed to be a non-linear function of age with Gaussian errors, as detailed below. Where possible analyses were conducted accounting for possible interactions among the genotypes, e.g. the Age of Sexual Maturity component was analyzed separately for the $WT \times WT$, $WT \times TG$ and $TG \times TG$ combinations, using an ANOVA decomposition to test for any reciprocal effects. A brief description of the analysis for each component follows below.

Age of Sexual Maturity: Based on the empirical distribution of the data, we assumed a Lognormal distribution for age of sexual maturity and analyzed the data with a linear model after log transforming the response variable. The location and scale parameters for the log-normal distribution were obtained by back-transformation using bias correction (Beauchamp and Olson, 1973).

Fecundity: Fecundity of females was assumed to follow a Poisson distribution with parameter Λ, denoting the expected number of eggs produced per female over the lifetime. To estimate Λ a generalized linear model with a log link function was specified.

Reciprocal effects of the genotype on fecundity was assessed by using the genotype of the mate as a factor.

Fertility: Fertility of males was assumed to follow a *Beta* distribution with parameters α and β, where the average fertility is given by $\frac{\alpha}{\alpha + \beta}$, the expectation from the *Beta* distribution.

Viability: Viability was modeled as an age dependent mortality function of the number of individuals entering a cohort using non-linear regression. We assumed a non-linear model with Gaussian residuals, with the mortality over age as an exponential function of cohort size and a constant mortality rate (z), which translates into $N_t = N_0 e^{-zt} + \epsilon_i$,where $t = 1, 2, 3\ldots$ are the age classes; z represents the mortality parameter, assumed constant for all age intervals; and $\epsilon_i \sim N(0, \sigma^2)$.

Mating Success: Since there were only two replicates for each combination of WT:TG proportions, the relative mating advantage was calculated for each replicate and averaged over the replicates for each combination. The average was compared to the expected proportions using a chi-square test. For each replicate the expected proportion of offspring genotypes was calculated as the product of the frequency of WT males and the number of offspring. The WT genotype was chosen since it was the recessive genotype and therefore the phenotype uniquely identifies the genotype. The ratio of the expected proportion to the observed proportion of WT offspring gives the relative mating advantage (disadvantage) of the WT genotype. The underlying assumptions are that there are no segregation distortion, probability of fertilization and hatching are the same for all genotypes.

All analysis of the experimental data were done using the statistical package R (Team, 2009). The parameters of the *Beta* distribution were estimated using a maximum likelihood approach by the `fitdstr` function from the `MASS` library. For nonlinear regression the `nls` function from the `nlme` package was used.

2.2.4

Non-Para

script in I

fitness com

and then a

each boots

in the para

dataset was

from the fitt

This bootstr

this case onl

Dependenci

components

used a multiv

ness compone

using linear co

and Lehtila, 2

Doak, 2002). T

dependencies a

variables. The

are not restricte

ber of existing r

more distributio

for a review). Br

random variables

1986). If $F(X)$ a

2.2.4 Monte Carlo Approach

Non-Parametric Bootstrap: Independence among vital rates We implemented a script in R to estimate the bootstrap distributions of the estimated parameters for the fitness components. For each fitness component the data were sampled with replacement and then analyzed with the corresponding statistical model to get parameter estimates for each bootstrap sample. We used 1000 bootstrap samples for estimating the uncertainty in the parameter estimates. Alternatively, for the viability component, since the available dataset was fairly small across the life-span of the experiment we resampled the residuals from the fitted model to generate the bootstrap distributions (Efron and Tibshirani, 1998). This bootstrap strategy may involve some extra model assumptions, thus it was used in this case only because of the limited experimental samples size.For estimating viability.

Dependencies using Copulas To assess the effect of dependencies among the fitness components we also extended the bootstrap methodology using a copula approach. We used a multivariate copula to incorporate dependencies among the joint distribution of fitness components. Previously dependencies have been accounted for in an ad-hoc fashion using linear correlations, either through incorporating pair-wise correlations (e.g. Ramula and Lehtila, 2005) or a transformation from the multivariate normal (e.g. Morris and Doak, 2002). The copula approach provides a formal statistical methodology to estimate dependencies as well generate samples from the joint distribution of correlated random variables. The principal advantages of this approach are that the marginals distributions are not restricted and can be from any family of distributions. Copulas encompass a number of existing multivariate distributions and provide a flexible framework for generating more distributions(see Genest and Favre, 2007; Nelsen, 2006; Embrechts et al., 2003, for a review). Briefly, a copula is a function that defines the joint distribution of a set of random variables based on their univariate marginal distributions (Genest and MacKay, 1986). If $F(X)$ and $G(Y)$ are the marginal distributions of two random variables X and

Y with a jo

function of t

the copula.

We used

samples fro

Gaussian Co

ponents with

history (Stea

where ρ repr

and 4 represe

respectively;

$-0.5, \rho_{24} =$

The copu

ing the unde

collected is r

tion we samp

0.5. For simp

mean and va

were based o

ples with 100

components.

the fitness par

ated in our ex

Y with a joint density $H(X, Y)$, then we can rewrite the joint distribution $H(X, Y)$ as a function of the marginals $H(X, Y) = C(F(X), G(Y))$, where the function $C(\cdot, \cdot)$ represents the copula.

We used the R package copula (Yan, 2007) for working with copulas. To generate samples from the joint multivariate distribution of the fitness components, we assumed a Gaussian Copula for the Fertility, Fecundity, Age of Sexual Maturity and Viability components with correlation structure as shown below, based on potential trade-offs in life history (Stearns, 1992):

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix}$$

where ρ represents the correlation coefficient among pair of fitness components; 1, 2, 3 and 4 represent the Fertility, Fecundity, Viability and Age of Sexual Maturity components respectively; and $\rho_{12} = \rho_{21} = 0, \rho_{13} = \rho_{31} = -0.5, \rho_{14} = \rho_{41} = 0.5, \rho_{23} = \rho_{32} = -0.5, \rho_{24} = \rho_{42} = 0.5$ and $\rho_{34} = \rho_{43} = -0.5$.

The copula approach allows us to generate replicate multivariate samples incorporating the underlying correlation structure, similar to a parametric bootstrap. The data we collected is not amenable for directly estimating the copula parameters and for illustration we sampled from a multivariate distribution with absolute values of correlations set to 0.5. For simplicity, we assumed the viability parameters were normally distributed, with mean and variance based on the estimates from the original data. All other parameters were based on the estimated marginal distributions of the data. A 1000 bootstrap samples with 100 observations each were generated from the joint distribution of the fitness components, and the same statistical models as the section above were used to estimate the fitness parameters. Sample sizes were unequal among the fitness components evaluated in our experiments. Therefore, to reproduce the marginal distributions in the copula

approach tha
sample sizes

Simulations
net fitness co
quency. The
reps for a pa
lations for ea
and error pro
instead of d
rates to ensu
age structure
nant eigenva
the populati
Any new re
zation based
2002). Our
was unchan
weekly fecu
viability to
individual g
(Wall, 1996
eter of the r
daily model
viability be
daily model
optimization

approach that were similar to the bootstrap distributions of estimates, we used adjusted sample sizes ($\sim$ 100) for all components.

Simulations For each set of bootstrap estimates of the fitness components we ran the net fitness component model to generate distributions of the evolution of transgene frequency. The estimated values for the mating success component was based only on two reps for a particular WT:TG combination. Therefore we ran three different sets of simulations for each estimated value of the WT:TG combination. To reduce model complexity and error propagation we used a reduced age-structure, so the model represented weekly instead of daily time steps. This necessitated re-parameterization of the estimated vital rates to ensure that the dynamics of the reduced model closely follows that of the original age structure. Under the theory of matrix population models (Caswell, 1989), the dominant eigenvalue (λ) of the projection matrix is representative of the long-term dynamics of the population and is based on the primitivity and irreducibility of the projection matrix. Any new representation should try to mimic the dynamics of the original parameterization based on λ, the generation time and the age distribution (Yearsley and Fletcher, 2002). Our re-parameterization was carried out as follows: Estimated fertility of males was unchanged and estimated fecundity of females was rescaled by a factor of 7 to reflect weekly fecundity instead of daily fecundity. Adult viability was rescaled by raising the viability to the 7th power, so that the overall survival of the adults over the adult stage of individual genotypes in the population remained the same. We used a Genetic Algorithm (Wall, 1996) to optimize the juvenile viability parameter such that the resultant λ parameter of the reparameterized model and the generation times were equivalent to that of the daily model, to reproduce the same dynamics. This resulted in the optimum WT juvenile viability being lower than the TG juvenile viability in the weekly model relative to the daily model. The juvenile viability was the parameter chosen to be modified during the optimization since it was the worst estimated fitness component due to small sample size i.e. there were only 6 data points over 2 reps for the TG genotype and 12 points over 4 reps for the WT genotype to estimate this component, over ~ 90 and ~ 60 day intervals respectively.

2.3 Results

Estimates of fitness components for the Transgenic (TG) and the Wildtype (WT) genotypes are given in Table 2.1. For all fitness components the WT had higher relative fitness values and therefore in this case it is expected that the transgene will be lost from the population. The mating success for each of the different proportions (WT:TG combinations) were different as there was some frequency dependence in mate choice (Table 2.1). Calculations of the estimated asymptotic growth rate (λ_{WT} =1.0422 and λ_{TG} =1.0213) based on average parameter values indicated that WT was the favored allele. Additionally, the WT males had greater relative mating advantage taking on three possible values (1.16, 1.93 and 5 times that of TG). The deterministic model based on the estimated values of the fitness components estimated the transgene lost (fixation of the WT allele W) within 16.17, 6.68, 3.69 generations with increasing mating advantage of the WT males respectively. All generation times are reported with respect to the WT genotype and calculated based on the average time for a female to replace herself (Caswell pg 110).

2.3.1 Independence among vital rates: Non-Parametric Bootstrap

Distribution of Parameters The distribution of the bootstrap estimates of the fitness component parameters are given with the associated summary statistics (Table 2.1; Fig. A.2.1). As mentioned above, not only does the WT genotype have higher fitness on average over all the individual components (Table 1), but also it has higher fitness under all combinations of the fitness components. This is shown by the low probability of the relative values (TG/WT) of the fitness components being < 1 for all sets of the bootstrap

estimates
it is expec
no possibi
of female
few variat
2.1). Fec
due to ove
shown). Th
negative bi
extra variat
components
the dominar
l_{rel}. The
~97% base
relatively sm
dynamics. T
% change in
2.3).

Transgene fr
given in Figur
advantage of t
decreases in a
the WT genot
stable to the va
flat surface 2-*d*
genotype (*Fig.*

estimates (Table 2). Since the WT genotype has a higher relative male mating advantage, it is expected that the transgene will be lost under all conditions (Fig. 2.1), i.e there is no possibility of a "Trojan Gene" effect (Muir and Howard, 1999). With the exception of female fecundity, the distributions of the fitness component estimates had a relatively low variation around the expected values as seen by the %CV of the estimates (Table 2.1). Fecundity estimates for both WT and TG had a greater than expected variance due to overdispersion relative to the Poisson distribution in the original data (results not shown). This can be easily accommodated by using a different probability model, e.g. the negative binomial, although in our case the nonparametric bootstrap method captures this extra variation (Table 2.1; %CV for fecundity). Uncertainty in the estimates of the fitness components results in variation in the estimated λs for each genotype,calculated from the dominant eigenvalue of the projection matrix, and also their relative values, hereafter λ_{rel}. The λ_{rel}s ranged from ~94%-99% ($\frac{\lambda_{TG}}{\lambda_{WT}}$ respectively Table 2.2) compared to ~ 97% based on the averages for the individual genotypes. While the shift in λ_{rel} is relatively small (~ ±3%) across all simulations, this induces substantial changes in system dynamics. These changes can be seen by the large %CV of the fixation times, relative to % change in λ_{rel} across all three scenarios of mating advantage for WT males (Table 2.3).

Transgene frequency The distribution of the time to fixation for the WT genotype is given in Figure 2.2 for the three different values of male mating success. As the mating advantage of the WT genotype increases the average time to fixation of the WT genotype decreases in a nonlinear fashion. Concurrently, the variance in the time to fixation of the WT genotype also decreases. For both genotypes the fixation times are relatively stable to the variation across any individual fitness component as seen by the relatively flat surface 2-dimensional kernel density estimates,except for the fecundity of the TG genotype (Fig. A.2.2 and A.2.3). This is also the case for the relative values (Fig. 2.3) and

of th

the W

A.2.3

2.3.2

Distrib

rameter

Table 2.

the multi

expected.

approach

dity are fo

this can be

bution of *r*

scales (Fig.

modify the

of their dist

lation (Rho)

scale the *com*

tive values *of*

also induces

components w

dispersion is rec

ative to the ind

~40% reductio

and is independent of the values of male mating success. The distribution of fixation times of the WT genotype show dependence only with respect to the distribution of the λs of the WT and TG genotypes and this relationship also holds for the λ_{rel}s (Fig. A.2.2, Fig. A.2.3 & Fig. 2.3).

2.3.2 Copulas: Dependence among fitness components

Distribution of Parameters By design the marginal distributions of the estimated parameters were elicited to be similar to that of the independent configuration (Fig. A.2.1; Table 2.1), however, the Copula induces correlations among the parameters that restricts the multivariate distributions to a reduced subset of possible configurations (Fig. 2.4). As expected, the pairwise joint distribution of the parameters are different under the copula approach even though the marginal distributions are the same. Since fertility and fecundity are for different sexes we did not simulate them to have any correlations, although this can be easily accommodated in this approach. The relationship of the joint distribution of *relative* fitness parameters are not similar to their relationship at the absolute scales (Fig. 2.5 vs Fig. 2.4). Including dependencies among the fitness components can modify the joint distribution of the relative fitnesses as revealed by pairwise comparison of their distributions (Fig. 2.5), which exhibit negative correlations (Spearman's correlation (Rho); Table 2.3b). For example, in our case we see that although in the absolute scale the correlation between fertility and sexual maturity is positive (Fig. 2.4), the relative values of these fitness components are negatively correlated (Fig. 2.5). Dependencies also induces changes in the marginal distributions of the relative fitness values for all components with increased dispersion (Table 2.2), except for sexual maturity where the dispersion is reduced. Additionally,dependencies restrict the possible values of λ_{rel}s, relative to the independence approach, wherein the distribution is more concentrated with $\sim 40\%$ reduction in the dispersion (Table 2.2).

Transgene F

n comparison

components

erage over th

independent

for the fixati

proach (Fig.

fixation is n

(Fig. S2). I

bution of th

fecundity a

observed w

(Fig. 2.3).

nents are n

(Fig. 2.3).

well when

2.3.3 V

The valid

drops to z

three scen

for every

prediction

both mes

Transgene Frequency The average time to fixation for the WT genotype was similar in comparison to the scenario under which there were no dependencies among the fitness components (WT mating success 1.93 vs 1.93(C) Table 2.3a). The trajectory of the average over the replicates with dependencies were similar to that of the average over the independent replicates (Fig. 2.1). With dependencies among vital rates the distribution for the fixation times changes with a reduced dispersion relative to the independence approach (Fig. 2.2). Comparison across individual components show that the probability of fixation is not related to the any of the components in both the WT and the TG genotypes (Fig. S2). In the case of the relative fitness components, it can be seen that the distribution of the fixation times show linear trends as a function of the relative fitness for the fecundity and sexual maturity components (Fig. 2.3). In the case of λ_{rel}s, the linear trend observed with dependencies has a greater slope compared to the independence scenario (Fig. 2.3). In all cases, the joint distribution of fixation times and relative fitness components are more concentrated and show lower dispersion than the independence scenario (Fig. 2.3). Finally, the distribution of fixation times have reduced mean and variance as well when dependencies are present among vital rates (Table 2.3a).

2.3.3 Validation with Mesocosm Data

The validation experiments had only two reps and in both reps the transgene frequency drops to zero within 10 generations. We extracted the 99% prediction intervals for all three scenarios of mating success for 10 generations, by calculating the generation time for every bootstrap replication based on the vital rates for that particular replication. The prediction intervals for transgene frequency with WT mating success=1.16 envelopes both mesocosm realizations, while the prediction intervals envelopes only one replicate when WT mating success=1.93 (Fig. 2.6). The scenario with the highest mating success predicts reduction in transgene frequencies to be much faster than the observed realizations. The results should be interpreted with some caution in the light of only two stochastic realizations available from the experimental results. The 99% intervals can be misleading if the density of the distribution is not symmetric, as in our case, and this is can be visualized by the vioplots (Fig. 2.6).

2.4 Discussion

In Chapter 1 it was shown that demographic stochasticity, arising from variation at the individual level, can alter predictions of the risk associated with transgene invasion. Specifically, for the "Trojan gene' effect demographic stochasticity can reduce extinction times relative to the deterministic model and more generally increase the persistence of the transgene in the system. Here we consider the effect of uncertainty in estimated parameters on model predictions applying the non-parametric bootstrap approach.

The bootstrap is a flexible resampling approach, in implementation and application, and its use is advocated to assess uncertainties in parameter estimation when using structured population models (Caswell, 1989; Wisdom et al., 2000; Morris and Doak, 2002). Models of transgene invasion have focused on the invasibility of the transgene (e.g Soboleva et al., 2003; Meagher, 2003; Garnier and Lecomte, 2006) and do not explicitly incorporate uncertainty associated with estimating parameters, due to a combination of lack of data and model complexity. This has already come under criticism in the context of model of population viability analysis (PVA) in conservation biology (Fieberg and Ellner, 2003). Our results, using the bootstrap procedure, show that uncertainty in estimation of model parameters contributes to considerable variation in predictions of the loss of the transgene (Fig. 2.1). Furthermore, bootstrapping is useful when evaluating the uncertainty of functions of model outputs (e.g. fixation time) which are analytically intractable.

This is especially relevant in the context of GEO invasion, where the overall population growth rate (λ), as well as genotype specific growth rates, (λ_g) are non-linear functions of the vital rates of all genotypes, unless simplifying assumptions are made regarding frequency dependence (Bodmer, 1965; Charlesworth, 1994) or density dependence (Benton and Grant, 1996). Using a bootstrap assessment of uncertainty in estimated vital rates, a re-analysis of data in avian populations showed predictions of extinction risk were modified (Saether and Engen, 2002). In our case, we see that variation in fixation times of the WT genotype are considerably higher in magnitude (Table 2.3), relative to variation in the estimated parameters (Table 2.1).

2.4.1 Dependence vs Independence

There is substantial theory from a life history evolution perspective, which links vital rates to fitness and elucidates the mechanistic basis underlying covariation in vital rates (Roff, 1992; Stearns, 1992). It is acknowledged that covariation among vital rates is an important determinant towards making realistic predictions and ignoring this can lead to spurious conclusions (Coulson et al., 2005). Using the copula approach, our predictions incorporating dependencies among fitness components show that prediction intervals can be over-estimated when dependencies are ignored (Fig. 2.6). Estimating covariation of vital rates has been fraught with difficulty and most work have utilized ad-hoc methods to incorporate correlations, usually using the well known Pearson correlation coefficient (e.g. Ramula and Lehtila, 2005). For instance, Morris and Doak (2002) recommend utilizing a multivariate Normal distribution incorporating correlations and transforming tbe resultant Normal marginal distributions to the distributions of interest. However, it is well known that correlations in the original scale are not an one-to-one transformation on the underlying normal scale. The Pearson correlation coefficient is appropriate only for elliptical distributions such as the Normal since it is a linear operator (Embrechts et al., 2003).

Recently, there has been a re-emergence of a less known statistical approach to model dependence among arbitrary distributions based on their univariate marginal distributions, the copula approach. Our results using this approach also show that correlations among the estimated vital rates can provide more realistic predictions by excluding regions of the parameter space that are highly unlikely (Fig. 2.5). An additional observation that stems from our results is that dependencies among the absolute values of fitness components induces correlation among relative fitness values. In our case, this occurs in the opposite direction to correlations on the absolute scale (Table 2.3b, Fig. 2.4 vs Fig. 2.5). While this may be a statistical artifact arising from ratios of the two random variables, this merits further study as the final determinant of selection is based on relative fitnesses and not absolute values.

2.4.2 Considerations of experimental methods

A decade ago there was little faith in the concept of combining models and short-term experiments to make predictions about transgene invasion into natural populations. The main contributions from invasion theory was considered to be on guidance towards setting up monitoring schemes and sampling programs that will be cost-effective in detecting a problem invasion (Kareiva et al., 1996). Comparisons of the prediction distributions versus trajectories of the validation data show surprisingly good agreement (Fig. 2.6). This is an optimistic outlook considering that there are only two replicates and also considering that correlations can modify the predictions. However, the present results do provide the promise that models combined with data can provide a realistic predictive framework for risk assessment of GEO's.

Although experimental data was available to estimate fitness components, there were pitfalls in our experimental design for the current study. Our experiments were good enough to provide estimates of the marginal distributions of fitness components as reflected in the agreement of our model predictions with the validation data (Fig. 2.6). However, the current design is not amenable to estimating the underlying correlation structure among vital rates and resulting in over or under estimates of the bounds when assessing uncertainty. Incorporating the Copula approach allows for a rigorous statistical methodology that provides a framework to extend current modeling methodology by accounting for covariation in vital rates and making predictions more realistic. This extension can be of great utility in the context of GEO risk assessment due to the necessity for making quantitative predictions of transgene invasion.

