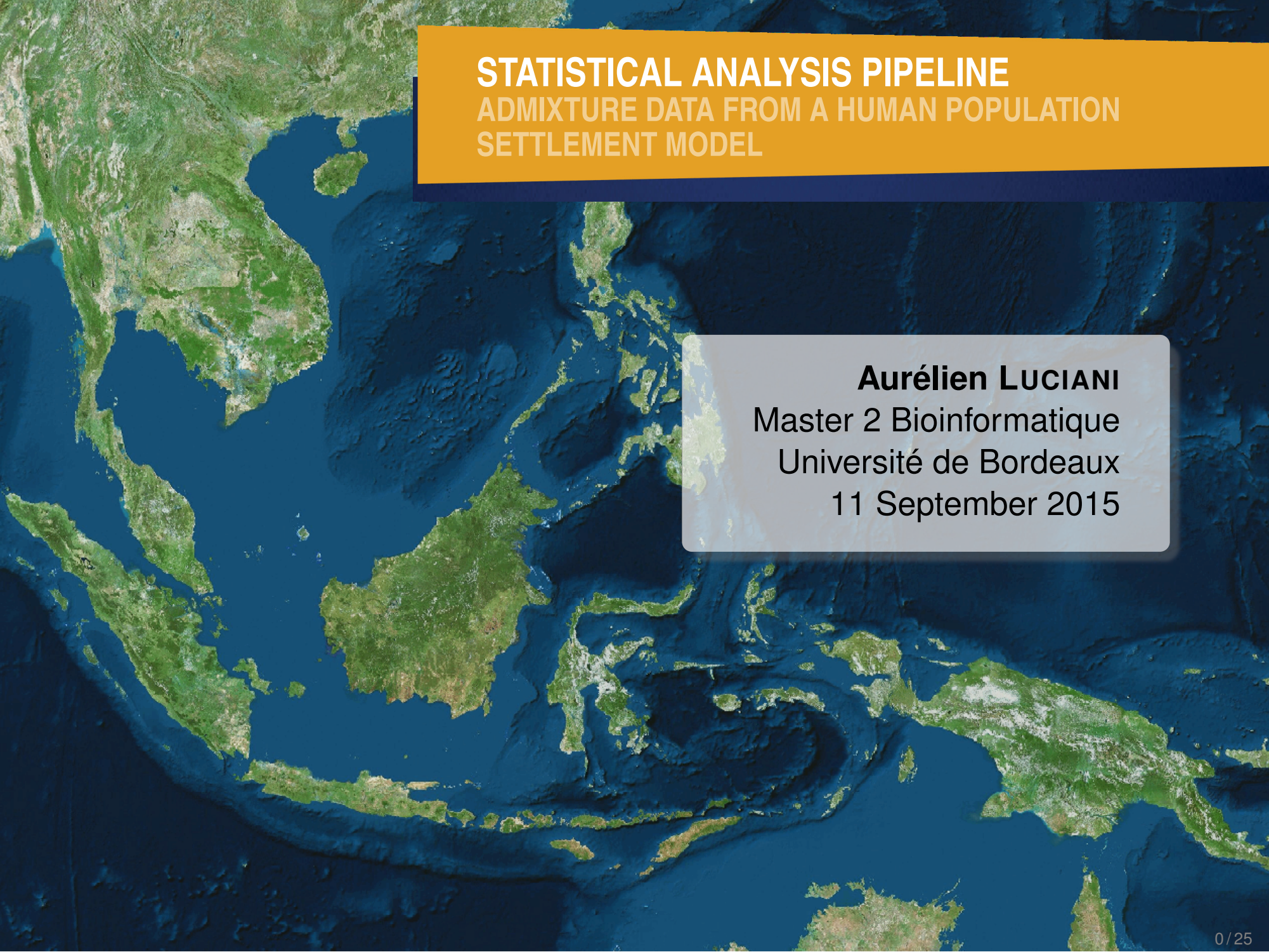


STATISTICAL ANALYSIS PIPELINE

ADMIXTURE DATA FROM A HUMAN POPULATION SETTLEMENT MODEL

Aurélien LUCIANI
Master 2 Bioinformatique
Université de Bordeaux
11 September 2015

INTRODUCTION

- Internship at Massey University (New Zealand)
- Computational Biology Research Group
- Part of a 2 years project
- Supervisors:
  - Murray Cox
  - Marie Noelle Beurton-Aimar

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

TABLE OF CONTENTS

Context
- Geological and anthropological context
- Previous papers
- Model

Measures and comparisons

Pipeline

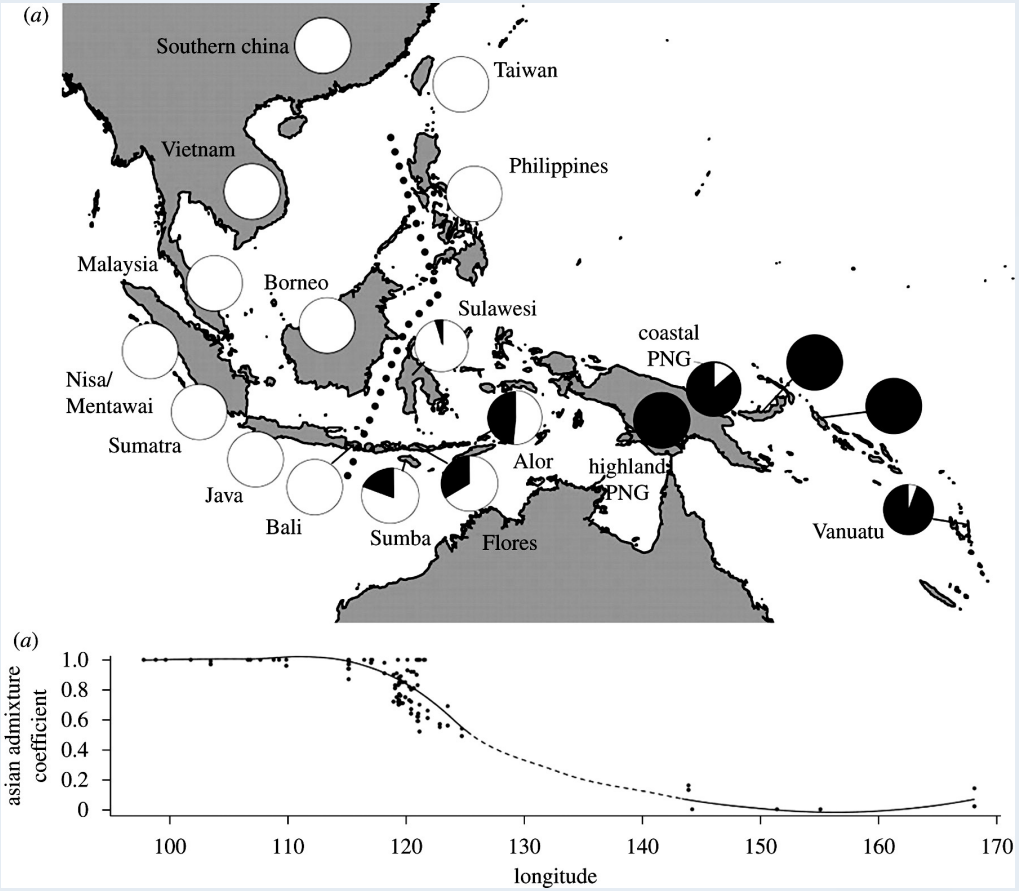
Statistical analysis


FIRST SETTLEMENT WAVE
AROUND 45,000 YEARS AGO
First humans
Ancestral Melanesian population
Hunter-gatherers

SECOND SETTLEMENT WAVE AROUND 4,500 YEARS AGO

- “Austronesian expansion”
- Two possible waves
- Admixture between populations
- What started it?
  - Agriculture knowledge?
  - Better navigation?