Data requirements for utilizing a Copula approach principally comprises of simultaneously collecting variables on the vital rates. The basic premise that there exists an unique Copula for continuous multivariate joint distributions provides an avenue for statistical estimation not only on the marginal distributions, but also on the underlying correlation structure. Likelihood based methods already exist (e.g in the `copula` package) to estimate the parameters for a copula based on the functional form of the copula and the marginal distributions. The necessity of jointly measurements for data collection can be easily accommodated when conducting laboratory experiments to estimate fitness components. For example, fertility (fecundity for females), sexual maturity and viability, in both adults and juveniles, can be measured on the same replicates, by estimating them concurrently for the entire cohort. While this would entail separate measurements for the two sexes the net fitness component model can be easily modified to accommodate the different sets with the added bonus of increasing predictive accuracy. Once the joint measurements are available, there are many methods currently available to estimate the joint distributions using the copula approach. This is an active area of research which includes development of methodology for making model choice based on different copulas. The proposed modifications to designing experiments can also be applied in general to even population viability assessments (PVA) using matrix population models.

2.4.3 Conclusions

Models to assess transgene introgression into natural populations need to incorporate ecological principles in a biologically meaningful manner. Concurrently, variation is ubiquitous in natural populations and contributes to uncertainty in the resultant estimates of population level parameters. In this study we show that accounting for the uncertainties using the non-parametric bootstrap can generate more realistic predictions of transgene fate. We also illustrate extensions to modeling dependencies among fitness components using copulas that can enhance predictions by incorporating biological relationships. In the event that data collected is not amenable to estimate the copula parameters, this can still be a powerful tool to assess the effect of different types of correlation patterns e.g. in sensitivity analysis. Additionally, demographic stochasticity can also be incorporated in the model (see Chapter 1) and while this will make predictions even more uncertain, this is certainly more realistic. Further extensions are possible such as linking forecasting data on experimental drivers with elements of the transition matrices (Gotelli and Ellison, 2006), thus emulating long-term environmental change. Thus evaluation of uncertainties can lead to methodological enhancements, in both data collection and modelling approaches, paving the way for effective decision-making and policy.

Wildty
Fecun
Fertilit
Viabili
Sexual
Maturi
λ
Mating

Transg
Fecund
Fertilit
Viabili
Sexual
Maturit
λ

Table 2.1
of increa
was domi
model.

Wildtype	2.5% Per.	1st Qu.	Median	Mean	3rd Qu.	97.5% Per.	std. dev	%CV
Fecundity	16.651	19.920	22.380	22.670	25.020	30.435	3.648	16.090
Fertility	0.465	0.474	0.478	0.478	0.483	0.491	0.007	1.416
Viability	0.991	0.992	0.992	0.992	0.992	0.992	0.000	0.032
Sexual Maturity	65.709	67.120	67.920	67.840	68.610	69.808	1.066	1.571
λ	1.036	1.040	1.042	1.042	1.044	1.048	0.003	0.300
Mating Success	rep1: 1.16	rep2: 1.93	rep3: 5					

Transgene	2.5% Per.	1st Qu.	Median	Mean	3rd Qu.	97.5% Per.	std. dev	%CV
Fecundity	3.840	5.616	6.779	6.838	7.892	10.536	1.723	25.205
Fertility	0.370	0.391	0.402	0.402	0.413	0.434	0.016	4.061
Viability	0.984	0.985	0.986	0.986	0.986	0.987	0.001	0.070
Sexual Maturity	90.970	94.710	96.430	96.370	98.180	101.500	2.709	2.811
λ	1.009	1.019	1.020	1.020	1.022	1.024	0.004	0.358

Table 2.1: Summary statistics of the bootstrap estimates of fitnesses and λ (the finite rate of increase) for the Wildtype (WT) and transgenic (TG) genotypes. The TG (*w* allele) was dominant in our case and this assumption was made for all fitness components in the model.

P
T
F
V
F
S
M
A
C
F
V
F
S
[illegible]
[illegible]

Table 2.2

of the W

ratio TG

parametr

same joi

ing in re

Parameter TG/WT	2.5%	1st Qu.	Median	Mean	3rd Qu.	97.5%	std.dev.	%CV
Fertility	0.7697	0.8175	0.8403	0.8414	0.8655	0.9091	0.0364	4.3253
Viability	0.9922	0.9932	0.9937	0.9937	0.9943	0.9952	0.0008	0.0772
Fecundity	0.1601	0.2400	0.2991	0.3095	0.3674	0.5282	0.0929	30.0009
Sexual Maturity	0.6630	0.6901	0.7042	0.7046	0.7192	0.7510	0.0224	3.1783
λ	0.9662	0.9764	0.9789	0.9783	0.9810	0.9850	0.0046	0.4688
Copula								
Fertility	0.7645	0.8133	0.8401	0.8413	0.8696	0.9242	0.0404	4.7988
Viability	0.9922	0.9931	0.9937	0.9937	0.9942	0.9952	0.0008	0.0783
Fecundity	0.1685	0.2454	0.3037	0.3178	0.3719	0.5655	0.1012	31.8354
Sexual Maturity	0.6678	0.6924	0.7064	0.7070	0.7208	0.7468	0.0208	2.9450
λ_C	0.9737	0.9773	0.9791	0.9791	0.9809	0.9844	0.0028	0.2867

Table 2.2: Summary statistics for the relative fitnesses and λ (the finite rate of increase) of the WT and TG genotypes for all scenarios. The relative fitnesses are based on the ratio TG/WT. Although, the average values are the same for the Copula and the Non-parametric bootstrap approaches, the distribution of the relative values are different. The same joint distribution of all the estimated components were used for the scenarios differing in relative mating success of the WT males.

3a). Summary Statistics of Fixation times

WT Mating Success	2.5%	1st Qu.	Median	Mean	3rd Qu.	97.5%	Std.dev	%CV
1.16	5.677	11.300	14.680	20.600	21.220	77.561	21.186	102.847
1.93	3.949	5.956	6.469	6.727	7.204	10.441	1.630	24.233
5	2.410	3.038	3.230	3.225	3.435	3.960	0.356	11.036
1.93(C)	5.324	6.036	6.480	6.764	7.148	9.965	1.227	18.145

3b). Correlations among relative Fitness Components

WT mating Success 1.93	Fertility	Viability	Sexual Maturity	Fecundity
Fertility	-	-0.00921	0.00035	0.00088
Viability	-	-	-	0.05483
Sexual maturity	-	-	-	-0.04022
Fecundity	-	-	-	-
WT mating Success 1.93 (C)	Fertility	Viability	Sexual Maturity	Fecundity
Fertility	-	-0.4622499	-0.4597631	0.015746
Viability	-	-	-	-0.37283
Sexual maturity	-	-	-	-0.385899
Fecundity	-	-	-	-

Table 2.3: Summary statistics of the fixation times for the WT genotypes. Fixation is defined as the time where the TG gene frequency < 0.001. The Scenarios indicate the relative mating success of the male WT genotype and the scenario suffixed with (C) was implemented using the Copula approach.

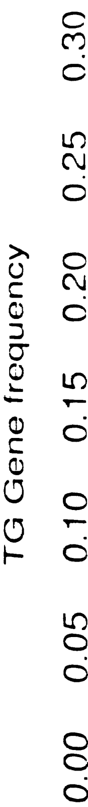

Figure 2.
based on
three dif
The traje
approach
text shou
average o
plementir
the WT g

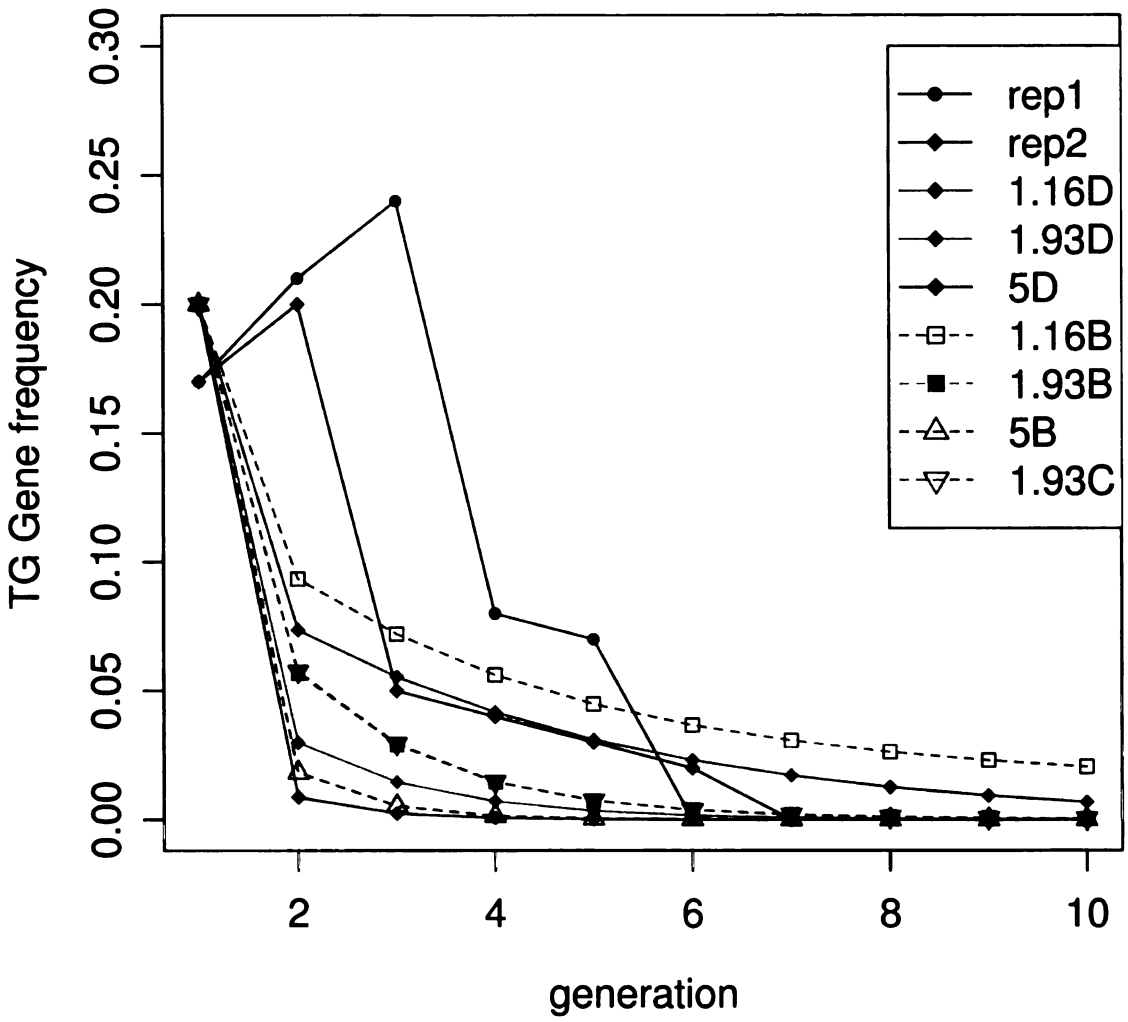

Figure 2.1: Evolution of the TG gene frequency over time for the deterministic models based on the expected values as well s the average over all bootstrap replications for the three different values of mating success of WT males (1.16, 1.93 and 5 respectively). The trajectory for the average over the bootstrap replications implementing the Copula approach overlays the trajectory for the corresponding independent approach. Legend text should be interpreted as follows: D - represents deterministic models, B - represents average over bootstrap replications, C - represents average over bootstrap replication implementing dependencies; 1.16, 1.93 & 5 - the numbers represent the mating success of the WT genotype in that scenario.

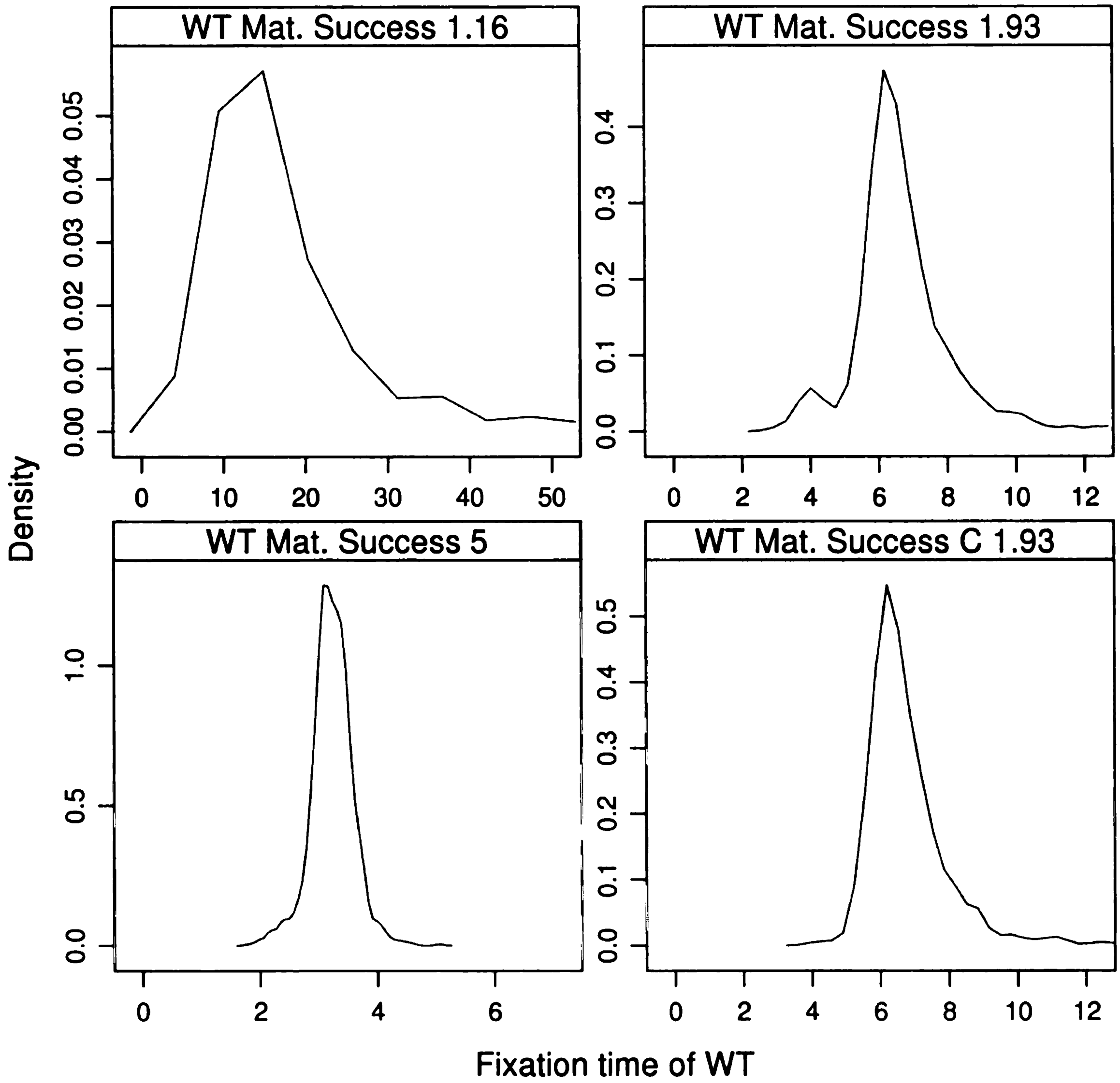

Figure 2.2: Distribution of the time to fixation of the Wildtype (WT) genotype defined as the time the TG allele frequency is < 0.001 for the three different mating success values of the WT genotype (1.16, 1.93 and 5 respectively).

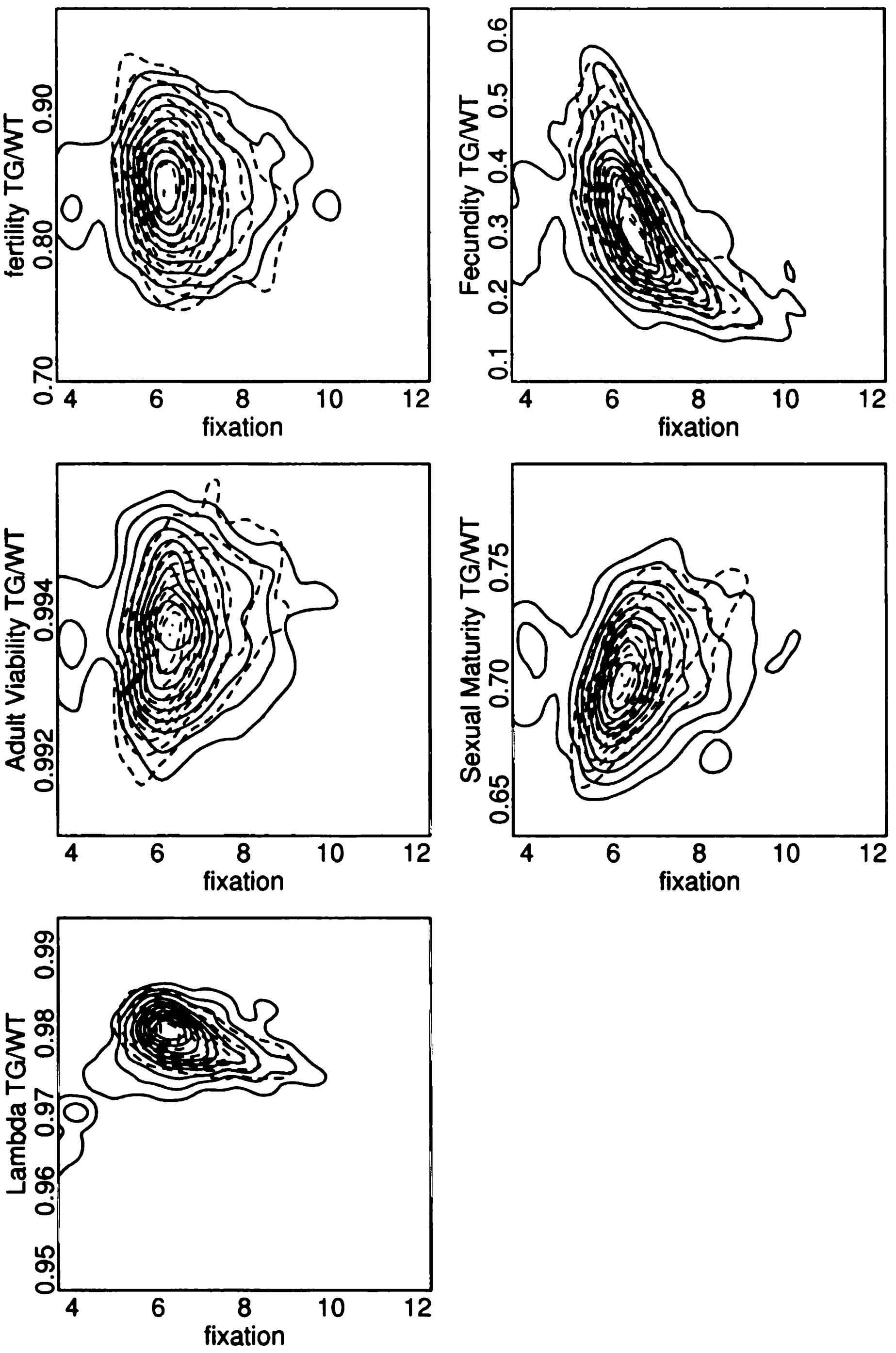

Figure 2.3: Distribution of the fixation times versus the distribution of relative fitness components and λ. Contours represent regions of the joint distribution using 2D kernel density estimates. Contours in red indicate the distributions in the scenario incorporating dependence using copula, while the contours in black represent the corresponding bootstrap scenario (i.e relative WT male mating success=1.93).

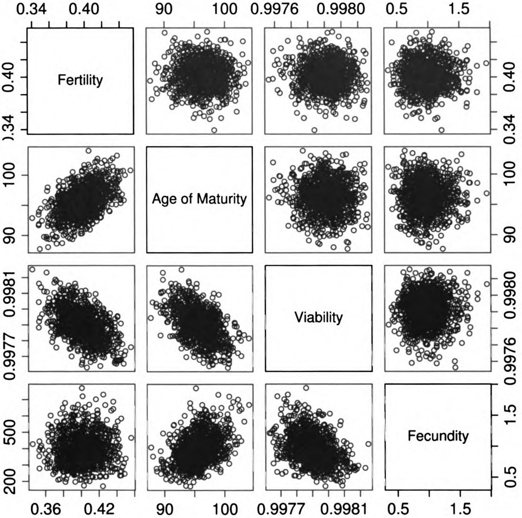

Figure 2.4: Pairwise scatterplots of the parameters showing the relationship between parameters under the independent bootstrap and the copula approach. The set of scatter plots above the diagonal show parameter combinations under independence. The set of scatterplots below the diagonal show dependence induced among the possible parameter combinations based on the specified correlations, here -0.5 between viability and all others, 0.5 between fecundity, fertility and the age of sexual maturity and 0 between fecundity ad ferility as they are among different sexes.

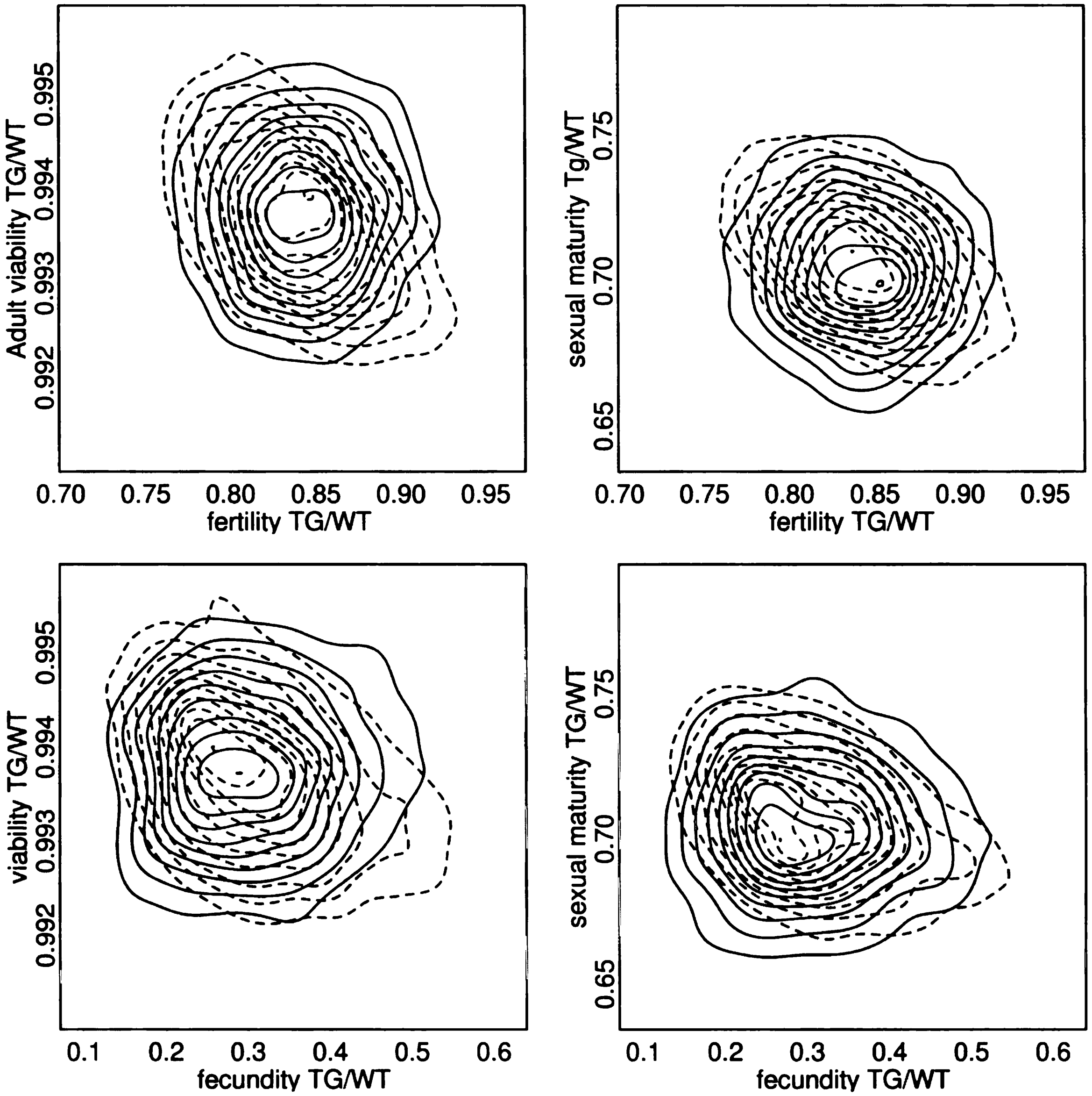

Figure 2.5: Comparison of *relative* fitness contours between bootstrap and copula generated parameter distributions. Contours represent the joint distribution using a 2D kernel density estimate. Contours connected by solid lines indicate the distributions in the scenario incorporating dependence using copula, while the contours drawn with dotted lines represent the corresponding bootstrap scenario (i.e relative WT male mating success=1.93). The contour levels indicate probabilities from 0.1 to 1 with a 10% difference in probability among adjacent levels.

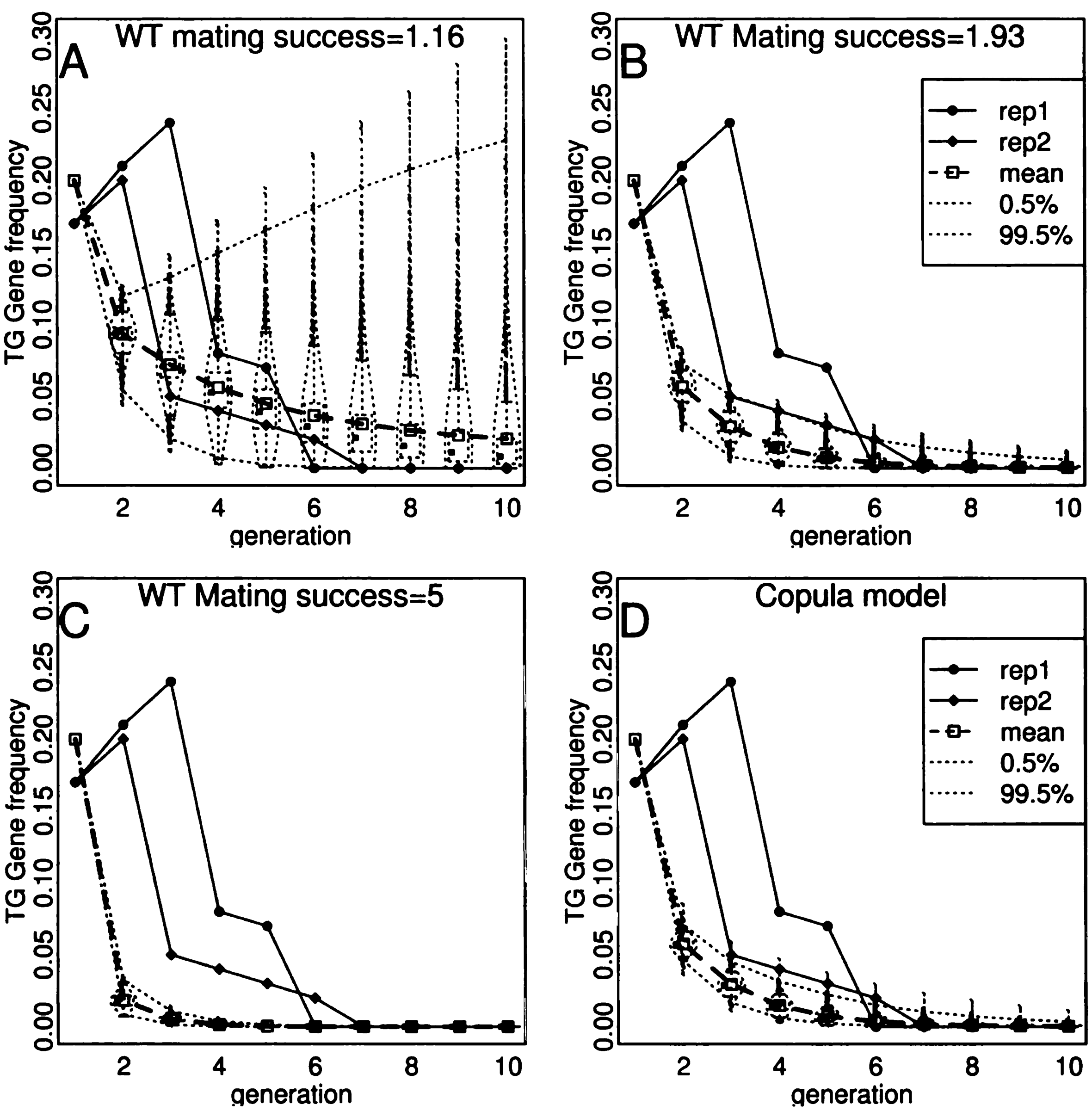

Figure 2.6: Comparison of the trajectories of gene frequency of the validation experiment from those generated from the model predictions, with the associated 99th percentiles of the prediction distribution at each generation. A) Represents the prediction distribution when the WT has a relative mating advantage of 1.16 B & D) the relative mating advantage of the WT =1.93 C) WT relative mating advantage =5 and D) Additionally, incorporates dependence using a copula approach. Legend text in the plots are: rep1 & rep2 - Mesocosm replicate,mean - Mean over 1000 bootstrap replicates, 0.5% & 99.5%- the 99% intervals over the bootstrap replications.

Chapter 3

Probabilistic Sensitivity Analyses of Models for Ecological Risk Assessment: A Case Study Using the Net Fitness Component Model for Assessing Risk of Genetically Engineered Fish

3.1 Introduction

Modern ecological risk assessments include modeling as a fundamental component to predict population trajectories and to quantify risks (Barnthouse and Sorenson, 2007). In the past decade advances in computing technology has promoted the emergence of simulation models as a primary tool utilized in this respect (e.g. Wisdom et al., 2000). Simulations models can be seen as another kind of experimental system in ecology and evolution, wherein a complex model formulation mimics the behavior of natural systems. This can help researchers investigate systems behavior by manipulations that would be impossible

otherwise, either logistically or ethically (Peck, 2004). For example, using simulations based on digital organisms, with computer programs that self-replicate, mutate, compete and evolve, it was shown how complex functions can originate by random mutation and natural selection over approximately 15,000 generations (Lenski et al., 2003).

Population level ecological risk assessment of anthropogenic impacts in natural environments has advanced towards using a combination of modeling informed through field data. While these models can be complex, there is impetus towards integrating both the genetics and demographic aspects through modeling approaches to better understand these impacts (Nacci and Hoffman, 2008). Similarly, ecological risk assessment for genetically engineered organisms (GEOs) is no different and employs models of varying complexity towards quantification of risks. GEOs that are considered to have potential ecological impacts are species that have natural wild populations and are difficult with regards to containment of the transgenes (e.g. plants, fish and insects). The last decade has seen the emergence of a variety of modeling approaches towards predicting the fate of the transgene introgression into natural populations with varying degrees of complexity. Models of introgression for transgenic crops evaluate gene flow based on pollen dispersal, for example, incorporating exponential distance and directional effects of wind mediated pollen dispersal (Meagher, 2003) or transport equations of atmospheric physics coupled with population dynamics and genetic models (Richter and Seppelt, 2004). Additionally, structured population models combining dispersal (Garnier and Lecomte, 2006) and environmental variability (Claessen et al., 2005) have also been employed in this context. Recently, AMELIE, a modeling framework combining both population structure, stochasticity and dispersal has been proposed for transgenic crops (Kuparinen and Schurr, 2007). Structured population models have also been used for assessing the risk of transgene introgression in GE fish populations (Muir and Howard, 2001) and with stochasticity adding further complexity (see Chapter 1).

The above models are examples of the increasingly complex models utilized for prediction in assessing risks associated with potential environmental consequences of transgene introgression into natural populations. The consequence of increased model complexity are twofold: the number of parameters in the model increases and the relationships among variables need not be linear. For example, in the proposed AMELIE (Kuparinen and Schurr, 2007) framework, the most sophisticated yet for transgenic crops, there are 29 parameters, and our simplified model has 10 parameters. The first step towards making sense of system behavior under these circumstances lies in extracting the relationship between variation in parameter values and the resultant variation in model output, known as uncertainty analysis (Saltelli, 2000). The next step is to examine the contributions of the parameters to the variation in the output and their relative importance, also known as sensitivity analysis. A wide array of methods exist for both local and global sensitivity analysis, as well as an extensive literature and one recent comprehensive review can be found in Saltelli (2000). Local sensitivities have been widely used for simple models, which usually can be analytically derived as they are derivative based, and thus are not computationally intensive. For example, elasticity analysis performed for population viability analysis (PVA, Caswell, 1989; Morris and Doak, 2002) using stage-structured models are local methods, however, this is simple only if the response is linear and monotonic. Global sensitivity analysis (GSA) has recently become prominent using variance-based techniques (an ANOVA like decomposition of the output variance) and provide insight on the relative importance of factors in complex model settings. Briefly, let $y = f(x_1, x_2 \ldots x_k)$ denote the output of a simulation model with $x_i\,(i = 1, 2, \ldots k)$ the input variables, with an joint probability distribution giving rise to uncertainty in y. If the input factors are orthogonal then the variance of y, $V(Y)$, for a model with k factors can be decomposed as:

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \ldots + V_{12\ldots k}$$

where $V_i = V(E(Y|X_i))$, $V_{ij} = V(E(Y|X_i, X_j)) - V_i - V_j$ and so on. A model with only V_i terms is called additive and the first order sensitivity indices are defined as $S_i = \frac{V_i}{V(Y)}$, which explains the amount of variance due to factor *i*. If the model is completely additive then $\sum_i S_i = 1$, else we can use another measure the total sensitivity index TI_i defined as:

$$T_i = \left(1 - \frac{V(E(Y|X_{-i}))}{V(Y)}\right) = S_i + \sum_{j \neq i} S_{ji} + \sum_{j \neq i, k \neq i} S_{ijk} + \ldots$$

which is the amount of variance explained by the factor X_i and all other interactions involving that factor. Both sensitivity indices have a quite general applicability, since they apply to a very wide range of non-linear models i.e. they are "model-free" and rely on only a few assumptions, namely the output has to be square integrable and the variance is an adequate measure of dispersion. An accessible introduction to GSA using variance based techniques can be found in Saltelli et al. (2008) as well as in two recent reviews introducing this approach to ecological applications (Cariboni et al., 2007; Fieberg and Jenkins, 2005).

Recent rapid technological advancements, both in computing and data generation (e.g. genomics), has facilitated wider use of complex models for biological applications and this is also true of applications in population ecology (e.g. PVA). However, increase in model complexity results in increased computational time, both in terms of computational burden as well the parameter combinations used. In the area of systems engineering complex computer models are the rule rather than the exception, especially as experimental evaluations are costly. For example aerospace applications (Gramacy and Lee, 2008) involve a large number of parameters during the prototype development phase and simulations as experiments are heavily utilized. This field is mature with formal methods for the design and analysis of computer experiments (e.g. see Fang and Sudjianto, 2006; Santner et al., 2003). Meta-modeling of the computer model is an approach which is used to increase efficiency when using computer experiments, known as an *emulator* in the literature. This involves using a *surrogate model* to simulate the computer model itself and is done using a wide variety of statistical regression techniques, from polynomial regression to spline smoothing methods (Fang and Sudjianto, 2006). The standard Bayesian approach to meta-modeling in the literature is to use a stationary Gaussian Process (GP)(Santner et al., 2003; Oakley and O'Hagan, 2004). The GP is conceptually straightforward and it easily accommodates prior knowledge in the form of covariance functions for the model output, as well as for input distributions (Oakley and O'Hagan, 2002), and returns estimates of predictive confidence. Mathematically, GPs are equivalent to many well known models (e.g. Bayesian linear models, spline models, large neural networks; see Rasmussen and Williams, 2005) and have also seen use in geostatistics as *kriging* and environmental modeling (e.g. Wikle et al., 1998). GP models may be easier to handle and interpret than, for example, neural networks (Plate, 1999).

The Bayesian framework is advantageous in that it is flexible to incorporate model extensions as well combine outputs across multiple models, e.g. using hierarchical modeling for population ecology (Clark, 2003) or Bayesian model averaging (BMA; Gelman, 2004). With increased computational power Bayesian approaches are becoming popular, and also recommended due to their flexibility, for assessments in conservation (Wade, 2000), population modeling (Millar and Meyer, 2000), population genetics (Beaumont and Rannala, 2004) and evolutionary biology (OHara et al., 2008).

In this study we apply Bayesian Gaussian Process (BGP) models for sensitivity analysis of a risk assessment model for transgene invasion, the net fitness component model

(Muir and Howard, 2001, 2002). We use experimental and simulation data from a previous study (see Chapter 2) to conduct GSA for the model. We use both the Monte Carlo approach (Saltelli et al., 2008) and the BGP model approach (Gramacy, 2007; Oakley and O'Hagan, 2004) for GSA analysis and evaluate the impact of sampling designs for parameter selection, specifically random sampling and Latin hyper-cube sampling. Given that the number of model evaluations are reduced using the BGP model, we also conduct a GSA by extending the previous dataset with two additional variables, the dominance effect of the transgene and the mating success of the wildtype, as a case study. We had not done this previously as there was insufficient information from the experimental data for those variables to estimate their underlying distribution in the population. We illustrate the flexibility of the BGP model by evaluating the effect of the uncertainty distributions for these two variables using GSA of the model.

3.2 Materials and Methods

3.2.1 Gaussian process modeling

The model description closely follows that of Gramacy (2007). Let **X** be a input matrix of k covariates with n observations and a single response variable Z. Then the hierarchical generative model for the stationary GP is as below:

The model output Z(response) depend on multivariate inputs **X** (explanatory variables) as

$$Z|\beta, \sigma^2, \mathbf{K} \sim N_n(\mathbf{F}\beta, \sigma^2\mathbf{K}) \qquad \sigma^2 \sim IG(\alpha_\sigma/2, q_\sigma/2)$$

Where β are linear trend coeffecients and $\sigma^2\mathbf{K}$ is the covariance of a mean-0 process,**K** is an correlation matrix and β is independent of $\sigma^2\mathbf{K}$ in the prior. Using a hierarchical approach the mean function can be modeled as

$$\beta|\sigma^2, \tau^2, \mathbf{W}, \beta_0 \sim N_m(\beta_0, \sigma^2, \tau^2, \mathbf{W}) \qquad \tau^2 \sim IG(\alpha_T/2, q_T/2)$$

$$\beta_0 \sim N_m(\mu, \mathbf{B}) \qquad \mathbf{W^{-1}} \sim W((\rho(V^{-1}), \rho)$$

$\mathbf{F} = \{1\ \mathbf{X}\}$ can be used to add covariates to the mean function, $\mathbf{W}$ is an $k \times k$ matrix representing the covariance of the mean function; N, IG and W are the normal, Inverse-Gamma and Wishart distribution. Constants μ, $\mathbf{B}$, $\mathbf{V}$, ρ, α_σ, q_σ,α_T, q_T are treated as known. The GP correlation function $\mathbf{K}$ is chosen from a seperable power family, thus assuming anisotropic correlations, with fixed power p_0, but known range and nugget parameters. Correlations take the form $\mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) = K^{*}(\mathbf{x_i}, \mathbf{x_j}) + g\delta_{i,j}$ where g is the nugget, $\delta_{i,j}$ is the Kronecker delta function and K^* is a true correlation representation from a parametric family as follows:

$$K^*(\mathbf{x_i}, \mathbf{x_j}|\mathbf{d}) = exp\left\{-\sum \frac{|x_{ki} - x_{kj}|^{p_0}}{d_k}\right\}$$

The implementation for the markov chain monte carlo sampling (MCMC) involves a Metropolis within Gibbs approach. The derivation of the full conditionals and the implementation of the MCMC is given in detail in Gramacy and Lee (2008) and all parameters except those within the $K(\cdot, \cdot)$ function can be sampled using the Gibbs step.

3.2.2 The net fitness component model

We used an age structured model with three genotypes (WW , Ww *and* ww), the Wild-type, heterozygote and the transgenic homozygote, respectively. The details of the model implementation is given in Chapter 1 and appendix 1. Hereafter the WW genotype will be referred to as the WT genotype and the *ww* genotype will be referred to as the TG genotype respectively for notational simplicity.

3.2.3 Monte Carlo simulations

We use three different sets of simulations to conduct GSA using the Monte Carlo approach. Previously, we used experimental data to evaluate uncertainty in estimates of fitness components (refer Chapter 2) and generated samples from the joint uncertainty distributions of the estimated fitness components. This was done using Monte Carlo approaches for each fitness component independently by non-parametric bootstrap. We use the data set from the non-parametric bootstrap for estimating sensitivity indices and this data set will hereafter will be referred to as MC-data. The data set had already been modified for a reduced age-structure using the genetic algorithm approach. The MC-data consisted of 1000 samples and we split it into two sets of 500 samples each to estimate the first order sensitivity indices (hereafter SI) and the total sensitivity indices (hereafter TI) using the Monte Carlo approach. Let A be the matrix of $N_1 \times k$ inputs, where N_1 represents the first 500 samples, and B the matrix of the other 500 samples and let ${}_{A}y_i$ represent the model output using the *i*th row of the A matrix, and similarly for ${}_{B}y_i$. Finally, we create a new matrix $C^{(i)}$ for each input variable x_i which contains the columns for that variable from A and all other columns from B and we denote the model output as ${}_{C}y_i$. Following Saltelli et al. (2008) we can estimate the SI_i as follows:

$$SI_i = \frac{Var(E(Y|X_i))}{V(Y)} = \frac{{}_{A}y \cdot {}_{C}y^{(i)} - \hat{f_0}^2}{{}_{A}y \cdot {}_{A}y - \hat{f_0}^2} = \frac{(1/N)\sum_{i=1}^{k} {}_{A}y_i \, {}_{C}y_i - \hat{f_0}^2}{(1/N)\sum_{i=1}^{k} {}_{A}y_i \, {}_{A}y_i - \hat{f_0}^2}$$

where the SI_i is the first order sensitivity index for variable x_i and is expressed as a fraction of $V(Y)$ that is explained by that variable;${}_{A}y$, ${}_{B}y$ and ${}_{C}y$ represent the vector of outputs for matrices A,B and C respectively. Thus the SI_i's range from 0-1. The TI$_i$ for an input i is calculated as follows from the evaluations of matrix B and C:

$$TI_i = 1 - \frac{Var(E(Y|X_{-i})}{V(Y)} = \frac{y_B \cdot y_C^{(i)} - \hat{f}_0^{\,2}}{y_A \cdot y_A - \hat{f}_0^{\,2}} = \frac{(1/N)\sum {}_B y_i \, {}_C y_i - \hat{f}_0^{\,2}}{(1/N)\sum {}_A y_i \, {}_A y_i - \hat{f}_0^{\,2}}$$

where the TI_i is the total sensitivity index for variable i and represents the sum of variances that includes the first order and all higher order interactions that involve the variable. If TI_i is 0 then it implies that the variable is non-influential with regards to the output. If it is close to the SI_i then it implies that the interactions with other variables are negligible in their influence on the output and the role of the variable is mainly additive.