CONTEXT



- Non-linear admixture changes

Figure: Cox et al. [2010]. *Proceedings of the Royal Society of London B: Biological Sciences*

CONTEXT

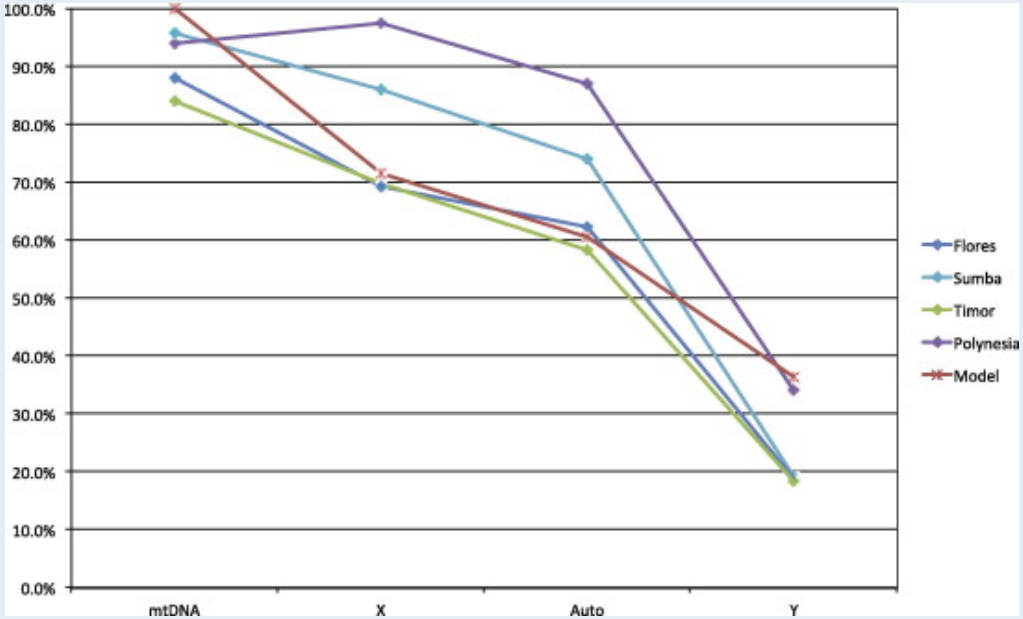


Figure: Lansing et al. [2011]. *Journal of Anthropological Archaeology*

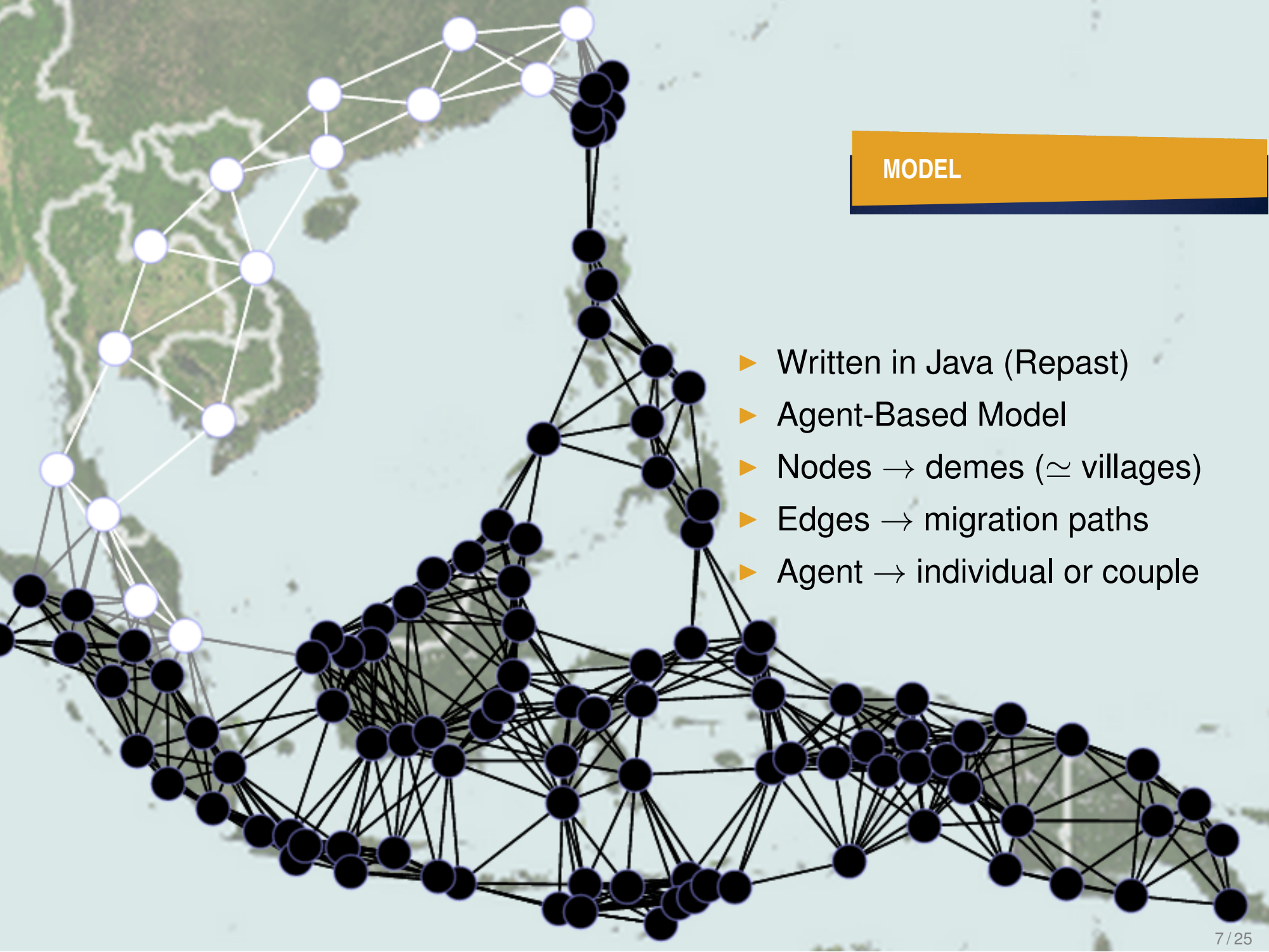
- Difference of admixture by type of DNA
- Gender-biaised admixture
- Implicit marriage rule?
- Melanesian ♂ — Asian ♀ marriages favoured?

PREVIOUS PAPERS

Questions:

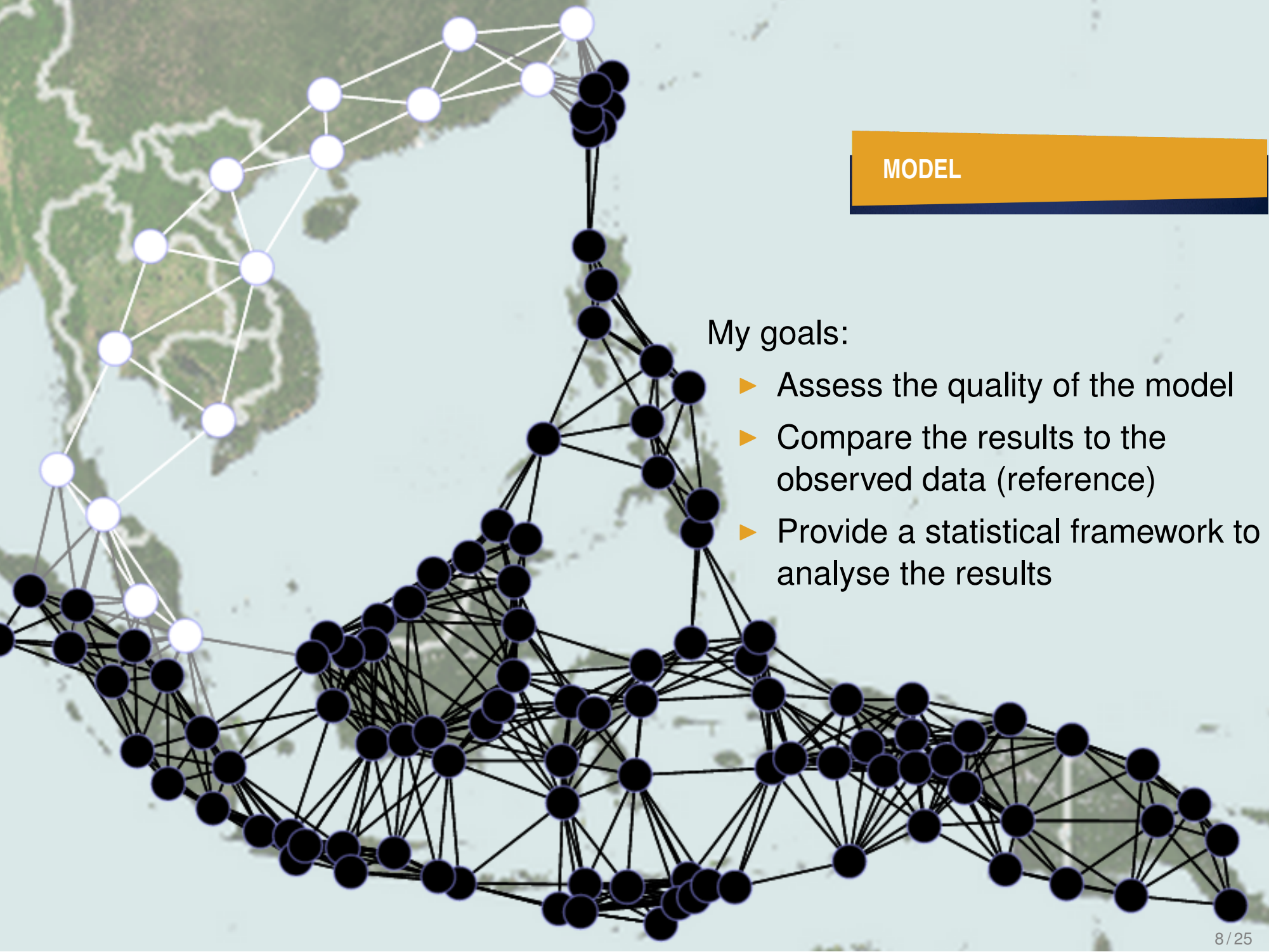
- Why the steep admixture gradient?
- Why the gender-biased admixture?
- What differences between the populations?

First step towards understanding $\rightarrow$ reproduce the scenario



MODEL

- Written in Java (Repast)
- Agent-Based Model
- Nodes $\rightarrow$ demes ($\simeq$ villages)
- Edges $\rightarrow$ migration paths
- Agent $\rightarrow$ individual or couple



MODEL

My goals:

- Assess the quality of the model
- Compare the results to the observed data (reference)
- Provide a statistical framework to analyse the results

TABLE OF CONTENTS

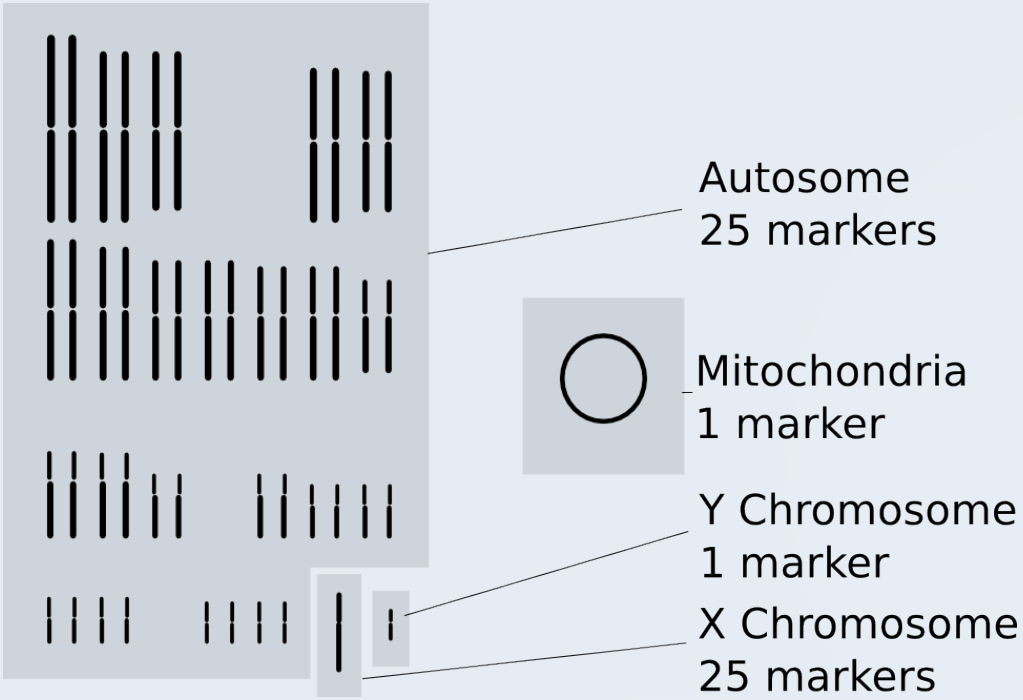
Context

Measures and comparisons
- Observed data
- Parameters
- Comparison functions
  - Mean Square Distance
  - Partial Mantel Correlation
  - Summary statistics

Pipeline

Statistical analysis

MEASURES & COMPARISONS



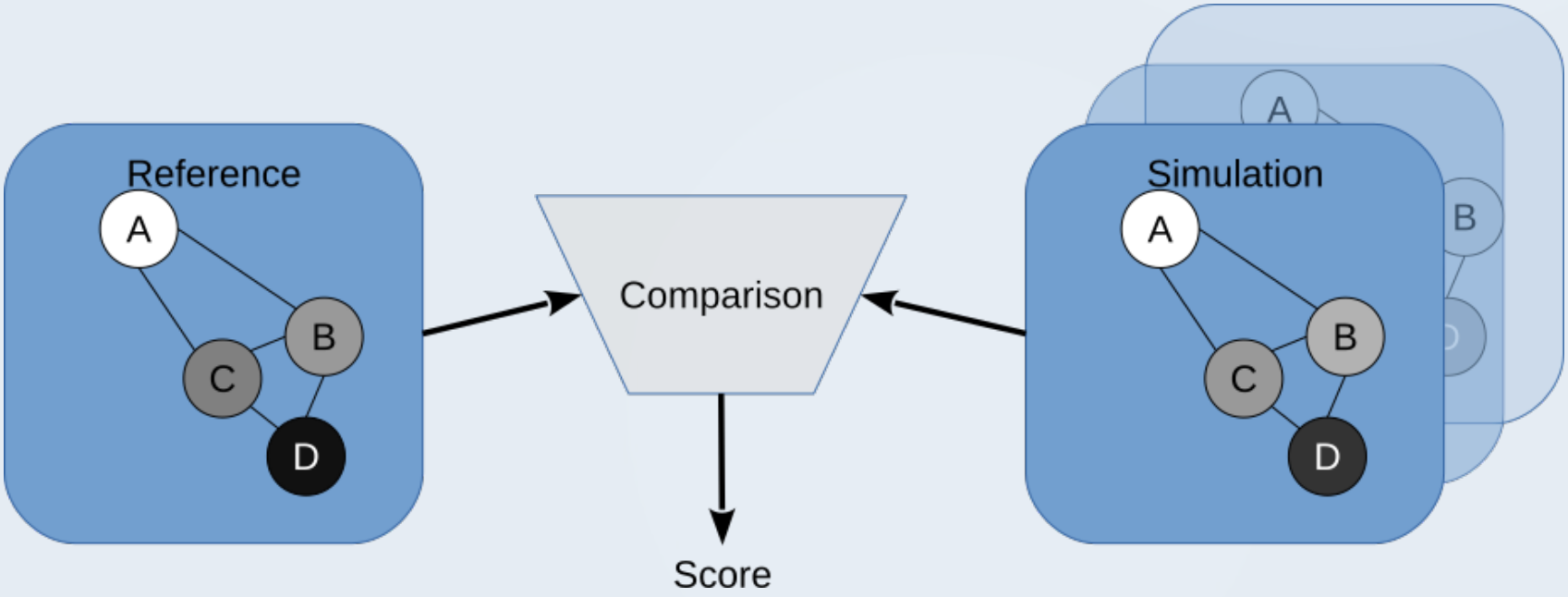
- Different parts of the DNA
- Associated with different origins
- 52 binary markers in total