Since the TG genotype is dominant in our case, both the transgenic homozygotes and heterozygotes have the same values for fitness components, and thus there were only ten input variables to the model. They were the five fitness components for the WT and TG genotypes, namely: fertility, fecundity, juvenile viability, adult viability and age of sexual maturity. We had to do a total of 500 × 10 + 2 evaluations to estimate all the SI and TI, and this will hereafter be referred to as MC2-data. Additionally, we used Latin hyper cube sampling to generate a space filling design of 2 sets of 200 samples each,for the 10 input variables. The combined data set of 400 samples will hereafter be referred to as the LHS-data. To sample from the marginal distributions of the fitness components we used an inverse transformation of the cumulative distribution function. We used this design to estimate the SI and TI, using the same protocol above, resulting in a total of 200 × 10 + 2 simulation, hereafter referred to LHS2-data.

3.2.4 Bayesian Gaussian Process (BGP) Approach

We used the BGP approach to fit a meta model to the output distributions using the following samples: a random sample of 300 observations each from the MC-data and the LHS-data. The models will hereafter be referred to as GP-MC and GP-LHS, respectively.

The data samples from the MC-data were used as the training samples and the remaining samples used as a test set to evaluate the predictive performance based on residual mean square error (rmse). We also used samples from the other data sets created as test sets to evaluate the performance of the GP model predictions for both the GP-MC and GP-LHS models. For example, we evaluated the predictions of the GP-MC with respect to the MC2-data as well as the LHS-data.

As a case study we evaluated the risk of transgene introgression using the BGP approach. We created a third simulation data set which included the dominance of the TG genotype and the mating success of the WT genotype as additional variables, hereafter referred to as FULL-data. The dominance effect of the TG genotype was chosen from an Uniform (0,1) distribution, and the mating success of the WT genotype from an Uniform (1,5) distribution. Simulations were conducted and the GP model, hereafter referred to as GP-FULL, was used for conducting global sensitivity analysis of the FULL-data.

Evaluations of global sensitivity indices, both first order and total, is simple once the GP model has been fit and we calculated the SI and TI for all data sets above for comparison. Additionally, using the GP-FULL model we calculated the sensitivity indices using two forms of uncertainty distributions of the dominance and the WT mating success variables. The first is the default uniform distribution and the second form for the uncertainty distributions used a scaled beta that reflects our prior information on the marginal distributions of those parameters. For both variables, we evaluated uncertainty distributions with modes close to the two extremes of the range of the data. For the dominance effect of the transgene, which ranged from 0-1, we used modes at 0.1 and 0.9 respectively. For the WT mating success, we used modes at 1.5 and 4.0 respectively. Since the GP approach used here is a fully Bayesian approach using MCMC, we also used diagnostics to assess convergence of the MCMC for the final models.

For all analysis we used the R statistical environment (Team, 2009) and the appropriate packages. BGP model fitting was done using the `tgp` package (Gramacy, 2007). The tgp package is extremely useful and provides functions for sensitivity analyses which lets us specify the uncertainty distribution for the matrices used to calculate the sensitivity indices. The sens functions were used to generate the global sensitivity indices, both first order and total. Additionally, the lhs package was used to generate the Latin hyper cube samples for the LHS-data. As the resulting values are on the unit hypercube, the inverse CDF of the marginal distributions from the MC-data was used to obtain the input values for LHS-data. We use the coda package to evaluate the convergence and serial correlation of the MCMC.

3.3 Results

3.3.1 Global Sensitivity analysis: Monte Carlo approach

The variance in the output as well as the estimated first order sensitivity indices (SI) and total sensitivity indices (TI) are given in Table 3.1 for all analyses using the MC-data and the LHS-data. For the MC-data the SI and TI indices are not within the admissible range (0-1) (Figure 3.1A and B), which occurs because the variance in output is not correctly estimated by the random sampling design (MC-data $var(y)$ = 4407.818 vs MC2-data $var(y)$ = 54210.4). The joint distribution from the random sampling does not encompass regions of parameter space that have very low probability and therefore would necessitate a larger number of runs. Even using the estimates of variances from the MC2-data runs does not improve results much. The SI for the MC2-data identifies the juvenile viability of the TG genotype as the most influential factor (Figure 3.1C). A space-filling design (e.g. the Latin hyper cube LHS-data) samples the entire parameter space and gives a much better estimate of the output variance (LHS-data $var(y)$ = 349242). For the LHS-data, the SI indices ranked age of sexual maturity of the WT genotype as the most influential input followed by juvenile viability and fecundity of both genotypes (Figure 3.1E). The

TI indices indicate that the effect of all the components identified as influential by the SI are not additive and that there are interactions among them (Figure 3.1B, D and F). All the components show higher order interactions as measured by the difference TI$_i$-SI$_i$, even factors that do not have any first order effects. Adult viability and fertility of both genotypes are only influential as interactions with the other components as measured by the difference in TI and SI. The TI indices are relatively consistent for both the MC-data and the LHS-data, however the SI indices vary among the two sampling designs.

3.3.2 Bayesian Gaussian Process (BGP) models

3.3.2.1 BGP Predictions

The GP approach is able to produce a near perfect fit for the MC-data as well as the LHS-data(rmse=0.0004 and rmse=0.00029 respectively) for the data on the transformed scale. For the MC-data the GP-MC model produced good predictions for the test sets from the MC-data and the MC2 data (rmse=25.3639, Figure 3.2C; rmse=56.4734, Figure 3.2E). The predictions are not very accurate for the LHS-data outside the range of training samples (rmse=487.5 Figure3.2D). However, when the input parameters were within the range spanned by the training samples the GP-MC predictions are accurate. For the LHS-data the GP-LHS model produced good predictions for both test sets, MC-data and MC2-data (Figure 3.3C and D) and the absolute deviations were < ±40 units for about 90% of the original values (Table 3.2). The GP-LHS model predictions are not very accurate at the upper limit of the fixation times, even in the case of the training set (Figure 3.3A and B). This is mainly caused by the effect of the reduced age structure resulting in reduction of age of sexual maturity for the WT genotype to only two values (9 and 10 weeks), with a very low probability for the larger age.

3.3.2.2 Global Sensitivity analysis: Gaussian Process approach

SI indices for the GP-MC model and the GP-LHS models show that the juvenile viability component, for both WT and TG genotypes, had the largest first order effect with the TG genotype having the highest effect (Figure 3.4; Table 3.1).

For the MC-data the GP-MC also ranks the fecundity of both genotypes and the sexual maturity of the TG genotype higher than the other components (Table 3.1). Additionally, we can estimate the main effects for the fitness components at no extra cost and deduce the relationship between individual fitness components and the response (Figure 3.4). Increasing juvenile viability of the TG increases the time to fixation of the WT, whereas increase in WT fecundity results in the decrease of the fixation time. Comparisons with the Monte Carlo approach reveal that the method is unable to detect the smaller effects, which is a function of both the sampling method used as well as the number of model evaluations. The Monte Carlo approach to calculating SI with the MC-data performs worse even though it takes a total of 6000 model evaluations compared to the GP approach for the same data which used 300 model evaluations as the test set. The total sensitivity indices reveal that all components have higher order effects and that the response is sensitive to combinations of the factors (Fig 3.4). This results is reflected in the Monte Carlo approach, however the TI estimates are much higher in that case.

For the LHS-data the GP-MC also ranks the juvenile viability components, for both the WT and TG genotypes, as the factor with the highest effect (Table 3.1; Figure 3.5). Additionally, now there is a clear indication that the age of sexual maturity and the fecundity components for both genotypes have first order effects as well (Figure 3.5). The main effects plots show the trends of the scaled response to the input. Fixation times increase or decreases alongwith the important factor identified by SI. For example, the slope of fixation time is positive with respect to TG viability. As for the MC-data the TI indices reflect that all components have higher order effects and the fertility and adult viability components are only important in interactions with the other components. Results compare much more favorably to the Monte Carlo approach with regard to the SI indices except for a difference in rank among the top three factors.

For evaluating the SI the GP methods provide more realistic estimates when the parameter space is not completely spanned, whereas the results are more consistent between approaches when the converse is true. The Monte Carlo approach suffers more from sample size compared to the GP approach. The TI estimates show differences in both magnitude and rank among the two methods, although one remarkable resemblance among both approaches is the consistency of those estimates for both data sets.

3.3.3 A Case study: Predictions for a model of transgenic Zebrafish

The GP-FULL model approach produces reasonable fit to the fixation times for the FULL-data in the transformed scale (rmse=$3.428e^{-04}$; Figure 3.6A). On the original scale the model is unable to fit three extreme observations which leads to a higher than expected rmse (rmse=171.942), although the fit improves once those observations are removed (rmse=20.717). This is the case for the test dataset as well (rmse=134.2; Fig 3.6B), where there is noticeable improvement once observations > 2000 are dropped (rmse=20.717; Figure 3.6C). Model predictions as measured by the mean absolute deviations from the true vales are approximately within ±40 days for 90% of the observations on the back-transformed scale (Table 3.1).

Sensitivity analysis of the model input variables reveals that mating success of the WT genotype is the most important factor followed by the juvenile viability and fecundity. However, all factors where the SI~0 exhibit interaction effects as revealed by the high difference in the TI-SI (Table 3.3). Changing the uncertainty distribution produced changes in the model sensitivity to the input factors and gives new insight into model behavior. Increasing the dominance of the transgene increased the SI of mating success of the WT genotype, while the opposite trend is noticed for the fecundity and juvenile viability components. The TI values are reduced at both values of dominance for all fitness components (Table 3.3, Figure 3.7). Shifting the WT mating success distributions results in increased SI values all components except for the mating success component itself which decreases under both conditions (Table 3.3; Figure 3.8). In the case of TI, shifting the WT mating success to lower values creates a corresponding reduction for most components. However, shifting mating success to the upper extreme results in the decrease of TI for most components, except for WT fecundity, fertility and juvenile viability which show increased TI values (Table 3.3). A significant result derived overall is that the WT mating success fails to have any main effect on the model when the distribution is shifted to the upper extreme (Figure 3.8 Main effects).

3.4 Discussion

Previously, we have shown that uncertainty arising from demographic stochasticity (see Chapter 1) and estimation of model parameters (see Chapter 2) can influence predictions of transgene introgression into natural populations. However, when parameters are unknown or difficult to estimate we had to make simplifying assumptions when using them in modeling due to computational and analysis constraints. For example, the experimental setup to estimate mating success in the previous study could not provide enough information on the underlying distribution for that component. It is not feasible to explore all possible candidate distributions using our simulation model and meta-modeling of computer experiments provides an attractive alternative to explore the parameter space in a reasonably efficient manner. In this study we explore the utility of using Bayesian Gaussian Process models for meta-modeling and illustrate their application for sensitivity analysis of the model.

Ecological risk assessments are based on increasingly complex modeling of ecological phenomena and model using a variety of statistical and computational approaches (Barnthouse and Sorenson, 2007), the case of risk assessment for genetically engineered organisms (GEOs) being one such example. Recent advances in computational power have fostered the use of complex models for ecological level phenomena and the possibility of using computer simulations as experimental surrogates (Peck, 2004). While the concept of using computationally demanding models are relatively new in ecological applications, this is the norm in areas such as engineering design wherein a mature literature exists towards design and analysis of simulations (Fang and Sudjianto, 2006; Santner et al., 2003). An important consideration in these areas is the development of a meta-model, an *emulator* that can be used to approximate the simulation model over most regions of the parameter space and this is usually constructed using statistical data-analytic methods. *Emulators* can therefore provide an efficient method of exploring model behavior and Bayesian Gaussian Process (BGP) models have recently become prominent in this area (Oakley and O'Hagan, 2004; Santner et al., 2003). Furthermore, the flexibility of a Bayesian approach makes these models attractive for consideration in ecological risk assessment. Our results, using the BGP approach for the net fitness component model of GEO risk assessment, show overall good performance as an *emulator* for prediction of simulation output using the fitted model (Figure 3.2, Figure 3.3, Figure 3.6 A, B and C). While the BGP models consistently had difficulty in predicting the times to fixation of the WT genotype at the extremes of the distribution, this can be attributed to the fact that this is a preliminary assessment of the model performance using the general setup of the R package `tgp`. We can expect further improvements from including modifications to our specific case and this is certainly an area for further research. Using the BGP model provides substantial improvement in the computing time to explore the parametric space as well as the scope for additional analysis such as sensitivity analysis.

A direct consequence of increased model complexity is that the number of variables and parameters underlying the model increases substantially, as well as the functional relationships among them. Model sensitivity analysis then provides insight as to what factors influence the output of interest and thus help narrow down important inputs (Saltelli, 2000). Most efforts for sensitivity analysis in ecological risk assessments have been derivative based, using analytical approaches (Caswell, 2008, 1989), and thus are confined to local perturbations of the parameters. Global sensitivity analysis (GSA), specifically variance based approaches, are attractive in that they can accommodate nonlinear/nonadditive effects and thus are "model free", although they can be computationally expensive since they use Monte Carlo methods for evaluations (Saltelli, 2000; Saltelli et al., 2008). These methods are however gaining prominence in ecological applications (Cariboni et al., 2007; Fieberg and Jenkins, 2005) as they can facilitate comparisons across studies and thus potential ability to learn collectively from multiple results. For example, a recent review of sensitivity analysis for spatial PVA's emphasized the need to standardize methods by applying GSA and develop tools for the same (Naujokaitis-Lewis et al., 2009). The Bayesian approach using BGP meta-modeling allows for calculation of the GSA sensitivity measures and sensitivity analysis to be undertaken using far fewer numbers of model runs in comparison to Monte Carlo methods (Oakley and O'Hagan, 2004). In our study, we initially used results from the previous simulations, based on a Monte Carlo approach (Saltelli et al., 2008), to conduct GSA and our estimates were not within admissible ranges (Figure 3.1), even for 500 runs of the model, whereas the BGP model produced reasonable estimates even with this data set (Figure 3.1)). Using the recommended Latin hyper cube sampling approach we arrived at reasonable estimates of sensitivity indices with the Monte Carlo approach, although this required 2400 simulations. The BGP approach provides very similar estimates using 300 simulations at a fraction of the computational cost.

An additional advantage to using the BGP model is that of prior specification on the uncertainty distribution of the input variables (Oakley and O'Hagan, 2002). This lets us use the fitted model for evaluating the effect of modifying underlying distributions when they are not available for the model parameters, without running further simulations. Previous studies for GEO risk assessment have utilized an uniform distribution within the range of the parameter space to generate random inputs (Claessen et al., 2005; Kuparinen and Schurr, 2007). For the case study of the net fitness component model, we evaluated two additional components for which we did not have experimental estimates of the underlying distribution. Results show that the modifying uncertainty distributions can result in modification of model behavior and the resultant sensitivity measures, although in a biologically meaningful manner, and thus provide additional insight on the effects of factors. For example, shifting the WT mating success distribution towards the upper extremes results in complete loss of importance for the component. This makes intuitive sense biologically, for above a certain threshold of WT mating success values the only deviation from complete loss of the transgene occurs if some other component is strong enough to counteract this effect.

The main caveat of a fully Bayesian approach, as used here, is ensuring that there is MCMC convergence.This can be problematic as inference on the BGP scales poorly with the number of data points typically requiring computing time that grows with the cube of the sample size. Storlie and Helton (2008) provide a comprehensive review of using smoothing methods such as GAM's and LOESS in the context of sensitivity analysis, though they do not provide explicit methods to conduct global sensitivity analysis. The Monte Carlo approach for sensitivity analysis is easy to implement but costly in the number of model runs required for analysis. More importantly, current methods of GSA are restricted to orthogonal or independent inputs, as there is considerable difficulty in evaluating the high dimensional integrals that arise in estimation when considering correlations among inputs. We have shown in context of risk assessment of GEOs that correlations among the fitness components exist due to to life history trade offs and have important ramifications for model predictions, and this is true more generally for most natural populations.

As illustrated by our results, the BGP approach enjoys good *emulator* properties and also provides means to conduct sensitvity and uncertainty analysis with significant savings in computation time. The `tgp` package (Gramacy, 2007) also provides additional functionality such as adaptive sampling and partitioning of the paramater space (Gramacy and Lee, 2008), which can be used for further analysis. Bayesian Gaussian Procss methods also enjoy the Bayesian advantage of being able to specify priors and thus possible model extensions using data from other sources. For example, it is possible to add an additional error term and model the noise as a second Gaussian Process, in addition to the deterministic output (Goldberg et al., 1998), and perhaps provide an avenue to model demographic stochasticity. Further extensions, within the Bayesian framework, can include combining environmental data using a time series approach for producing realistic predictions of temporal trajectories (Gotelli and Ellison, 2006)

Analysis	MC		LHS		GP-MC		GP-LHS	
Component	S.I.	TI	S.I.	TI	S.I.	TI	S.I.	TI
J_W	0	1.0505	0.1117	0.9569	0.1237	0.6036	0.1212	0.5210
J_T	0.311	0.9317	0.1552	0.9267	0.1987	0.7067	0.1661	0.6046
A_W	0	1.0987	7e^{-04}	1.063	0.0308	0.3834	0.0230	0.3918
A_T	0	1.0977	0.0018	1.0527	0.0281	0.3854	0.0277	0.3956
Fr_W	0	1.0988	8e^{-04}	1.0639	0.0255	0.3868	0.0244	0.3930
Fc_W	0	1.0979	0.1103	0.9546	0.0486	0.4420	0.0592	0.4397
Fr_T	0	1.0988	6e^{-04}	1.0618	0.0317	0.3916	0.0230	0.3975
Fc_T	0.0016	0.9767	0.0894	0.9658	0.0359	0.4286	0.0426	0.4357
S_W	0	1.071	0.2175	0.8967	0.0313	0.5109	0.1015	0.5710
S_T	0.0022	1.0264	0.0079	1.0223	0.0543	0.4441	0.0583	0.4501

Table 3.1: The mean of the first order Sensitivity indices (SI) and the total sensitivity indices (TI). The subsecript for the fitness components represent the genotypes: W- WT genotype and T - TG gneotype. The fitness components are as follows: J - Juvenile viability, A- adult viability, Fr- Fertility of male, Fc- Fecundity of females and S- age of sexual maturity.

GP-MC prediction quantiles

test set	25%	50%	75%	90%
model	$6.5e^{-05}$	0.00015	0.00033	0.00056
training	0.4390	1.0892	2.5824	7.4562
LHS-data	2.0151	6.3550	21.8629	121.9078
MC2-data	0.6864	2.0021	5.9359	20.5598

GP-LHS prediction quantiles

test set	25%	50%	75%	90%
model	$5.82e^{-5}$	0.00014	0.00026	0.00039
training	0.6382	1.5568	6.9838	38.3397
MC-data	3.3608	8.3949	15.2161	33.9908
MC2-data	3.9021	8.8574	20.8094	39.6219

GP-FULL prediction quantiles

test set	25%	50%	75%	90%
model	0.0002	0.0004	0.0007	0.0010
training	2.1915	5.1806	10.6909	24.4652
test	4.4369	9.9096	22.3818	49.4247

Table 3.2: Summary statistics for the fitted models GP-MC GP-LHS GP-FULL. Quantiles repesent absolute deviations of predicted values from the true values. The **model** represents the deviations from the fitted values on the transformed scale and the **training** set represent the deviations from the fitted values back transformed to the original scale

Fitness	Both		Dominace				Mating Success			
	Uniform		mode=0.1		mode=0.9		mode=1.5		mode=4	
	SI	TI	SI	TI	SI	TI	SI	TI	SI	TI
J_W	0.089	0.405	0.109	0.377	0.115	0.367	0.127	0.405	0.188	0.525
J_T	0.025	0.301	0.022	0.225	0.026	0.220	0.030	0.242	0.032	0.288
A_W	0.023	0.299	0.024	0.230	0.024	0.216	0.029	0.243	0.031	0.294
A_T	0.023	0.298	0.023	0.210	0.023	0.198	0.026	0.218	0.027	0.267
Fr_W	0.022	0.299	0.025	0.270	0.025	0.257	0.037	0.293	0.043	0.345
Fc_W	0.103	0.422	0.110	0.373	0.098	0.342	0.108	0.379	0.191	0.527
Fr_T	0.023	0.299	0.023	0.232	0.024	0.217	0.029	0.242	0.030	0.294
Fc_T	0.024	0.304	0.024	0.224	0.024	0.218	0.028	0.243	0.031	0.292
S_W	0.048	0.358	0.031	0.238	0.039	0.240	0.035	0.259	0.044	0.321
S_T	0.025	0.299	0.024	0.209	0.024	0.198	0.026	0.222	0.027	0.269
M_W	0.364	0.693	0.330	0.642	0.364	0.680	0.255	0.558	0.046	0.309
Dom	0.039	0.330	0.044	0.262	0.024	0.199	0.047	0.285	0.071	0.366

Table 3.3: Summary of Global sensitivity analysis for the case study. First order (SI) and total (TI) sensitivity indices are presented for the FULL-data. Uniform column represent uniform distributions for both the mating successand dominance components, while the Dominace and mating success columns represent the modifications of the uncertainty distributions.

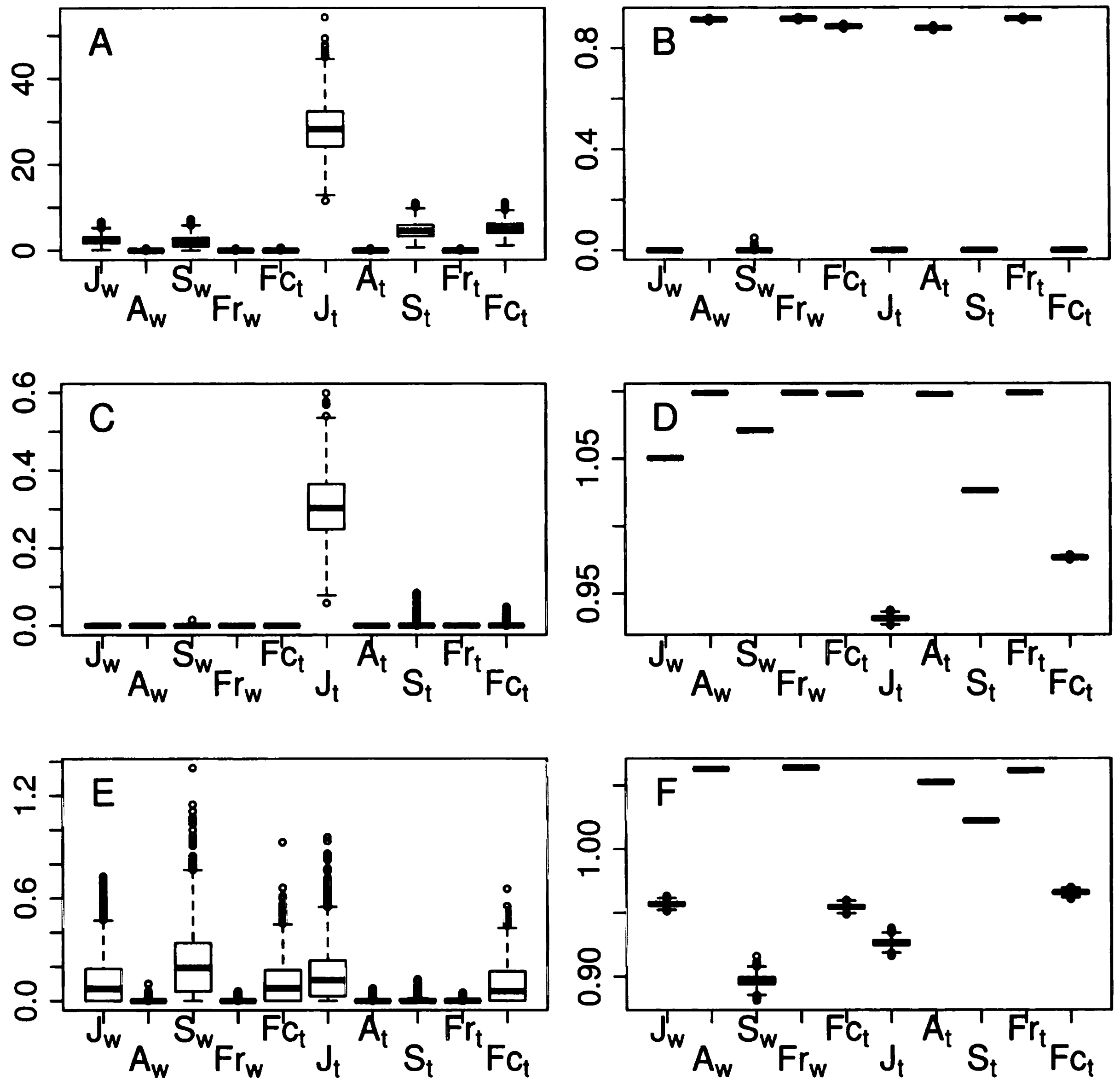

Figure 3.1: Global sensitivity analysis using the Monte carlo approach. A) and B) First order (SI) and total sensitivity (TI) indices for the MC-data respectively C) and D) SI and TI using the full variance from the MC2-data. E) and F) SI and TI for sensitivity analysis using the LHS-data.

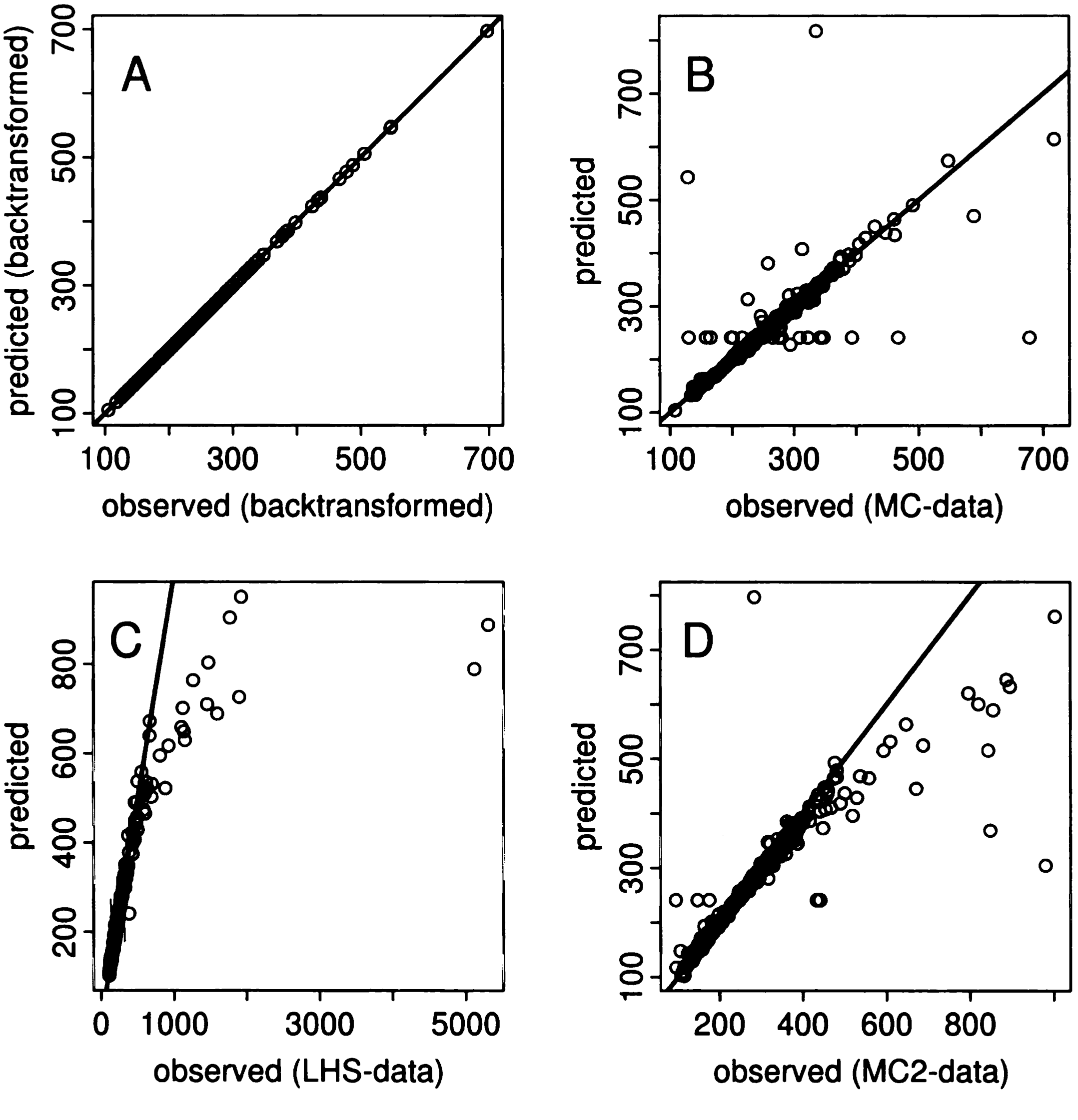

Figure 3.2: Bayesian gaussian process model predictions using the model fitted for the MC-data. A) Predictions from the GP-MC for the data used for model fitting after transformation B) Backtransformed predicted values to the orginal scale for the fitted values and the response C) Model predictions from the test set of the MC-data D) Predictions for the LHS-data as the test set E) Predictions using MC2-data as the test set F) observed response versus backtransformed response

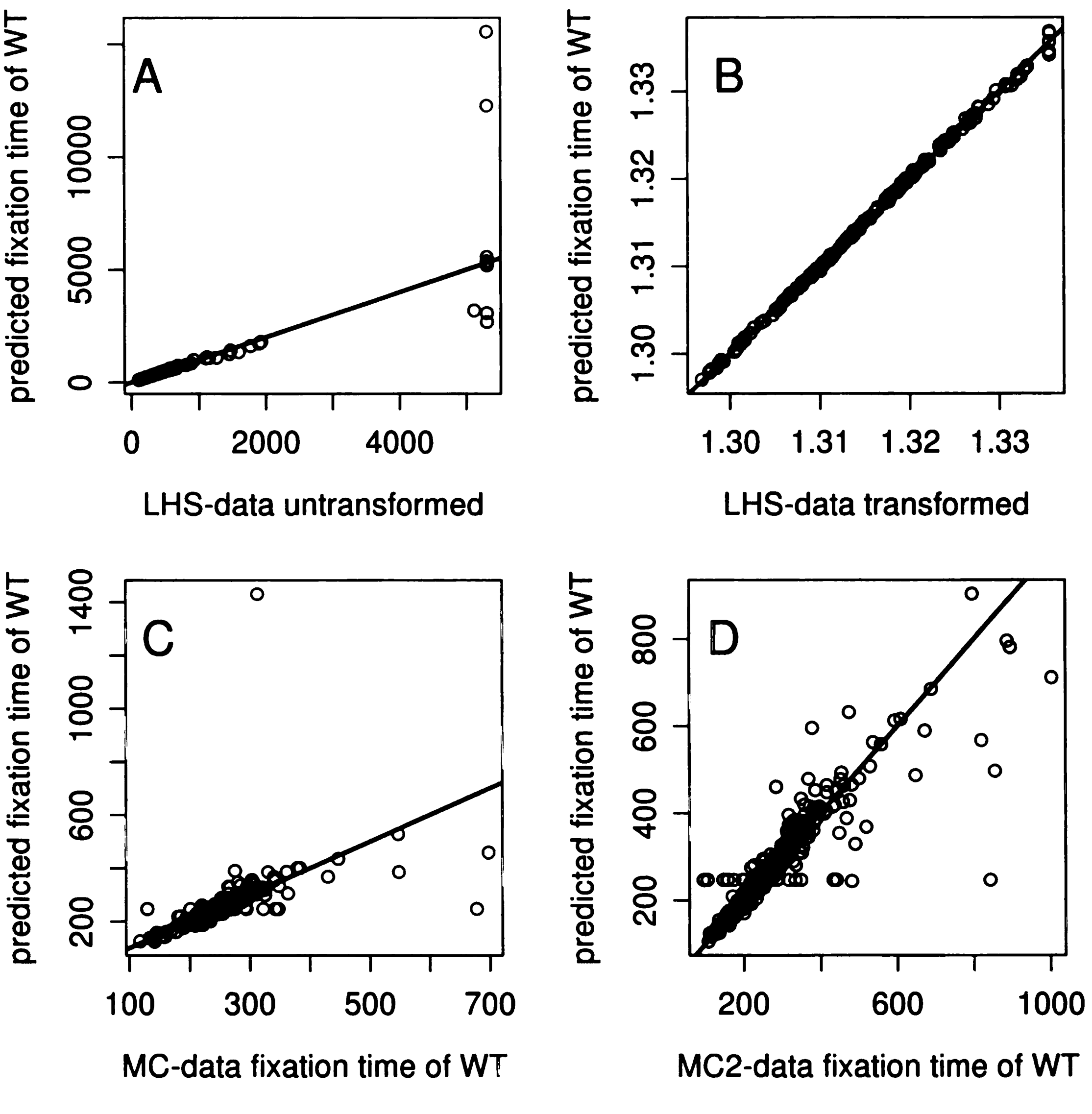

Figure 3.3: Bayesian gaussian process model predictions using the model fitted for the LHS-data. a) Predictions from the GP-LHS for the data used for model fitting after back-transforming predicted values b) Model predictions before backtransforming the response to the original scale c) Predictions using MC-data as the test set d) Predictions using the MC2-data as the test set

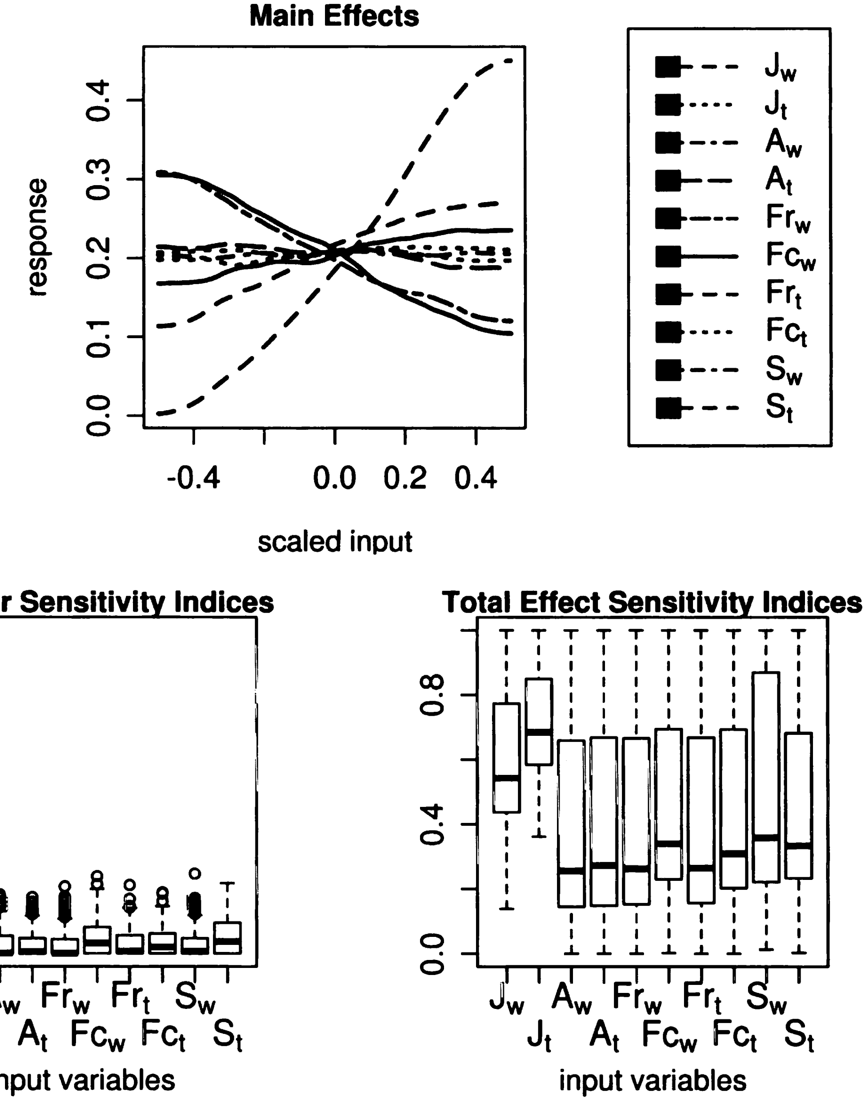

Figure 3.4: Sensitivity analysis using the Bayesian gaussian process model for the MC-data

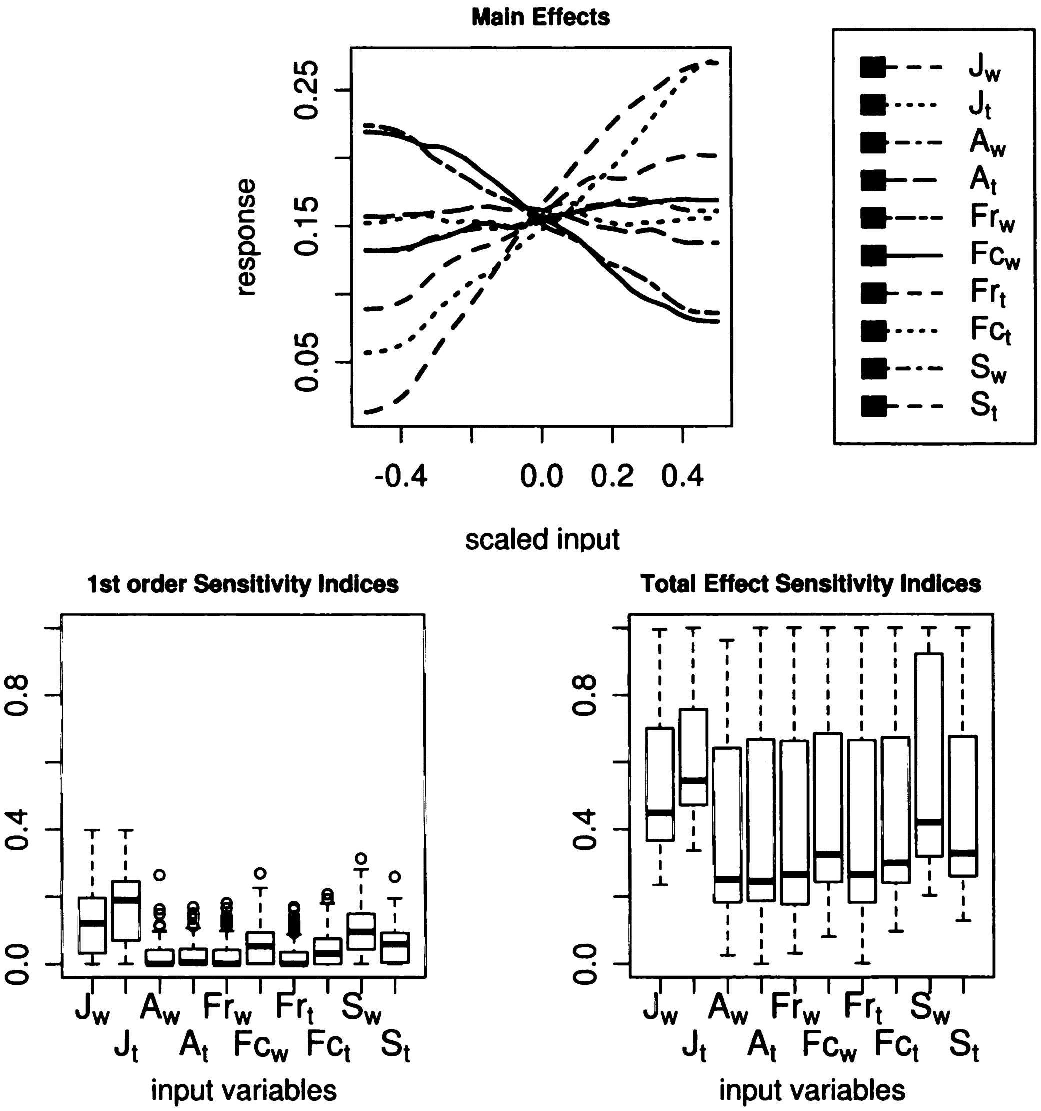

Figure 3.5: Sensitivity analysis using the Bayesian gaussian process model for the LHS-data

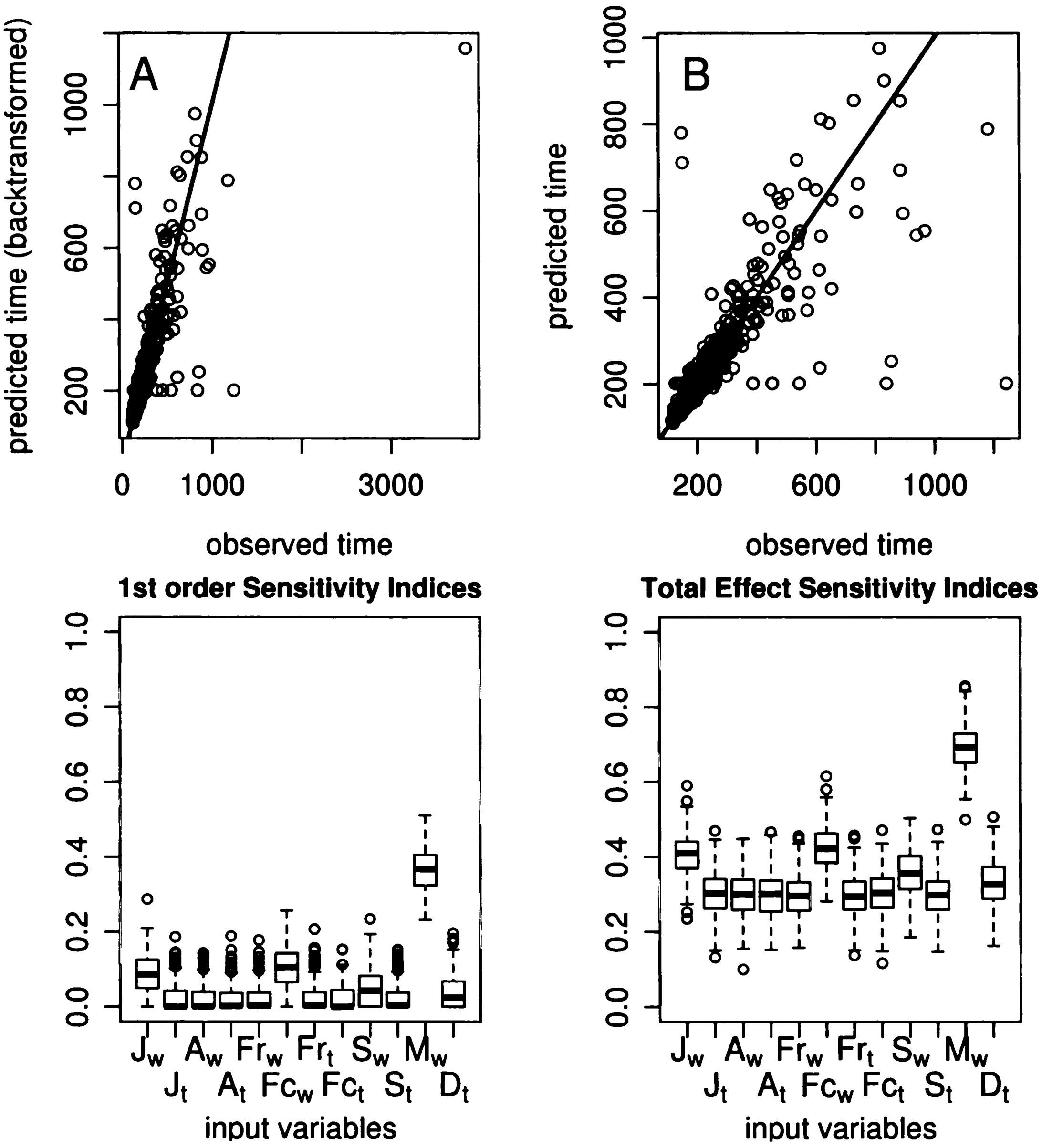

Figure 3.6: Prediction plots and Sensitivity analysis using the Bayesian gaussian process model for the FULL-data