MEASURES & COMPARISONS

Parameter	Estimated	Comment
Migration prob.	$0.0 < x \leq 1.0$	prob. to start migrating for a Melanesian agent
Migration prob. ratio	$1.0 \leq x \leq 4.0$	corresponding ratio for an Asian agent
Fecundity	$2.5 < x < 8.0$	Poisson law mean for a Melanesian agent
Fecundity ratio	$1.0 \leq x \leq 2.0$	corresponding ratio for an Asian agent
Marriage threshold	$0.0 \leq x \leq 0.25$	affects marriages rules
Growth rate	$0.0 < x < 0.001$	limiting rate of pop. growth
Number of agents	$100 \leq x < 400$	pop. size in each deme, initially
Graph	{...}	composition of the graph (nodes and edges)
Starting distribution	{...}	distribution of pop. in the graph

Table: Summary of the changeable model parameters.

MEASURES & COMPARISONS



MEASURES & COMPARISONS

Mean Square Distance

$$MSD = \frac{\sum_{i=1}^{n} (AdRef_i - AdSim_i)^2}{n}$$

$$0 \leq MSD \leq 1$$

MEASURES & COMPARISONS

Mean Square Distance

$$MSD = \frac{\sum_{i=1}^{n} (AdRef_i - AdSim_i)^2}{n}$$

$$0 \leq MSD \leq 1$$

Example

Island	A	B	C	D
Reference	1.0	0.5	0.4	0.0
Simulated	1.0	0.4	0.3	0.2
Distance	0.0	0.1	0.1	0.2

$$MSD = 0.1$$

MEASURES & COMPARISONS

Mantel Partial Correlation

$$cor = mantel.partial\left(M_{Simulated}, M_{Reference}, M_{geographical}\right)$$

$$-1 \leq cor \leq 1$$

MEASURES & COMPARISONS

Mantel Partial Correlation

$$cor = mantel.partial\left(M_{Simulated}, M_{Reference}, M_{geographical}\right)$$

$$-1 \leq cor \leq 1$$

Example

$$mantel.partial\left(\begin{bmatrix} 0.0 & 0.5 & 0.6 & 1.0 \\ 0.5 & 0.0 & 0.1 & 0.5 \\ 0.6 & 0.1 & 0.0 & 0.4 \\ 1.0 & 0.5 & 0.4 & 0.0 \end{bmatrix}, \begin{bmatrix} 0.0 & 0.6 & 0.7 & 0.8 \\ 0.6 & 0.0 & 0.1 & 0.2 \\ 0.7 & 0.1 & 0.0 & 0.1 \\ 0.8 & 0.2 & 0.1 & 0.0 \end{bmatrix}, \begin{bmatrix} 0.0 & 300 & 250 & 400 \\ 300 & 0.0 & 120 & 150 \\ 250 & 120 & 0.0 & 100 \\ 400 & 150 & 100 & 0.0 \end{bmatrix}\right)$$

$$cor = 0.72$$

MEASURES & COMPARISONS

Summary statistics:

- Mean Square Distance
  - Autosomal admixture
  - X-Chromosomal admixture
- Partial Mantel Correlation
  - Autosomal admixture
  - X-Chromosomal admixture