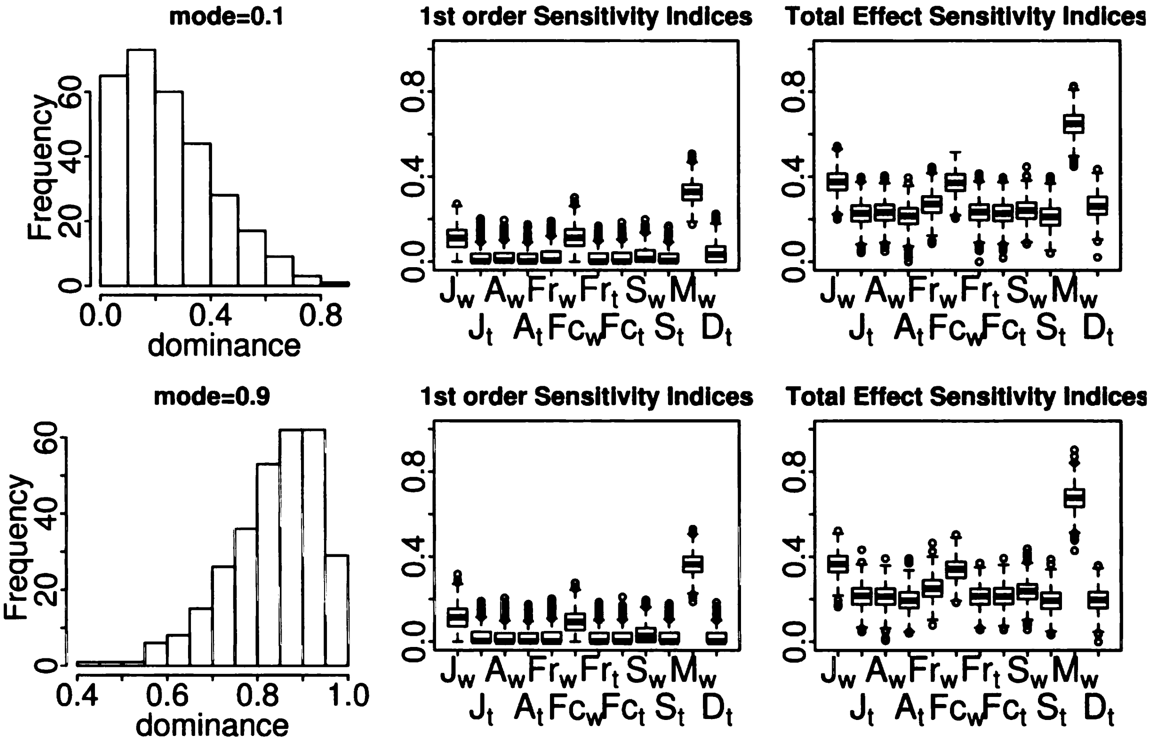

Figure 3.7: Sensitivity analysis using the Bayesian gaussian process model for the FULL-data using different uncertainty distributions for the dominance parameter

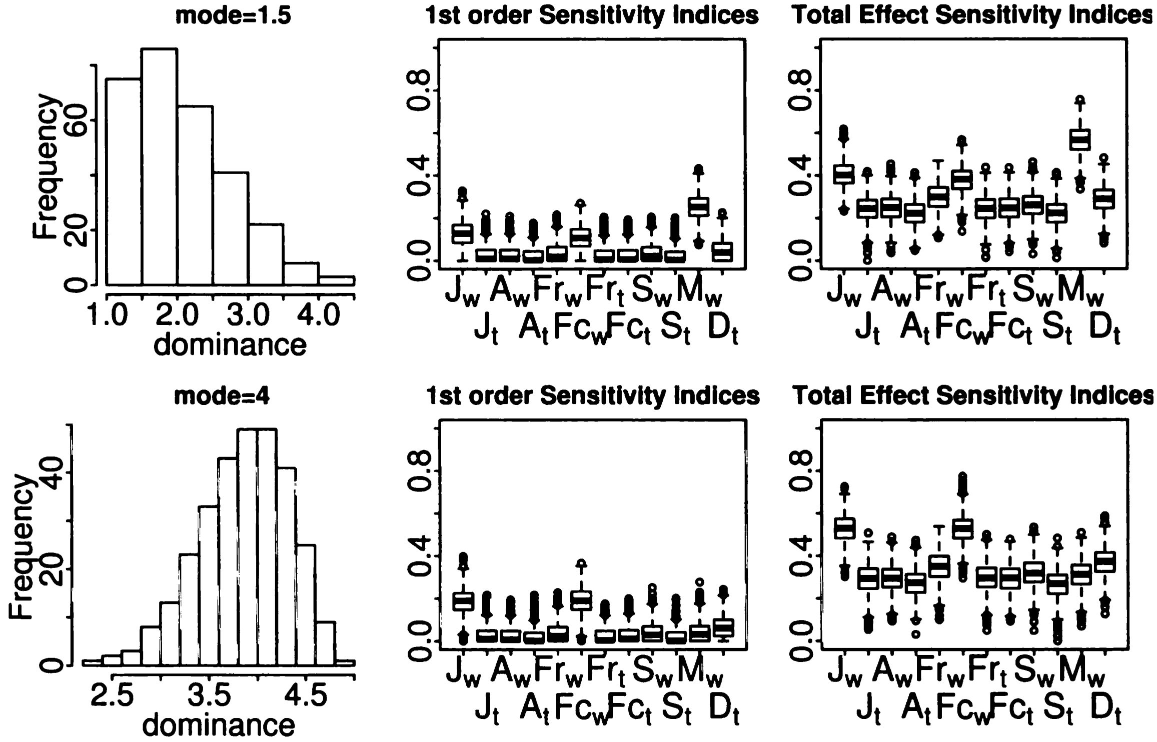

Figure 3.8: Sensitivity analysis using the Bayesian gaussian process model for the FULL-data using alternate uncertainty distributions for the WT mating success parameter

Concluding Remarks

This dissertation work focused on improvement of methods to predict transgene introgression into natural populations using modeling approaches. Models to assess transgene introgression into natural populations need to incorporate ecological principles in a biologically meaningful manner by accounting for variation that generates uncertainty in the predictions. Variation is ubiquitous in natural populations arising intrinsically due to chance events, i.e. demographic stochasticity, resulting in variability in estimates of population level parameters. The overarching objective of this research was to assess the impact of different sources of variability on model predictions, thus providing realistic bounds for predictions that can be subsequently used for policy and regulation.

Chapter 1 of this dissertation looked at the effect of demographic stochasticity on predictions, and the relative contributions of fitness components to the resultant variability. This was accomplished by conducting stochastic simulations of transgene introgression for combinations of fitness components contributing to demographic stochasticity. All simulations were conducted using a stochastic re-implementation of the net fitness component model (Muir and Howard, 2001, 2002) and results were assessed for two sets of predictions: the transgene frequency in the populations and the time to extinction of the natural population. Results from this work show that demographic stochasticity not only contributes to uncertainty in predictions, but also that it can reduce the average time to extinction. More importantly, assessment of gene frequencies revealed that demographic stochasticity can elevate persistence of the transgene far beyond predictions within a purely deterministic framework. Additionally, it was seen that stochasticity in the juvenile viability component contributed the most to variability in predictions, while stochasticity in the fecundity and mating success component did not have any substantial influence. Thus this work highlights the importance of using a stochastic framework towards making realistic predictions when using modeling approaches for risk assessment.

The results of Chapter 2 underscores the impact of uncertainties in estimated parameters on model predictions. Transgenes used for genetically engineering organisms can be derived from sources that are phylogenetically disparate. This is the case for our model organism, a genetically modified zebrafish (the Glo-Fish), where the gene producing florescence is derived from a sea-anemone. This necessitates the evaluation of fitness components using laboratory experiments, at least for the transgenic individual, since the possibility of the same genotype occurring under natural conditions is almost nil. Using the non-parametric bootstrap to assess the uncertainty in parameter estimates and incorporate it into model predictions, our results here show that we can can generate more realistic predictions of transgene fate as validated against a mesocosm experiment. Given that the experimental settings used to estimate the fitness components did not allow us to study the interdependencies among them, we also implemented a copula approach to assess the effect of such correlations. This application lent insights into the limitations of the current experimental set-up and we proposed enhancements that can facilitate better estimation. In the event that data collected is not amenable to estimate the copula parameters, this can still be a powerful tool to assess the effect of different types of correlation patterns within a sensitivity analysis framework.

The direct simulation methods applied in Chapters 1 and 2 posed significant computational burdens, which will increase with extensions to the current modeling framework. For example, linking forecasting data on experimental drivers with elements of the transition matrices (Gotelli and Ellison, 2006) for emulating effects of long-term environmental change. Therefore, to exhaustively evaluate model behavior we applied global sensitivity analysis methods, which are computationally demanding when based on Monte Carlo methods. The results using these sensitivity measures provide measures on the importance of individual fitness components on model behavior over different regions of the parameter space, which can be used to direct future research. For example, the model is not very sensitive to mating success of the wildtype genotype when the distribution of mating success is ≥ 3. Thus further model evaluations in this region can be fixed at a single value for the mating success component. To alleviate this computational burden, in chapter 3 we applied a meta-modeling approach using Bayesian Gaussian Process models, which can be used not only to derive sensitivity measures but also to explore differing forms of uncertainty distributions of the parameters. Our results show that this approach performs very satisfactorily and the choice of the Bayesian paradigm can be of immense utility for further extending the model, e.g. to incorporate demographic stochasticity.

The results from this dissertation are not specific to the transgenic system and can be broadly applied in areas of conservation and demography. The work on demographic stochasticity highlights the need to examine evolutionary trajectories combining genetics and population dynamics, for which the transgenic system was a suitable model as the exact genetic modification underlying the phenotype is known. This work used two less known methodological aspects, the copula approach and the Bayesian Gaussian Process meta-modeling, that are widely suited to modeling applications in ecology. Thus results from this research can be used to enhance not only the predictive framework for transgene introgression, but also to foster methodological developments in broader areas of ecology.

Appendix A

Description of the model

Let $n^{g}_{a,t}$ represents the number of individuals in a particular age class a (where $a = 0, 1, 2, \ldots, m$) of genotype g ($g = WW, Ww$ or ww) at time t. The composition of the total population in an age class at any given time t can be written as the sum of numbers across genotypes $N^{total}_{a,t} = \sum_g N^{g}_{a,t}$. The dynamics of the population vector of age classes $\overrightarrow{N^{total}_t}(= \sum_g \overrightarrow{N^{g}_t})$, where $\overrightarrow{N^{g}_t}$ represents the population vector of an individual genotype g at time t, can be represented using a matrix approach

$$\overrightarrow{N^{total}_{t+1}} = \begin{bmatrix} A^{WW} & A^{Ww} & A^{ww} \end{bmatrix} \times \begin{bmatrix} \overrightarrow{N^{WW}_t} \\ \overrightarrow{N^{Ww}_t} \\ \overrightarrow{N^{ww}_t} \end{bmatrix} \tag{A.1}$$

where A^{g} is the projection matrix for genotype g. In the deterministic case, the projection matrix A^{g} is composed of parameters for fecundity and survival (see SI Table S2 for details on the parameters used in the Monte Carlo runs). In the simplest case, i.e., for a single genotype with m age classes we can write,

$$\overrightarrow{N_{t+1}} = A \times \overrightarrow{N_t} = \begin{bmatrix} 0 & 0 & f_1 & \cdots & f_k \\ s_1 & 0 & 0 & \cdots & 0 \\ 0 & s_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_{m-1} \end{bmatrix} \times \begin{bmatrix} n_0 \\ \vdots \\ \vdots \\ n_m \end{bmatrix} \qquad \text{(A.2)}$$

where the f_i's denote the fecundity functions/parameters,k is the num ber of fecund age classes and the s_i's denote the survival parameters, $\overrightarrow{N_t}$ is the population vector at time t (For further details on analysis of the dynamics of the population in the deterministic case refer to Caswell (Caswell, 1989)). This can be extended as a partitioned matrix involving the fecundity functions and the survival functions for all three genotypes as in equation (2) (See Methods). As seen in equation (2), the survival functions can be partitioned to form individual submatrices for each genotype ($\mathbf{S^g}$) which are independent of each other. However, the fecundity functions are nonlinear because they depend on all three genotypes (both the male and female fractions) as underlined below for the WW genotype

$$\begin{aligned} N^{WW}_{t+1} = & \underline{\left[N^{WW}_{f,t} f^{WW}_t sr^{WW}_{WW \times WW} p^{WW}_t +\right]} \\ & \underline{\left[N^{WW}_{f,t} f^{WW} sr^{WW}_{WW \times Ww} p^{Ww} + \dots\right]} + \\ & s^{WW}_0 N^{WW}_{0,t} + \dots + s^{WW}_{a,t} N^{WW}_{a,t} \end{aligned} \qquad \text{(A.3)}$$

where N^{WW}_{t+1} is the population size of the wildtype at time $t+1$,f^g is the fecundity of females of genotype g, s^g_a represents the survival transitions from age classes, $sr^g_{g_1 \times g_2}$ represents the proportion of offspring of genotype g produced by mating of females of genotype g_1 with males of genotype g_2, and $p^g_t = \dfrac{m^g N^g_{m,t}}{\sum_g m^g N^g_{m,t}}$ represents the relative mating success of males of genotype g with mating advantage m (this assumes that mating advantage can be predicted from marginal frequencies of males, i.e. does not interact with genotype of female). Therefore from equations (2) and (3) it is seen that the individual projections matrices for each genotype is not a linear function of the population for that genotype alone, and the overall projection matrix A is a nonlinear function of population size across all genotypes. When the transgenic genotype has a higher relative mating advantage, the fecundity functions (i.e. the first two lines of the R.H.S in equation (3)) are a decreasing function of time whereby the number of WW offspring produced from the $Ww_f \times WW_m$ or $Ww_f \times Ww_m$ matings are lower than the number of Ww and ww offspring produced by the $WW_f \times Ww_m$ or $WW_f \times ww_m$ matings.

Appendix B

Stochastic implementation for the fitness components

B.1 Fecundity of females

The number of eggs laid by each female is assumed to follow a Poisson distribution with parameter Λ, the expected number of eggs laid by an individual female. Assuming that the eggs laid by each female in a time step are independent and identically distributed (*iid*) random variables, for each time step the number of females that reproduce per genotype over all ages ($N^g_{f,rep} = \sum_a n^g_{f,a}$) were calculated and the total number of eggs produced by the available females in a genotype g was simulated as a Poisson variable with parameter $N^g_{f,\mathrm{rep}}\Lambda$. It is easily verified that the sum of n *iid* Poisson variables with parameter Λ is also Poisson, with parameter $n\Lambda$ (Casella and Berger, 1990)

B.2 Mating success of males

Relative mating success is defined as the relative proportion of individuals that mate weighted by the mating advantage that an individual of a particular genotype enjoys.

Let

$$f_{sex}^{g} = \frac{N_{sex}^{g}}{\sum_g N_{sex}^{g}}$$

denote the relative frequency of individuals of a particular genotype g in a specific sex. Then the relative mating success M_{sex}^{g} is a function of the mating advantage m_{sex}^{g} and the relative frequency f_{sex}^{g} where

$$M_{sex}^{g} = \frac{m_{sex}^{g} \times f_{sex}^{g}}{\sum_g m_{sex}^{g} \times f_{sex}^{g}}$$

Assuming that the number of individuals of a particular sex that are available for reproduction at any time is $N_{rep,\,sex}^{g}$, the realized number of individuals is the re-weighted number with the mating successes as weights, that is $N_{rep,\,sex}^{real,\,g} \sim M_{sex}^{g} \times N_{rep,\,sex}^{g}$, in the deterministic context. In the stochastic model, after we calculate the relative mating success M^{g} as given above, we assume the number of individuals of a particular sex that actually mate as a random draw from the available N_{sex}^{total}, where $N_{sex}^{total} = \sum_g N_{sex}^{g}$, from a multinomial distribution with parameters given by the M_{sex}^{g}'s, i.e,

$$N_{rep,\,sex}^{g,\,real} \sim Mult(N_{sex}^{total}; M_{sex}^{WW}, M_{sex}^{Ww}, M_{sex}^{ww})$$

and the probability of mating is calculated as

$$p_{sex}^{g} = \frac{N_{rep,\,sex}^{real,\,g}}{\sum_g N_{rep,\,sex}^{real,\,g}}$$

All females are assumed to reproduce and mating occurs at random, i.e., the probability of a female pairing with individual male from any genotype depends only on the relative proportion of each genotype and their mating advantage.

B.3 Survival from age class a to $a + 1$

The event of an individual surviving from age a to $a + 1$ can be considered a Bernoulli random variable with probability of survival s_a. Assuming independence across individuals and a constant survival parameter, the number of individuals surviving from age a to $a + 1$ in one time step is a binomial random variable, i.e. $N_{a+1,t+1} \sim Bin(N_{a,t}; s_a)$.

Appendix C

Analytical results on effect of stochasticity in each fitness component on variation

C.1 Fecundity

For stochasticity in fecundity of feeadtotoc males, assuming that all eggs are fertile the number of offspring entering the population at time t ($N^g_{0,t}$) is a Poisson variable with

$$E\left(N^g_{0,t}\right) = Var\left(N^g_{0,t}\right) = N^g_{f,rep,t}\Lambda.$$

If $N^g_{f,rep,t}\Lambda \gg 0$, the Poisson distribution tends to a Gaussian process as $N^g_{0,t} \sim N(\mu,\sigma)$, where $\mu = N^g_{f,rep,t}\Lambda$ and $\sigma = \sqrt{N^g_{f,rep,t}\Lambda}$, and it is again simple to see that the %CV declines exponentially as the product $N^g_{f,rep,t}\Lambda$ increases. In our simulations $\Lambda = 8.8$ and for $N^g_{f,rep,t} = 50$ for an arbitrary time t we can calculate the 99% intervals as (387,496), which implies that the probability the number of offspring produced is smaller than 387 or greater than 496 is 0.01. As $N^g_{f,rep,t}$ increases this bound decreases and so we can expect stochasticity in the fecundity component should make significant contributions only when the population size is very small

C.2 Mating Success

Given the way we calculate relative mating success (see Appendix B) for a genotype g, the realized number of males is a random variable from a multinomial distribution with parameters M_m^g, the relative mating success, and, $\sum_g N_{rep,m}^g$, the total number of males present. Conditional on the parameters for the other genotypes, we know that the realizations of males for genotype g_1 , is a Binomial random variable i.e

$$N_{rep,m}^{real,g_1} \sim Bin\left(\sum_g N_{rep,m}^g ; M_m^{g_1} \mid M_m^{g_2}, M_m^{g_3}\right)$$

with expectation and variance given by:

$$E\left[N_{rep,m}^{real,g_1}\right] = M_m^{g_1} \sum_g N_{rep,m}^g$$

and

$$Var\left[N_{rep,m}^{real,g_1}\right] = M_m^{g_1}\left(1 - M_m^{g_1}\right)\sum_g N_{rep,m}^g$$

However, in the model the final probability of mating with a male of genotype g by any female is given by the $p_m^g = \dfrac{N^{real,g_1}}{\sum_g N_{rep,m}^g}$, whereby taking expectations and variances we have

$$E\left[p_m^g\right] = E\left[\frac{N_m^{real, g_1}}{\sum_g N_{rep, m}^g}\right] = \left(\sum_g N_{rep, m}^g\right)^{-1} E\left[N_m^{real, g_1}\right] \qquad \text{(C.1)}$$

and

$$Var\left[p_m^g\right] = Var\left[\frac{N_m^{real, g_1}}{\sum_g N_{rep, m}^g}\right] = \left(\sum_g N_{rep, m}^g\right)^{-2} Var\left[N_m^{real, g_1}\right] \qquad \text{(C.2)}$$

From equations (4) and (5) we can see that the contribution of the mating success component to the variance will be really low and therefore we can expect stochasticity in this component to only influence the mean of the process unless the number of reproducing males are very low (Fig. 1, Fig. 3, scenario M).

C.3 Viability

In the case of adult and juvenile viability, since we assume that $N_{a+1, t+1} \sim Bin(N_{a, t}; s_a)$, where $N_{a, t}$ is also $\sim Bin(N_{a-1, t-1}; s_{a-1})$ and so on, the unconditional distribution of $N_{a+1, t+1}$ is $\sim Bin(N_{a-1, t-1}; s_a s_{a-1})$ with expectation

$$E\left[N_{a+1, t+1}\right] = s_a s_{a-1} N_{a-1, t-1}$$

and variance

$$Var\left[N_{a+1, t+1}\right] = s_a s_{a+1}\left(1 - s_a s_{a+1}\right) N_{a-1, t-1}.$$

Since $N_{a-1, t-1}$ itself is a radom variable, we can extend this all the way to the the first age class at time $t - a$ and therefore write the expectation and variance at a particular time

t for a particular age class a as:

$$E\left[N_{a+1,t+1}\right] = \left(\left(\prod_{k=1}^{a} s_k\right) N_{0,t-a}\right) \tag{C.3}$$

and

$$Var\left[N_{a+1,t+1}\right] = \left(\left(\prod_{k=1}^{a} s_k\right)\left(1 - \prod_{k=1}^{a} s_k\right) N_{0,t-a}\right) \tag{C.4}$$

where $N_{0,t-a}$ represents the number of offspring born at an earlier time $t - a$. From equations (3),(6) and (7) we can see that the overall variance of the population at time $t + 1$ for a genotype g is given by $Var\left[N^g_{t+1}\right] = \sum_{a=1}^{m} Var\left[N^g_{a,t+1}\right]$, assuming that the covariances among age-classes are independent. if the $N_{0,t-a}$ are not very different across age classes, then contributions to variance from each age classes is based on the product

$$\left(\prod_{k=1}^{a} s_k\right)\left(1 - \prod_{k=1}^{a} s_k\right) \tag{C.5}$$

That is the contribution of each age class to variance in population size depends on the survivorship upto that particular age class. The variance is therefore a weighted sum of the offspring produced from time $t-a$ to time $t-1$. Hence, the highest contributions to the overall variance are from those ages where the expression in equation [8] is maximized, which usually occurs in age classes in the middle of the age distribution. For example, in the daily model

$$\left(\prod_{k=1}^{a} s_k\right)\left(1 - \prod_{k=1}^{a} s_k\right)$$

is maximized when $a = 11$ and is > 0.05 in the interval $a : (2, 43)$. Therefore, we expect the most contribution to variance to occur from these earlier age classes.

Appendix D

Additional plots for Chapter 1

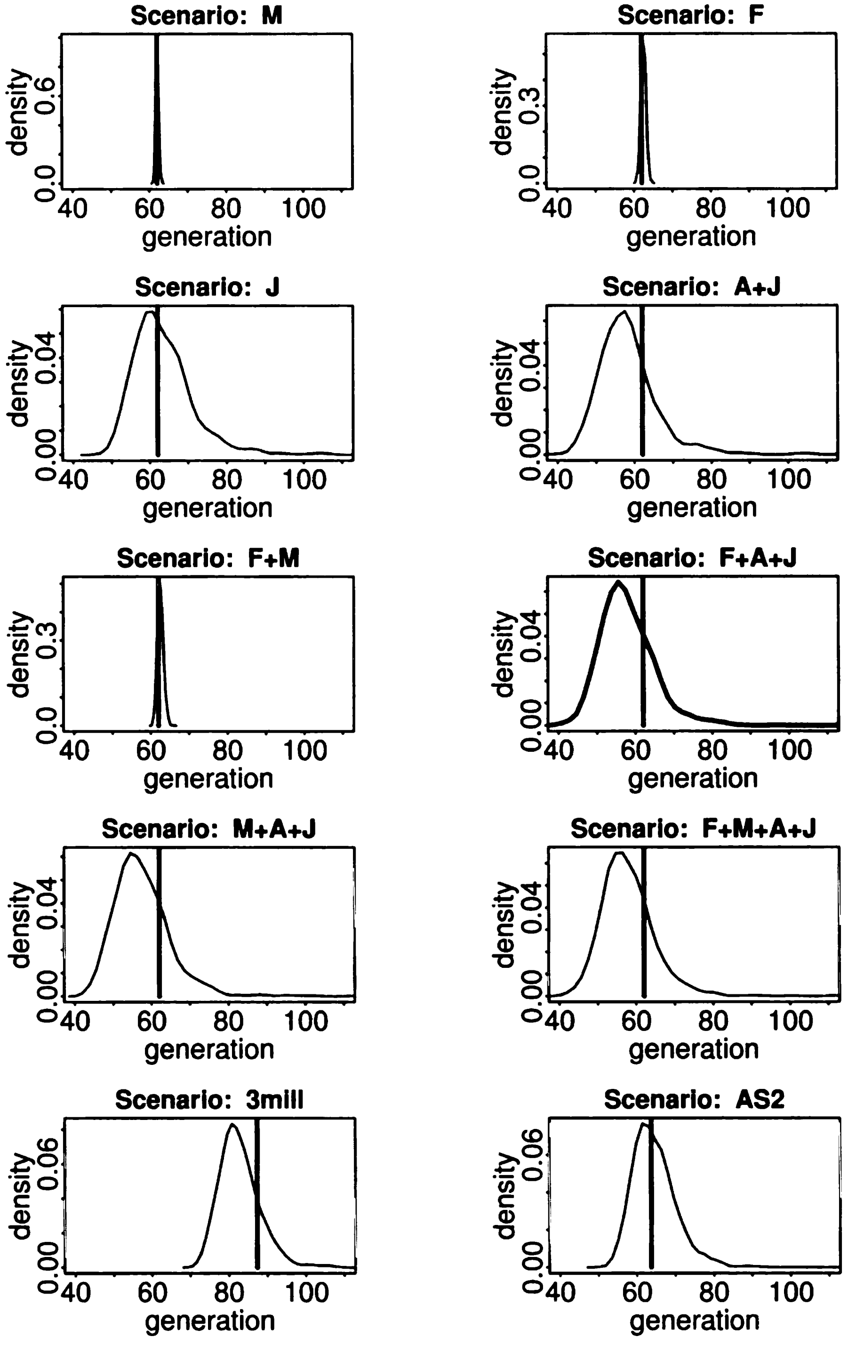

Figure A.1.1: Distribution of extinction times for all scenarios at a quasi-extinction threshold of 100 individuals. A kernel density estimate using a Gaussian kernel is used for smoothing. The vertical bars represent the time to extinction for the deterministic model.

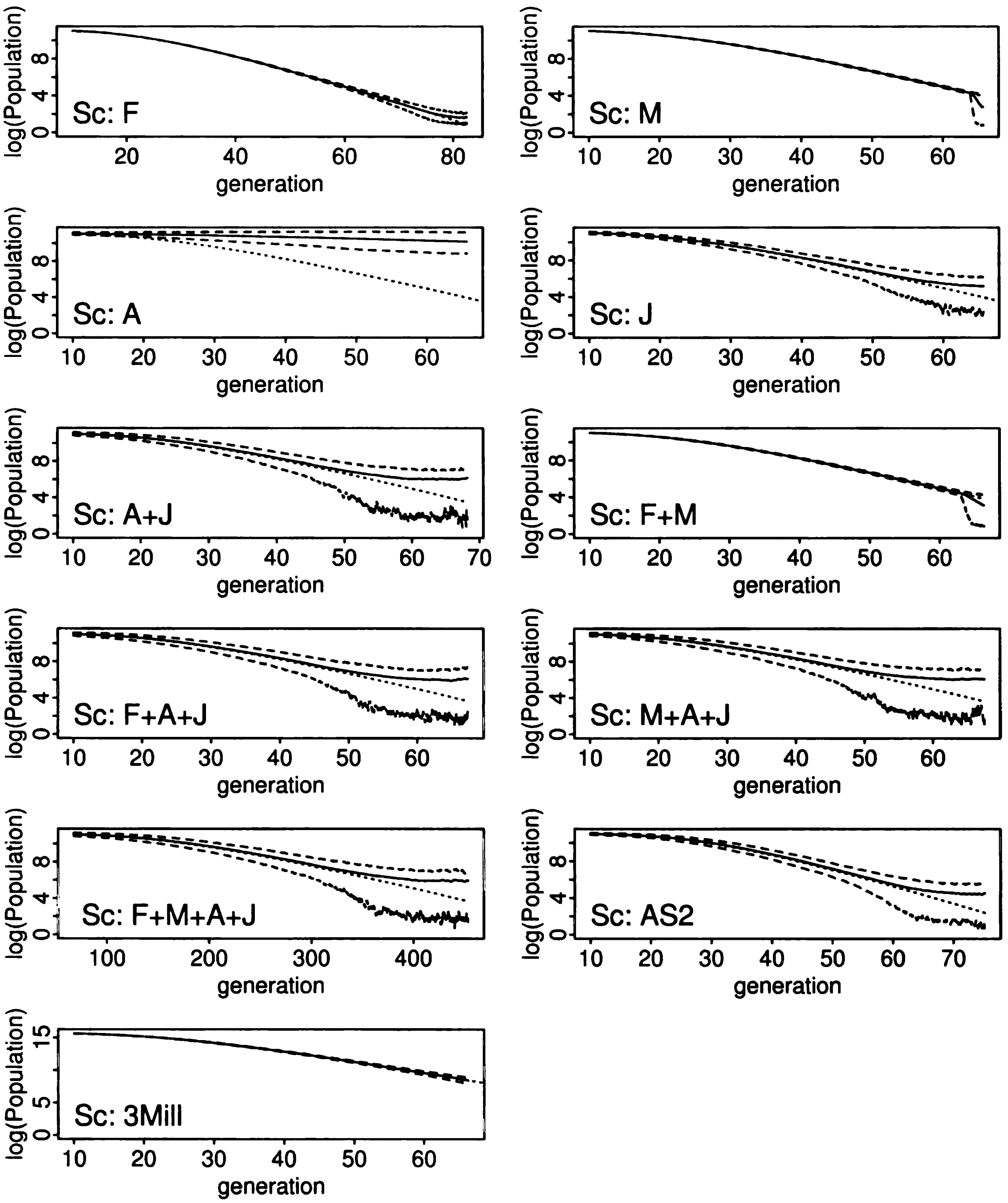

Figure A.1.2: The evolution of average total population size over time for all scenarios and the quantiles of population size over time. The dashed blue line represents the evolution from the deterministic model, while the green lines represent the 5% and 95% quantiles respectively. The red line represents the average population size over 1000 runs at each time step. Population numbers have been log transformed for better visualisation.

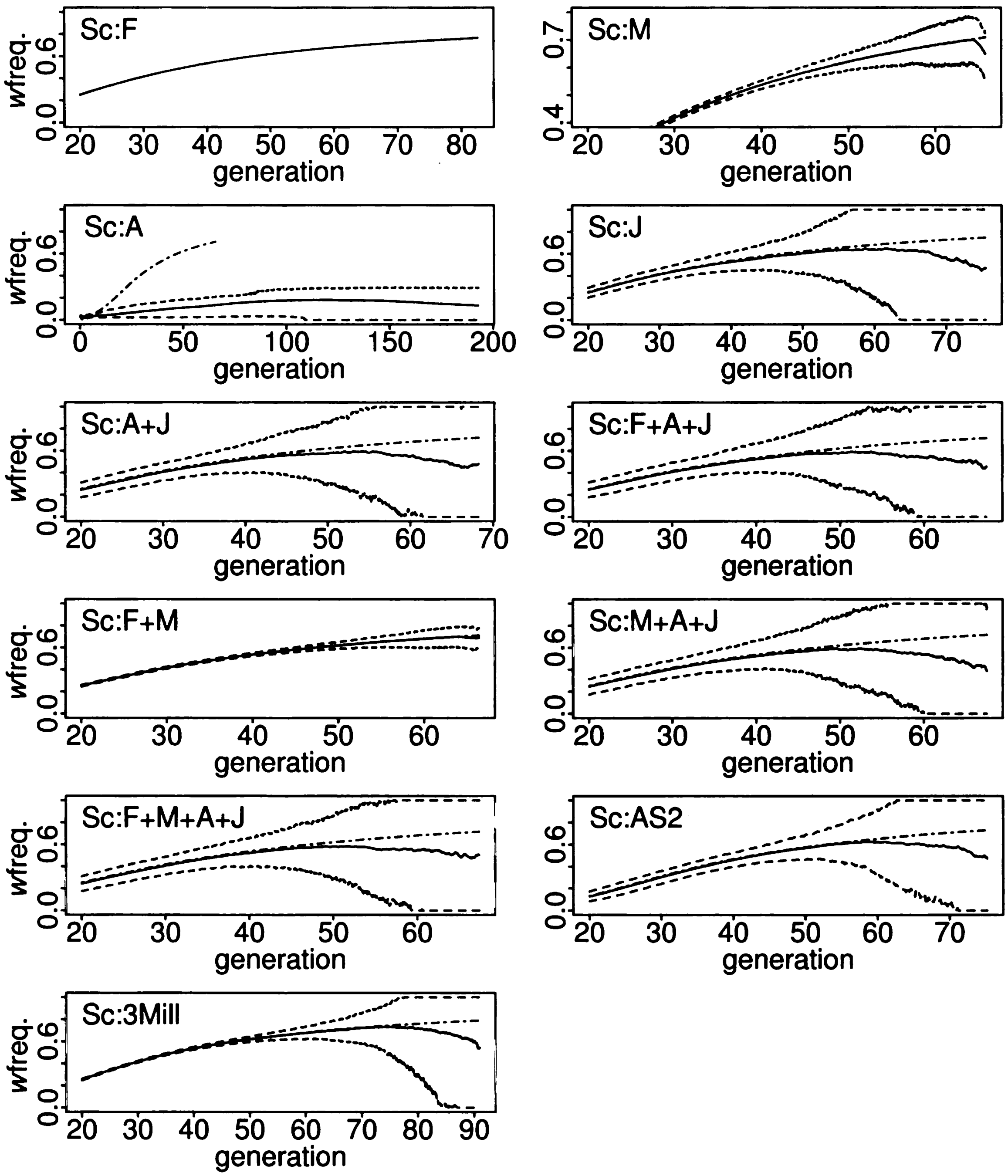

Figure A.1.3: The evolution of average transgene frequency over time for all scenarios and the quantiles of the transgene frequency at time step. The dashed blue line represents the evolution from the deterministic model, while the green lines represent the 5% and 95% quantiles respectively. The red line represents the average population size over 1000 runs at each time step.

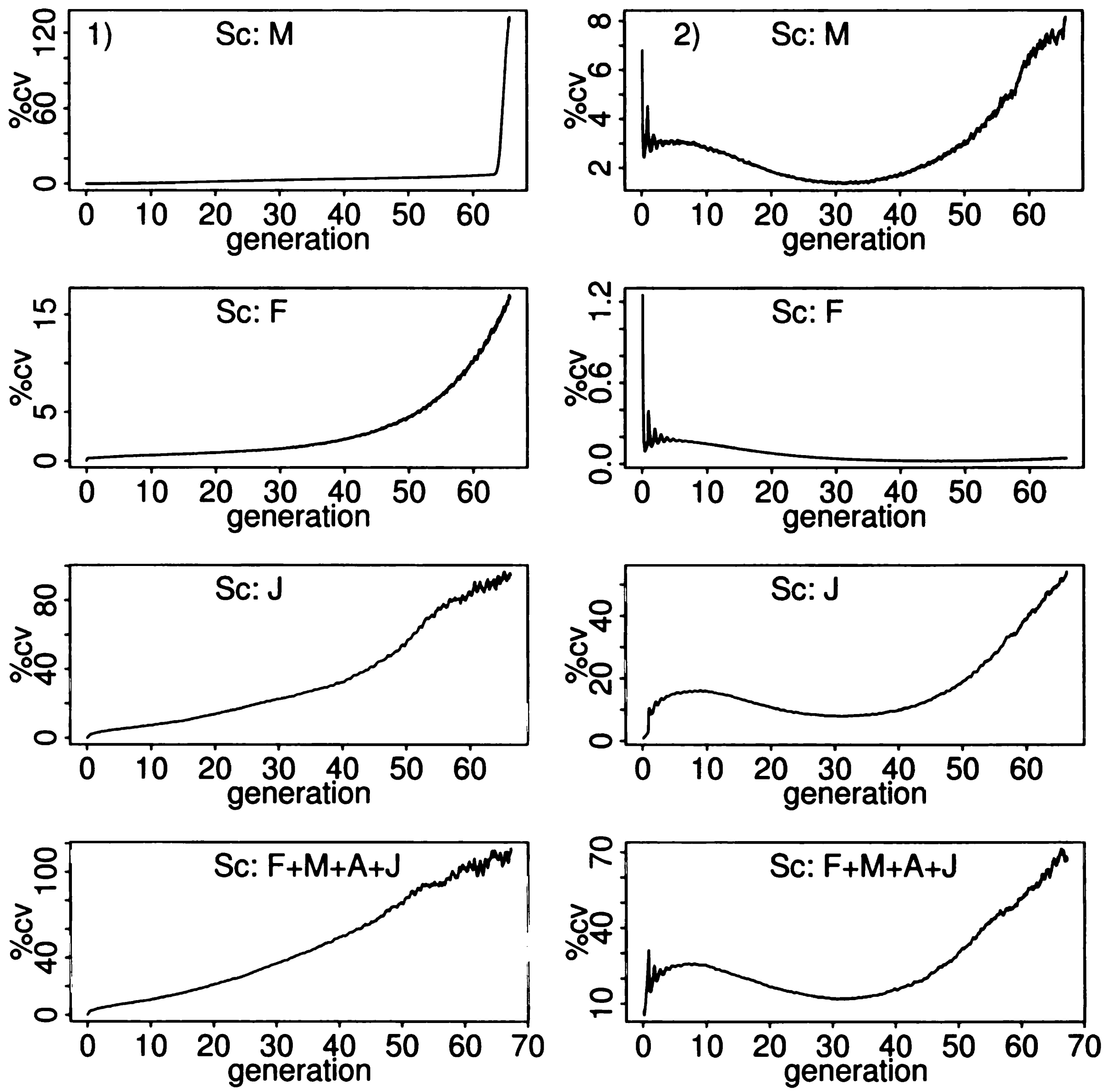

Figure A.1.4: The evolution of the coeffecient of variation(%CV) for the runs over time for scenarios F, M, J and F+M+A+J. Limit for the x-axis is chosen as the generation where at least 100 populations were not extinct. 1) The y-axis represents the trajectory of the %CV for the total population size. 2) The y-axis for the right column represents the trajectory of the %CV of the transgene gene frequency.

Appendix E

Additional plots for Chapter 2

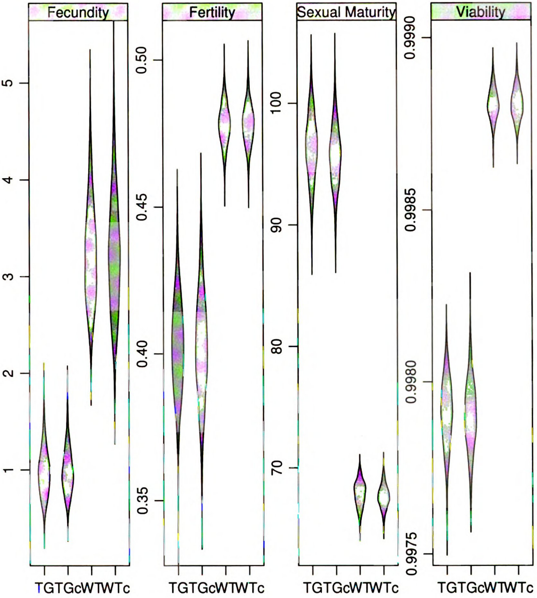

Figure A.2.1: Marginal distribution of the bootstrap parameter estimates for the Wildtype (WT) genotype and the transgenic type (TG). Marginal distributions for the method using Copulas simulating dependence are labeled with a "c" subscript.The The same set of parameters were used for three different values of realtive mating success of the WT males (1.16.1.93 and 5 respectively).