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

TABLE OF CONTENTS

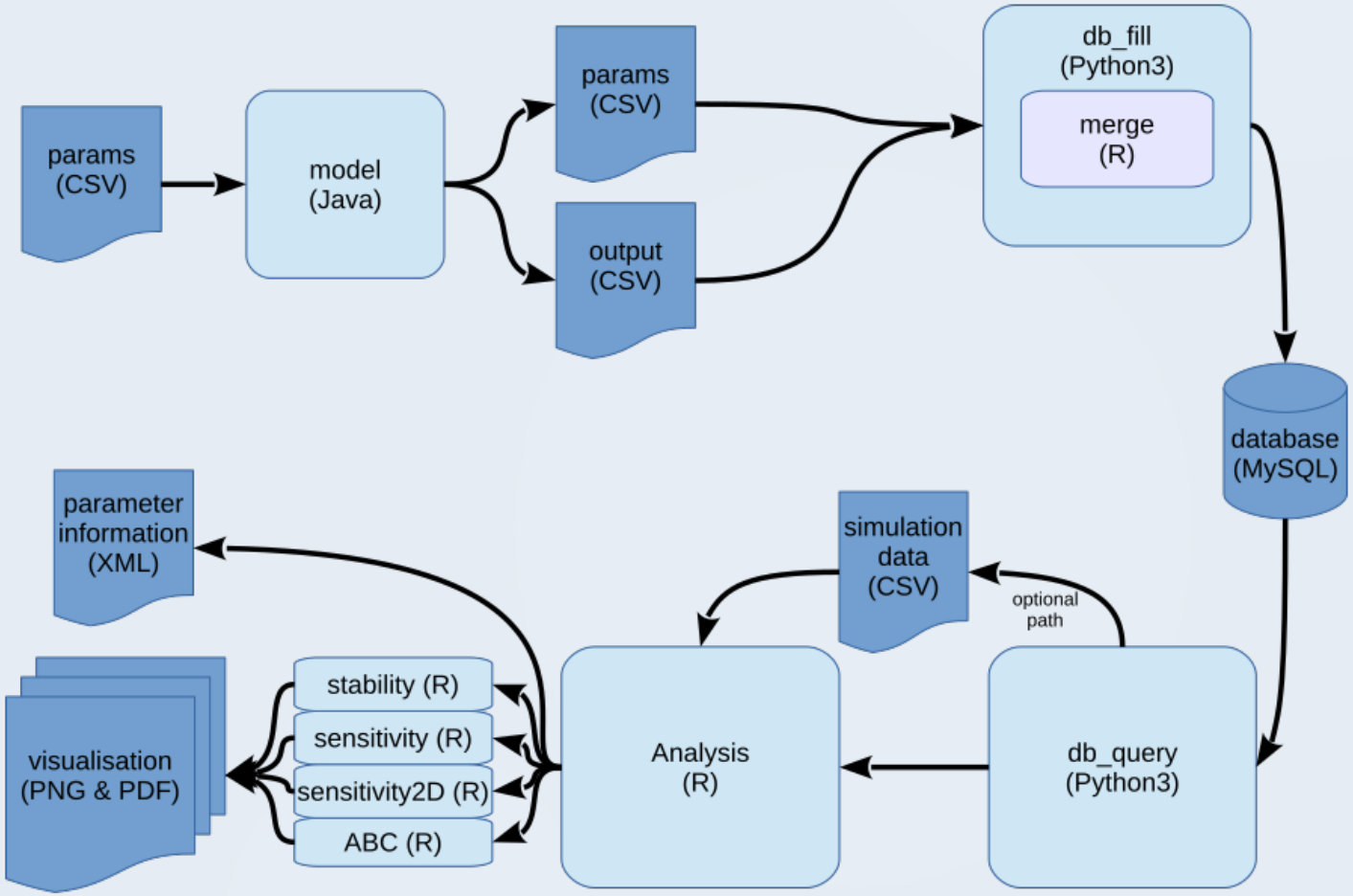
Context

Measures and comparisons

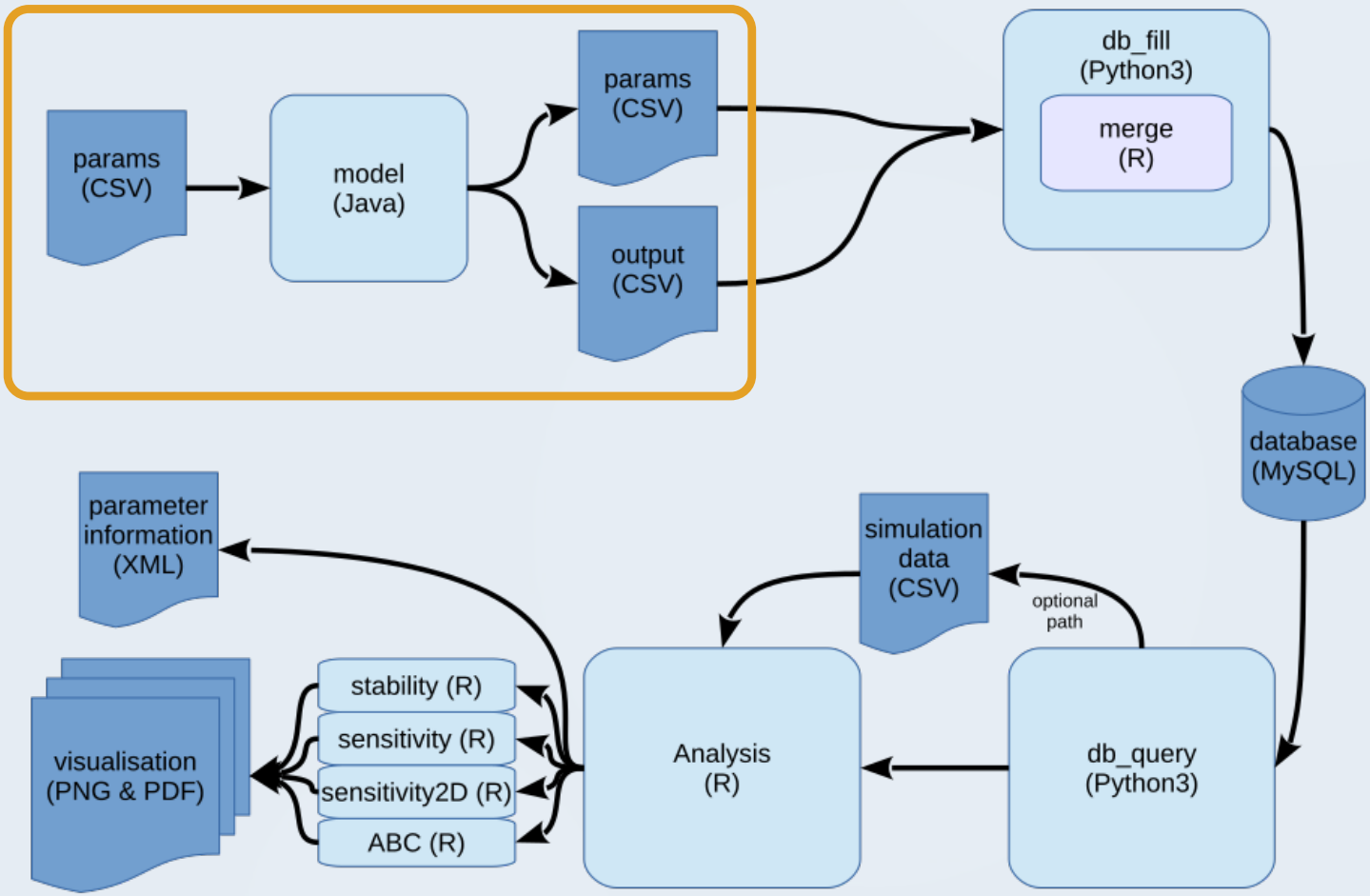
Pipeline
- Overview

Statistical analysis

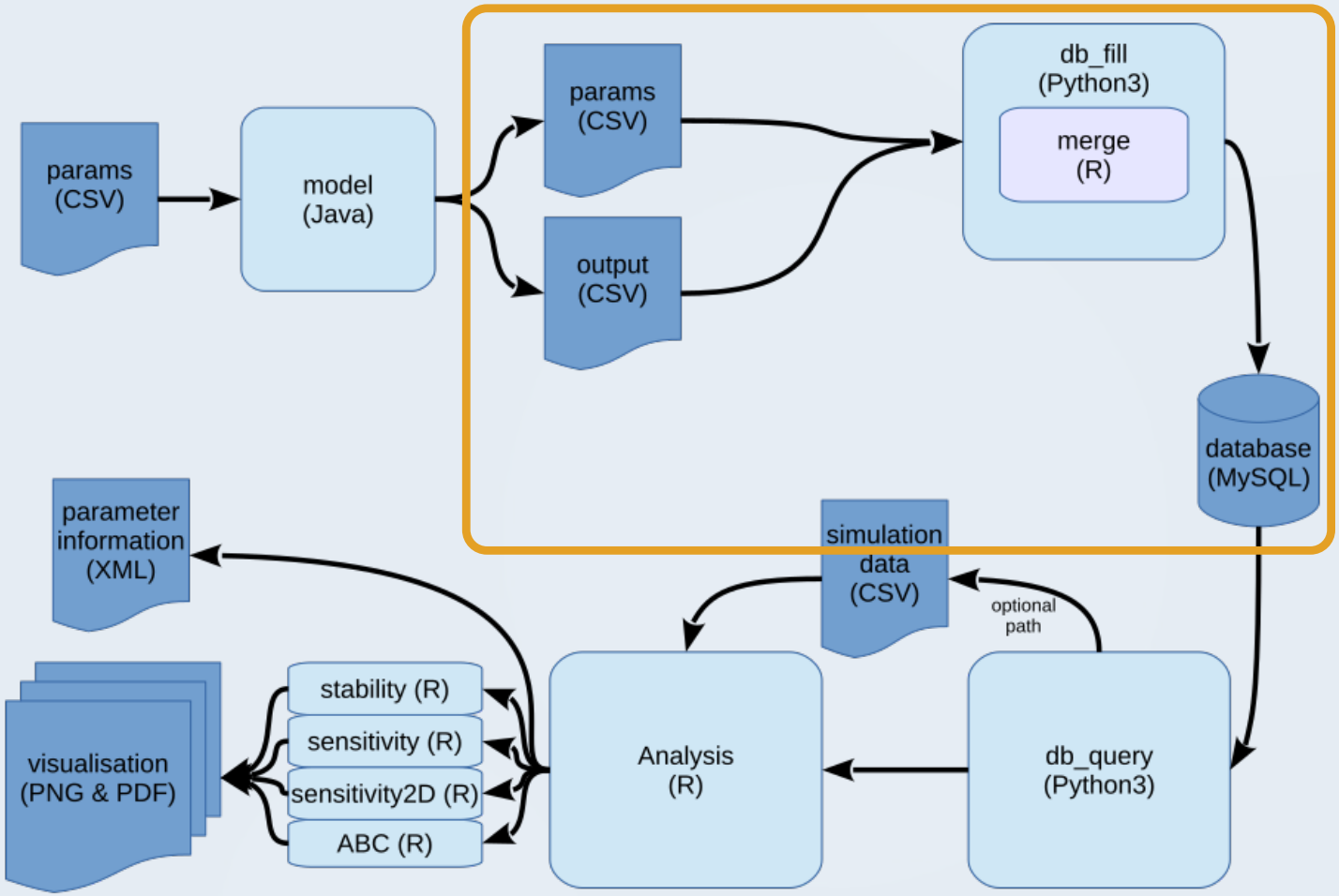
PIPELINE


params
(CSV)
model
(Java)
params
(CSV)
output
(CSV)
db_fill
(Python3)
merge
(R)
database
(MySQL)
parameter
information
(XML)
simulation
data
(CSV)
optional
path
visualisation
(PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis
(R)
db_query
(Python3)