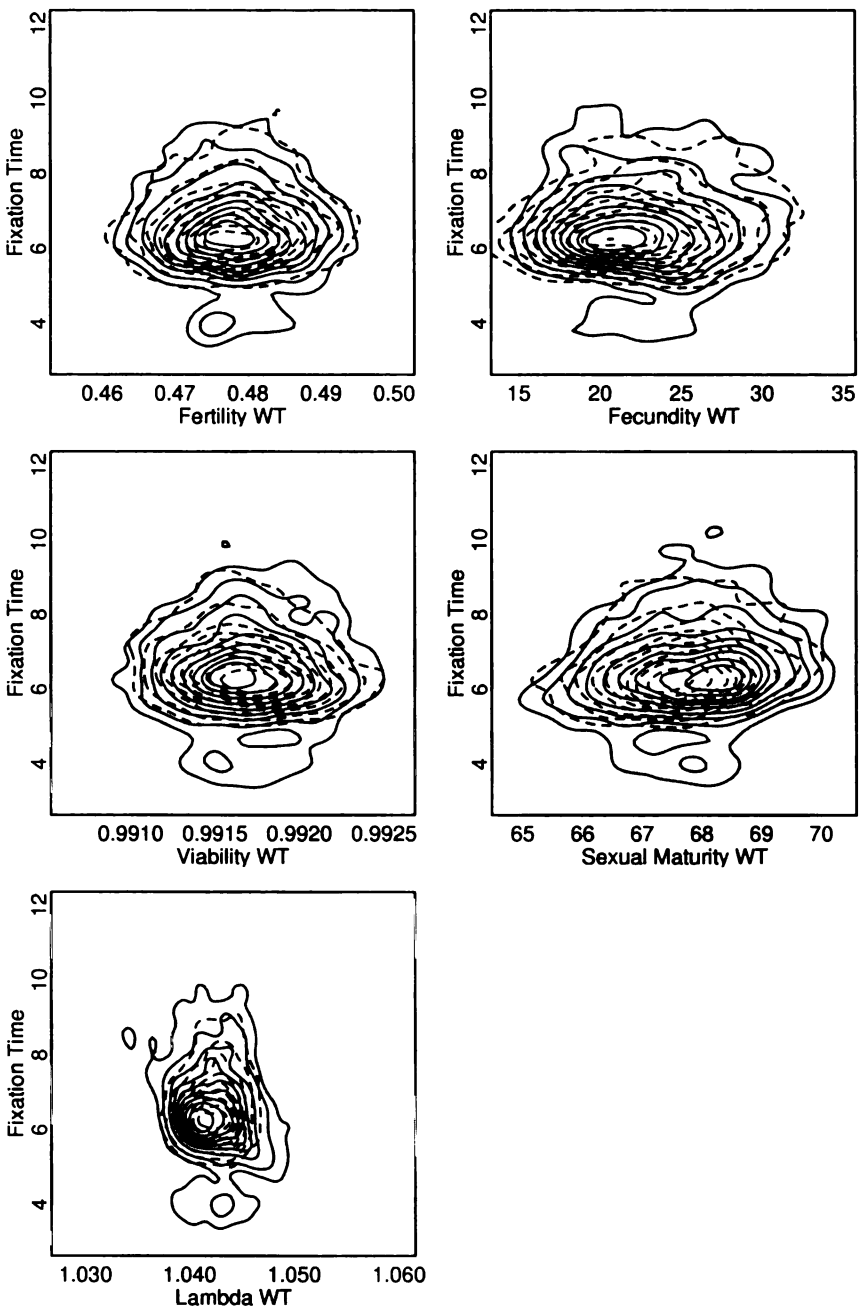

Figure A.2.2: Comparison of the fixation time as a function of fitness components for the Wildtype(WT) genotype. The dashed contour lines represent the scenario incorporating dependence using the copula approach

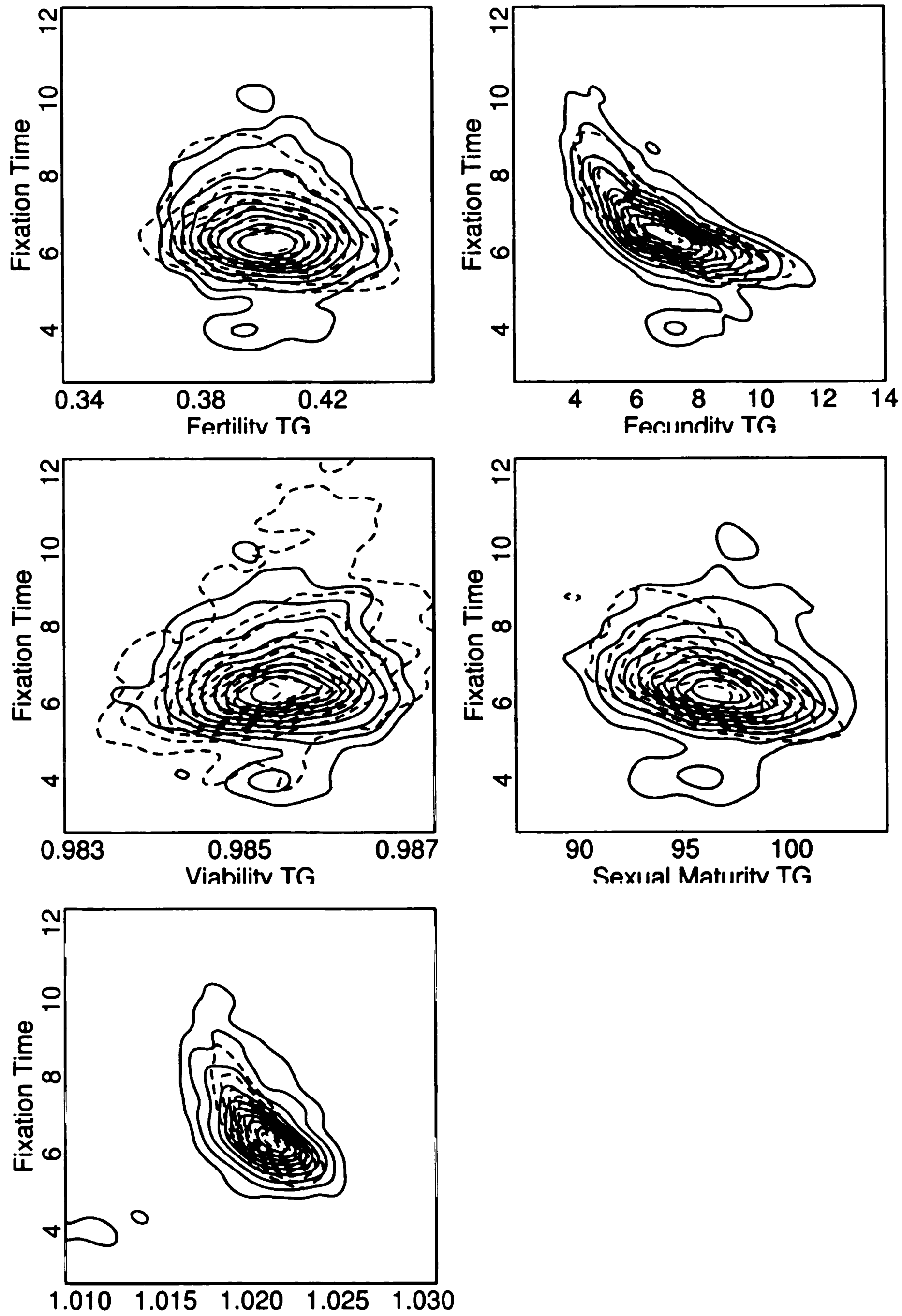

Figure A.2.3: Comparison of the fixation time as a function of fitness components for the transgenic (TG) genotype. The dashed contour lines represent the scenario incorporating dependence using the copula approach

BIBLIOGRAPHY

Aerni, P., 2004. Risk, regulation and innovation: The case of aquaculture and transgenic fish. *Aquatic Sciences* **66**:327–341.

Andow, D. and A. Hilbeck, 2004. Science-based risk assessment for nontarget effects of transgenic crops. *BioScience* **54**:637–649.

Andow, D. and C. Zwahlen, 2006. Assessing environmental risks of transgenic plants. *Ecology Letters* **9**:196 – 214.

Barnthouse, L. and M. Sorenson, editors, 2007. *Population-Level Ecological Risk Assessment*. CRC, Boca Raton, FL.

Beauchamp, J. J. and J. S. Olson, 1973. Corrections for bias in regression estimates after logarithmic transformation. *Ecology* **54**:1403–1407.

Beaumont, M. and B. Rannala, 2004. The Bayesian revolution in genetics. *Nature Reviews Genetics* **5**:251–261.

Benton, T. G. and A. Grant, 1996. How to keep fit in the real world: Elasticity analyses and selection pressures on life histories in a variable environment. *The American Naturalist* **147**:115–139.

Bodmer, W. F., 1965. Differential fertility in population genetics models. *Genetics* **51**:411–424.

Boyce, M. S., C. V. Haridas, C. T. Lee, and T. N. S. D. W. Group, 2006. Demography in an increasingly variable world. *Trends Ecol Evol* **21**:141–148.

Cariboni, J., D. Gatelli, R. Liska, and A. Saltelli, 2007. The role of sensitivity analysis in ecological modelling,. *Ecological Modelling* **203**:167–182. to read.

Casella, G. and R. L. Berger, 1990. *Statistical Inference*. Wadsworth & Brooks/Cole Pub. Co., Pacific Grove, Calif.

Caswell, H., 1989. *Matrix Population Models: Construction, Analysis, and Interpretation*. Sinauer Associates, inc., sunderland Mass.

Caswell, H., 2008. Perturbation analysis of nonlinear matrix population models. *Demographic Research* **18**:59–115.

Charlesworth, B., 1994. *Evolution in Age-Structured Populations.* Cambridge studies in mathematical biology. Cambridge University Press, Cambridge.

Claessen, D., C. Gilligan, P. Lutman, and F. van den Bosch, 2005. Which traits promote persistence of feral GM crops? Part 1: implications of environmental stochasticity. *Oikos* **110**:20 – 29.

Clark, J. and B. Whitelaw, 2003. A future for transgenic livestock. *Nature Reviews Genetics* **4**:825 – 833.

Clark, J. S., 2003. Uncertainty and variability in demography and population growth: A hierarchical approach. *Ecology* **84**:1370–1381.

Coulson, T., J.-M. Gaillard, and M. Festa-Bianchet, 2005. Decomposing the variation in population growth into contributions from multiple demographic rates. *Journal of Animal Ecology* **74**:789–801.

Devlin, R., M. D'Andrade, M. Uh, and C. Biagi, 2004. Population effects of growth hormone transgenic Coho Salmon depend on food availability and genotype by environment interactions. *PNAS* **101**:9303 – 9308.

Devlin, R., L. Sundstrom, and W. Muir, 2006. Interface of biotechnology and ecology for environmental risk assessments of transgenic fish. *Trends in biotechnology* **24**:89 – 97.

Efron, B. and R. Tibshirani, 1998. *An Introduction to the Bootstrap.* Chapman & Hall/CRC, New York.

Embrechts, P., F. Lindskog, and A. J. McNeil, 2003. Modelling Dependence with Copulas and Applications to Risk Management, chapter 8. Elsevier.

Engen, S., R. Lande, and B. Saether, 2005a. Effective size of a fluctuating age-structured population. *Genetics* **170**:941 – 954.

Engen, S., R. Lande, B. Saether, and H. Weimerskirch, 2005b. Extinction in relation to demographic and environmental stochasticity in age-structured models. *Mathematical Biosciences* **195**:210 – 227.

Ewens, W., 2004. *Mathematical Population Genetics.* 2nd ed., Springer-Verlag, New York.

Fang, K. and A. Sudjianto, 2006. *Design and Modeling for Computer Experiments.* Chapman & Hall/CRC.

Ferriere, R., U. Dieckmann, and D. Couvet, 2004. *Evolutionary Conservation Biology.* Cambridge studies in adaptive dynamics. Cambridge University Press, Cambridge.

Fieberg, J. and S. P. Ellner, 2003. Using PVA for management despite uncertainty: Effects of habitat, hatcheries, and harvest on Salmon. *Ecology* **84**:1359–1369.

Fieberg, J. and K. J. Jenkins, 2005. Assessing uncertainty in ecological systems using global sensitivity analyses: A case example of simulated wolf reintroduction effects on elk. *Ecological Modelling* **187**:259–280.

Fox, G. and B. Kendall, 2002. Demographic stochasticity and the variance reduction effect. *Ecology* **83**:1928–1934.

Gaggiotti, O., 2003. Genetic threats to population persistence. *Annales Zoologici Fennici* **40**:155–168.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi, 2003. Gnu Scientific Library: Reference Manual. Network Theory Ltd.

Garnier, A., A. Deville, and J. Lecomte, 2006. Stochastic modelling of feral plant populations with seed immigration and road verge management. *Ecological modelling* **197**:373–382.

Garnier, A. and J. Lecomte, 2006. Using a spatial and stage-structured invasion model to assess the spread of feral populations of transgenic oilseed rape. *Ecological modelling* **194**:141 – 149.

Gelman, A., 2004. *Bayesian Data Analysis.* 2nd ed., Chapman & Hall/CRC, Boca Raton, FL

Genest, C. and A. Favre, 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal Of Hydrologic Engineering* **12**:347–368.

Genest, C. and J. MacKay, 1986. The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician* **40**:280–283.

Goldberg, P., C. Williams, and C. Bishop, 1998. Regression with input-dependent noise: A gaussian process treatment. pages 493–499 *in* M. Kearns, M. Jordan, &

T. Solla, editors. *Advances in Neural Information Processing Systems* MIT press, Cambridge, MA.

Gotelli, N. J. and A. M. Ellison, 2006. Forecasting extinction risk with nonstationary matrix models. *Ecol Appl* **16**:51–61.

Gramacy, R., 2007. tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software* **19**:6.

Gramacy, R. and H. Lee, 2008. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**:1119–1130.

Grant, A. and T. G. Benton, 2000. Elasticity analysis for density-dependent populations in stochastic environments. *Ecology* **81**:680–693.

Hails, R. and K. Morley, 2005. Genes invading new populations: A risk assessment perspective. *Trends in Ecology & Evolution* **20**:245 - 252.

Hedrick, P., 2001. Invasion of transgenes from salmon or other genetically modified organisms into natural populations. *Canadian Journal Of Fisheries And Aquatic Sciences* **58**:841–844.

Inchausti, P. and J. Halley, 2003. On the relation between temporal variability and persistence time in animal populations. *Journal of Animal Ecology* **72**:899–908.

Jones, A., S. Arnold, and R. Burger, 2003. Stability of the g-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. *Evolution* **57**:1747–1760.

Kareiva, P., I. Parker, and M. Pascual, 1996. Can we use experiments and models in predicting the invasiveness of genetically engineered organisms? *Ecology* **77**:1670–1675.

Kareiva, P. M., J. G. Kingsolver, and R. B. Huey, 1993. *Biotic Interactions and Global Change*. Sinauer Associates, Sunderland, Mass.

Knibb, W., 1997. Risk from genetically engineered and modified marine fish. *Transgenic Research* **6**:59–67.

Koons, D. N., C. J. E. Metcalf, and S. Tuljapurkar, 2008. Evolution of delayed reproduction in uncertain environments: A life-history perspective. *Am Nat* **172**:797–805.

Kuparinen, A. and F. M. Schurr, 2007. A flexible modelling framework linking the spatio-temporal dynamics of plant genotypes and populations: Application to gene flow from transgenic forests. *Ecological Modelling* **202**:476 – 486.

Lande, R., 1993. Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *The American Naturalist* **142**:911–927.

Legendre, S., J. Clobert, A. Moeller, and G. Sorci, 1999. Demographic stochasticity and social mating system in the process of extinction of small populations: The case of passerines introduced to New Zealand. *American Naturalist* **153**:449–463.

Lenski, R. E., C. Ofria, R. T. Pennock, and C. Adami, 2003. The evolutionary origin of complex features. *Nature* **423**:139–144.

Meagher, T., 2003. Using empirical data to model transgene dispersal. *Philosophical Transactions: Biological Sciences* **358**:1157–1162.

Millar, R. B. and R. Meyer, 2000. Bayesian state-space modeling of age-structured data: Fitting a model is just the beginning. *Canadian Journal Of Fisheries And Aquatic Sciences* **57**:43–50.

Morris, W. and D. Doak, 2002. *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis*. Sinauer Associates.

Muir, W., 2004. The threats and benefits of GM fish. *Embo Reports* **5**:654–659.

Muir, W. and R. Howard, 1999. Possible ecological risks of transgenic organism release when transgenes affect mating success: Sexual selection and the trojan gene hypothesis. *PNAS* **96**:13853–13856.

Muir, W. and R. Howard, 2001. Fitness components and ecological risk of transgenic release: A model using Japanese Medaka (*Oryzias latipes*). *American Naturalist* **158**:1–16.

Muir, W. and R. Howard, 2002. Assessment of possible ecological risks and hazards of transgenic fish with implications for other sexually reproducing organisms. *Transgenic Research* **11**:101–114.

Muir, W. and R. Howard, 2004. Characterization of environmental risk of genetically engineered (ge) organisms and their potential to control exotic invasive species. *Aquatic Sciences* **66**:414–420.

Nacci, D. and A. Hoffman, 2008. Genetic variation in population-level ecological risk assessment, pages 93–113. *in* L.W. Barnthouse, W.R. Munns Jr & M.T. Sorensen, editors. *Population-Level Ecological Risk Assessment*, CRC, Boca Raton, FL, USA.

National Research Council, 2004. Biological confinement of genetically engineered organisms. National Academies Press.

Naujokaitis-Lewis, I. R., J. M. R. Curtis, P. Arcese, and J. Rosenfeld, 2009. Sensitivity analyses of spatial population viability analysis models for species at risk and habitat conservation planning. *Conserv Biol* **23**:225–229.

Naylor, R., K. Hindar, I. A. Fleming, R. Goldburg, S. Williams, J. Volpe, F. Whoriskey, J. Eagle, D. Kelso, and M. Mangel, 2005. Fugitive salmon: Assessing the risks of escaped fish from net-pen aquaculture. *BioScience* **55**:427–437.

Nelsen, R., 2006. *An Introduction to Copulas.* Springer Verlag.

Oakley, J. and A. O'Hagan, 2002. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**:769–784.

Oakley, J. E. and A. O'Hagan, 2004. Probabilistic sensitivity analysis of complex models: A bayesian approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **66**:751–769.

OHara, R. B., J. M. Cano, O. Ovaskainen, and C. A. J. S. Teplitsky, 2008. Bayesian approaches in evolutionary quantitative genetics. *Journal of Evolutionary Biology* **21**:949–957.

Peck, S., 2004. Simulation as experiment: A philosophical reassessment for biological modeling. *Trends in Ecology & Evolution* **19**:530 – 534.

Pilson, D. and H. Prendeville, 2004. Ecological effects of transgenic crops and the escape of transgenes into wild populations. *Annual Review Of Ecology Evolution And Systematics* **35**:149 – 174.

Plate, T., 1999. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models. *Behaviormetrika* **26**:29–50.

Pollard, J., 1966. On the use of the direct matrix product in analysing certain stochastic population models. *Biometrika* **53**:397–415.

Ramula, S. and K. Lehtila, 2005. Importance of correlations among matrix entries in stochastic models in relation to number of transition matrices. *Oikos* **111**:9–18.

Rasmussen, C. E. and C. K. I. Williams, 2005. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA.

Richter, O. and R. Seppelt, 2004. Flow of genetic information through agricultural ecosystems: A generic modelling framework with application to pesticide-resistance weeds and genetically modified crops. *Ecological Modelling* **174**:55 – 66.

Roff, D. A., 1992. *The Evolution of Life Histories. Theory and Analysis.* Chapman & Hall, New York.

Saether, B., S. Engen, A. Moeller, H. Weimerskirch, M. Visser, W. Fiedler, E. Matthysen, M. Lambrechts, A. Badyaev, P. Becker, et al., 2004. Life-history variation predicts the effects of demographic stochasticity on avian population dynamics. *American Naturalist* **164**:793–802.

Saether, B.-E. and S. Engen, 2002. Pattern of variation in avian population growth rates. *Philosophical Transactions of the Royal Society B: Biological Sciences* **357**:1185–1195.

Saether, B.-E., S. Engen, J. E. Swenson, O. Bakke, and F. Sandegre, 1998. Assessing the viability of scandinavian brown bear, *Ursus arctos*, populations: The effects of uncertain parameter estimates. *Oikos* **83**:403–416.

Saltelli, A., 2000. *Sensitivity Analysis.* Wiley Series in Probability and Statistics. Wiley.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, 2008. *Global Sensitivity Analysis: The Primer.* Wiley-Interscience.

Santner, T., B. Williams, and W. Notz, 2003. *The Design and Analysis of Computer Experiments.* Springer.

Shpak, M., 2007. Selection against demographic variance in age structured populations. *Genetics* **107**:.080747.

Snyder, R. E., 2003. How demographic stochasticity can slow biological invasions. *Ecology* **84**:1333–1339.

Soboleva, T., P. Shorten, A. Pleasants, and A. Rae, 2003. Qualitative theory of the spread of a new gene into a resident population. *Ecological Modelling* **163**:33 – 44.

Stearns, S. C., 1992. *The Evolution of Life Histories.* Oxford University Press, Oxford.

Storlie, C. B. and J. C. Helton, 2008. Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. *Reliability Engineering & System Safety* **93**:28 – 54.

Team, R. D. C., 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tiedje, J., R. Colwell, Y. Grossman, R. Hodson, R. Lenski, R. Mack, and P. Regal, 1989. The planned introduction of genetically engineered organisms - ecological considerations and recommendations. *Ecology* **70**:298–315.

Tuljapurkar, S., 1990. Population dynamics in variable environments: lecture notes in biomathematics no. 85.

Tuljapurkar, S., 1997. Stochastic Matrix Models, pages 59–89. *in* S. Tuljapurkar and H. Caswell, editors *Structured-Population Models in Marine, Terrestrial, and Freshwater Systems*, Population and Community Biology Series. Chapman & Hall.

Tuljapurkar, S., C. C. Horvitz, and J. B. Pascarella, 2003. The many growth rates and elasticities of populations in random environments. *The American Naturalist* **162**:489502.

Wade, P. R., 2000. Bayesian methods in conservation biology. *Conservation Biology* **14**:1308–1316.

Wall, M., 1996. GAlib: A C++ Library of Genetic Algorithm Components, version 2.4, Documentation Revision B. Massachusetts Institute of Technology.

Wikle, C., L. Berliner, and C. N., 1998. Hierarchical bayesian space-time models. *Environmental and Ecological Statistics* **5**:117–154.

Wisdom, M. J., S. L. Mills, and D. F. Doak, 2000. Life stage simulation analysis: Estimating vital-rate effects on population growth for conservation. *Ecology* **81**:628–641.

Yan, J., 2007. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software* **21**:1–21.

Yearsley, J. M. and D. Fletcher, 2002. Equivalence relationships between stage-structured population models. *Mathematical Biosciences* **179**:131–143.