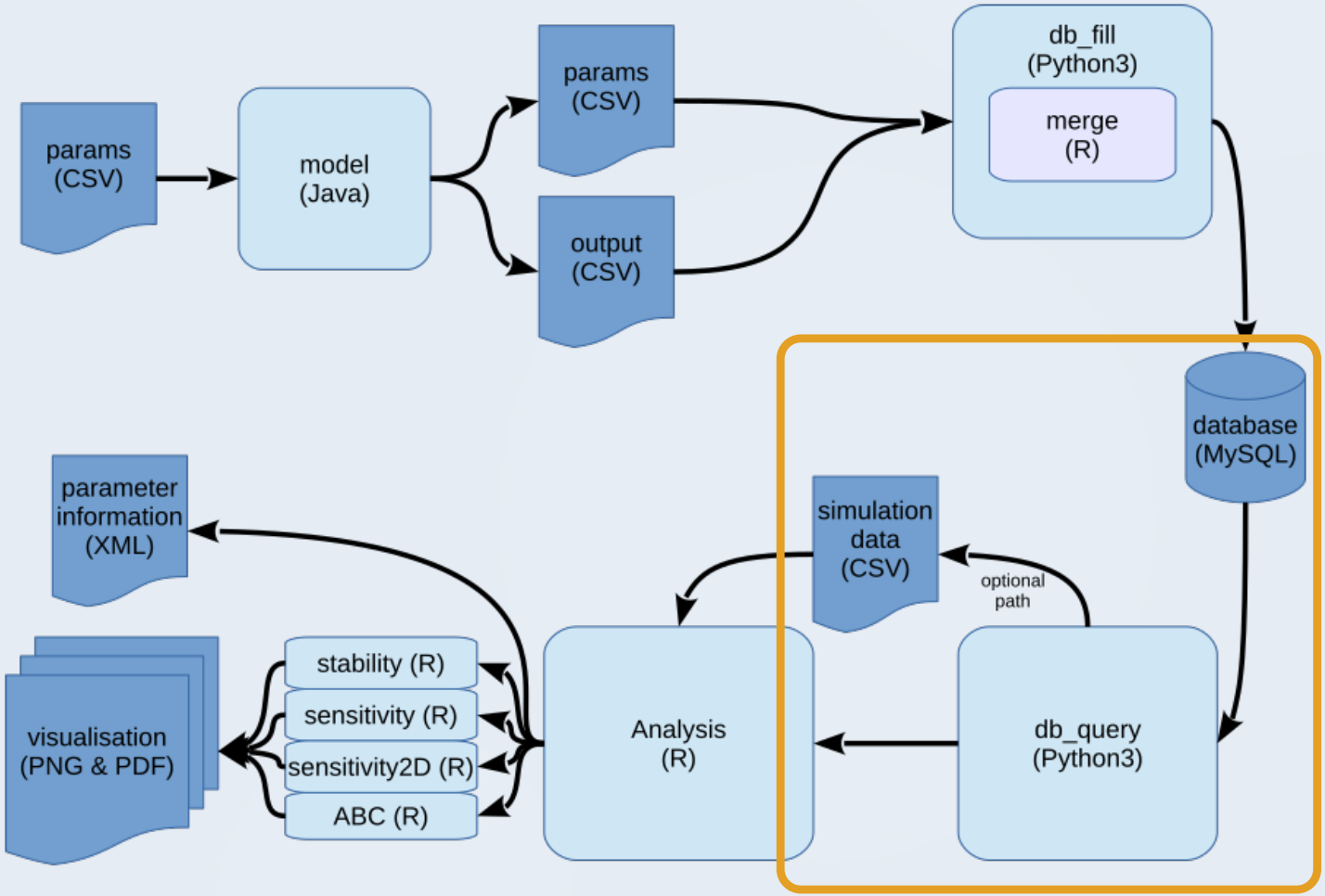
PIPELINE


params
(CSV)
model
(Java)
params
(CSV)
output
(CSV)
db_fill
(Python3)
merge
(R)
database
(MySQL)
parameter
information
(XML)
simulation
data
(CSV)
optional
path
visualisation
(PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis
(R)
db_query
(Python3)

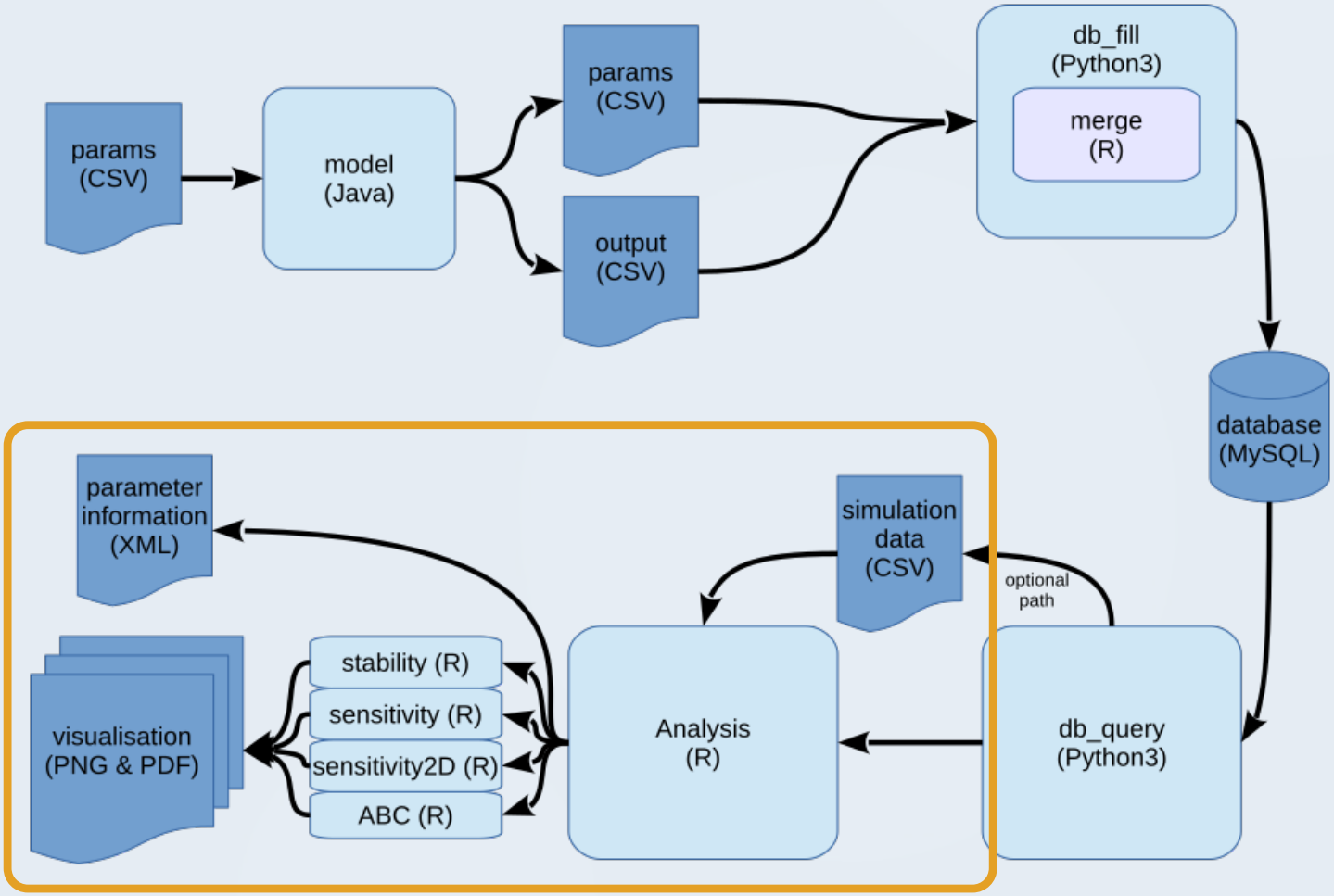
PIPELINE


params (CSV)
model (Java)
params (CSV)
output (CSV)
db_fill (Python3)
merge (R)
database (MySQL)
parameter information (XML)
simulation data (CSV)
optional path
visualisation (PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis (R)
db_query (Python3)

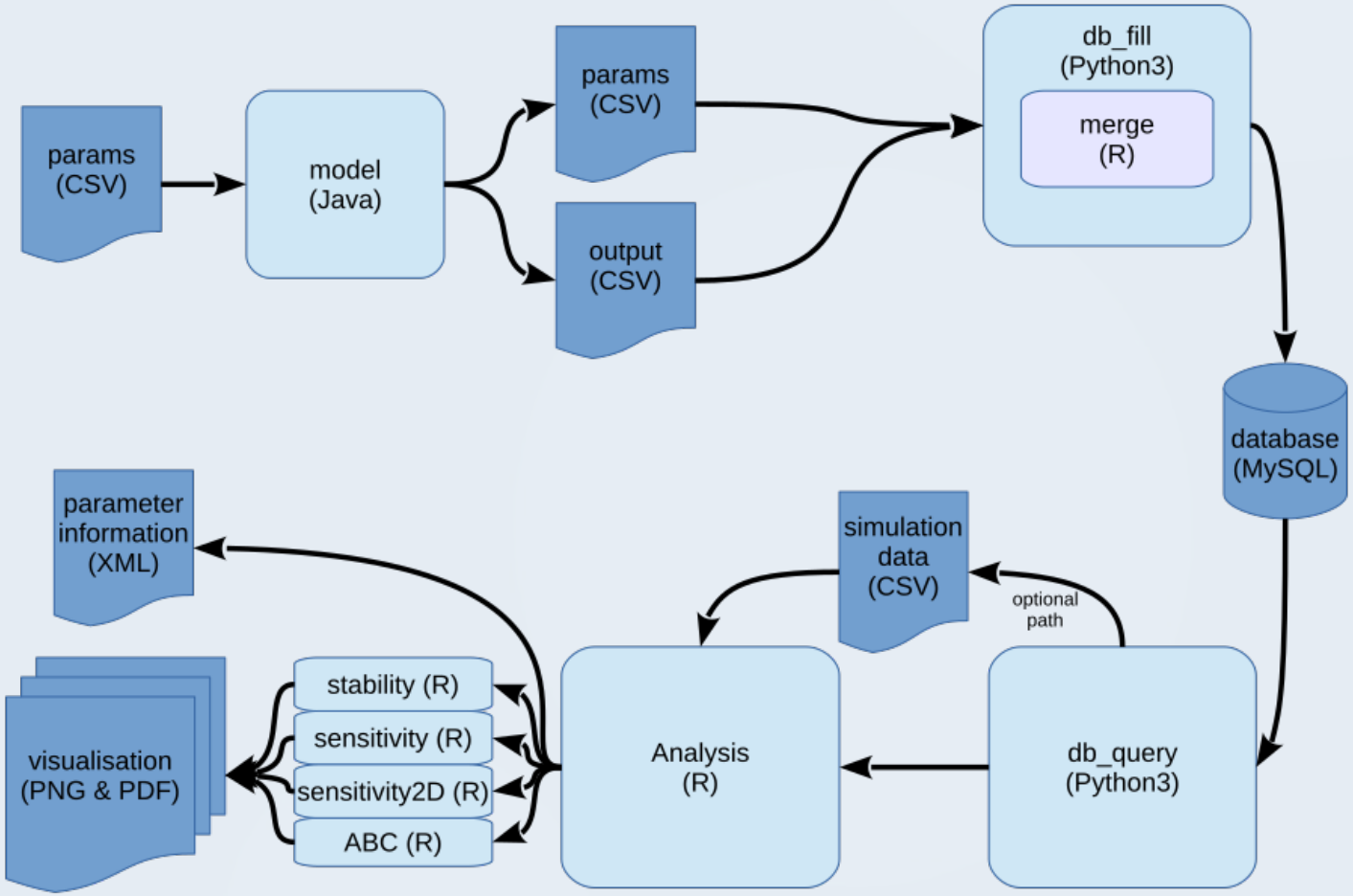
PIPELINE


params
(CSV)
model
(Java)
params
(CSV)
output
(CSV)
db_fill
(Python3)
merge
(R)
database
(MySQL)
parameter
information
(XML)
simulation
data
(CSV)
optional
path
visualisation
(PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis
(R)
db_query
(Python3)

PIPELINE


params
(CSV)
model
(Java)
params
(CSV)
output
(CSV)
db_fill
(Python3)
merge
(R)
database
(MySQL)
parameter
information
(XML)
simulation
data
(CSV)
optional
path
visualisation
(PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis
(R)
db_query
(Python3)

PIPELINE


params
(CSV)
model
(Java)
params
(CSV)
output
(CSV)
db_fill
(Python3)
merge
(R)
database
(MySQL)
parameter
information
(XML)
simulation
data
(CSV)
optional
path
visualisation
(PNG & PDF)
stability (R)
sensitivity (R)
sensitivity2D (R)
ABC (R)
Analysis
(R)
db_query
(Python3)

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

TABLE OF CONTENTS

Context

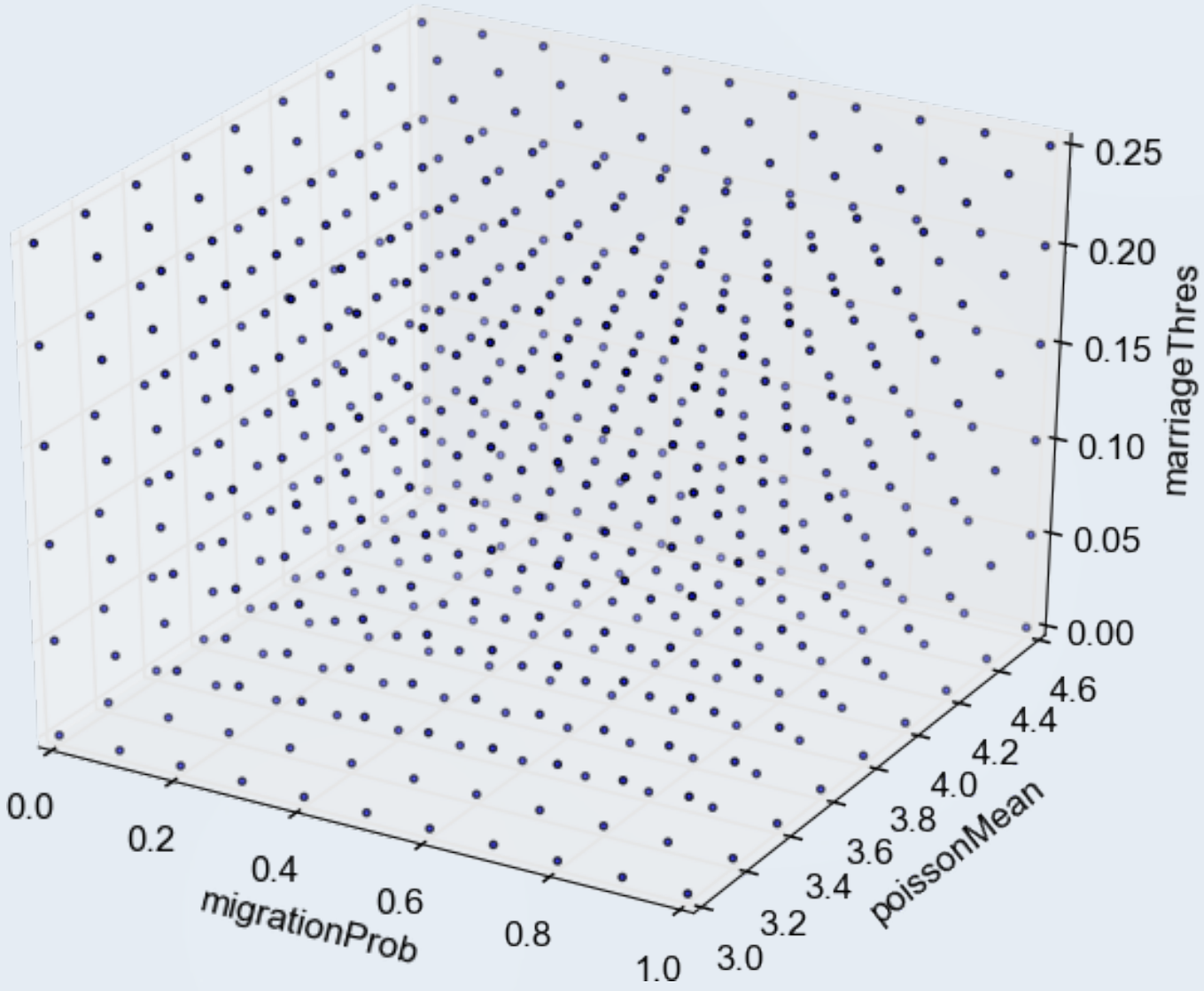
Measures and comparisons

Pipeline

Statistical analysis
- Grid search
- Approximate Bayesian Computation

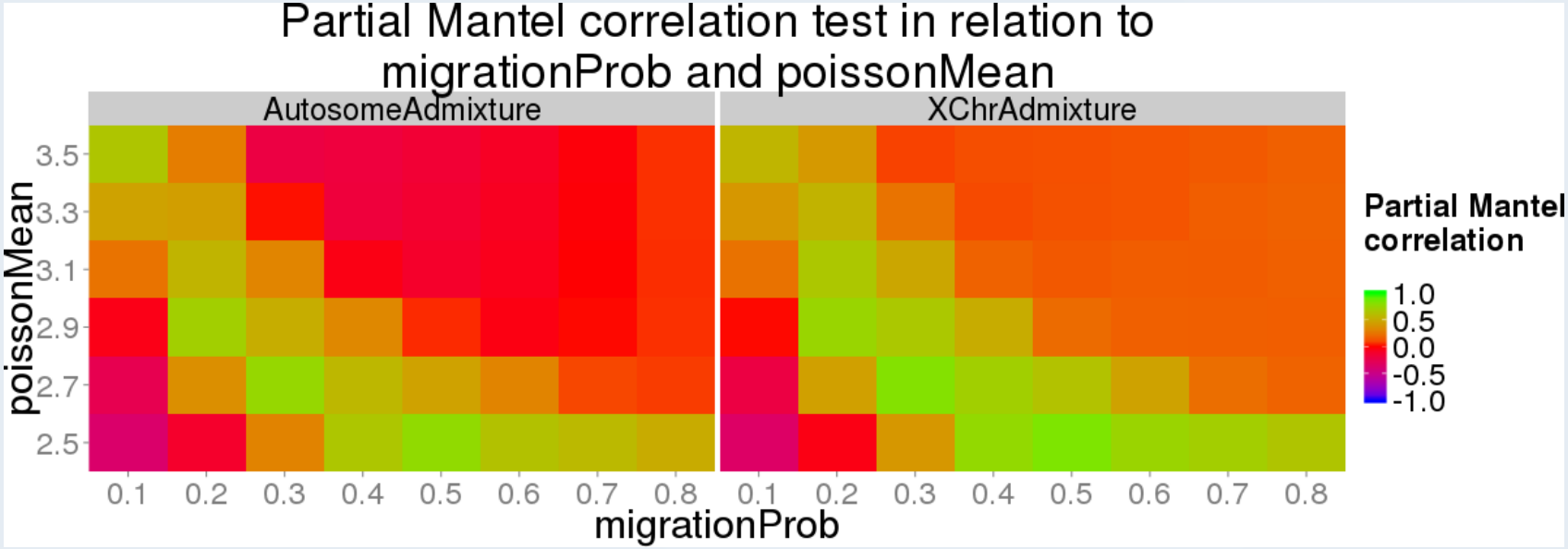
STATISTICAL ANALYSIS

GRID SEARCH

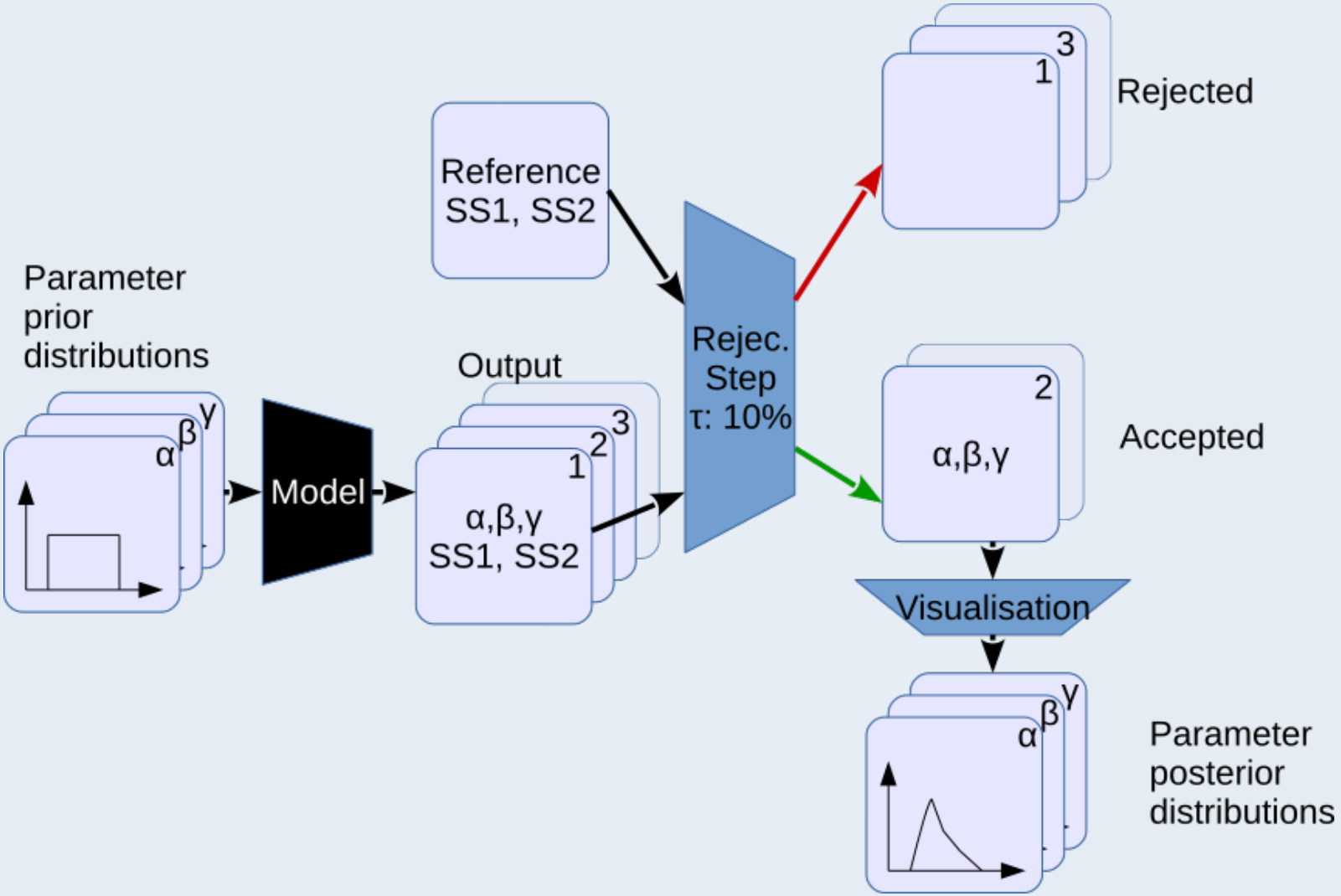


STATISTICAL ANALYSIS

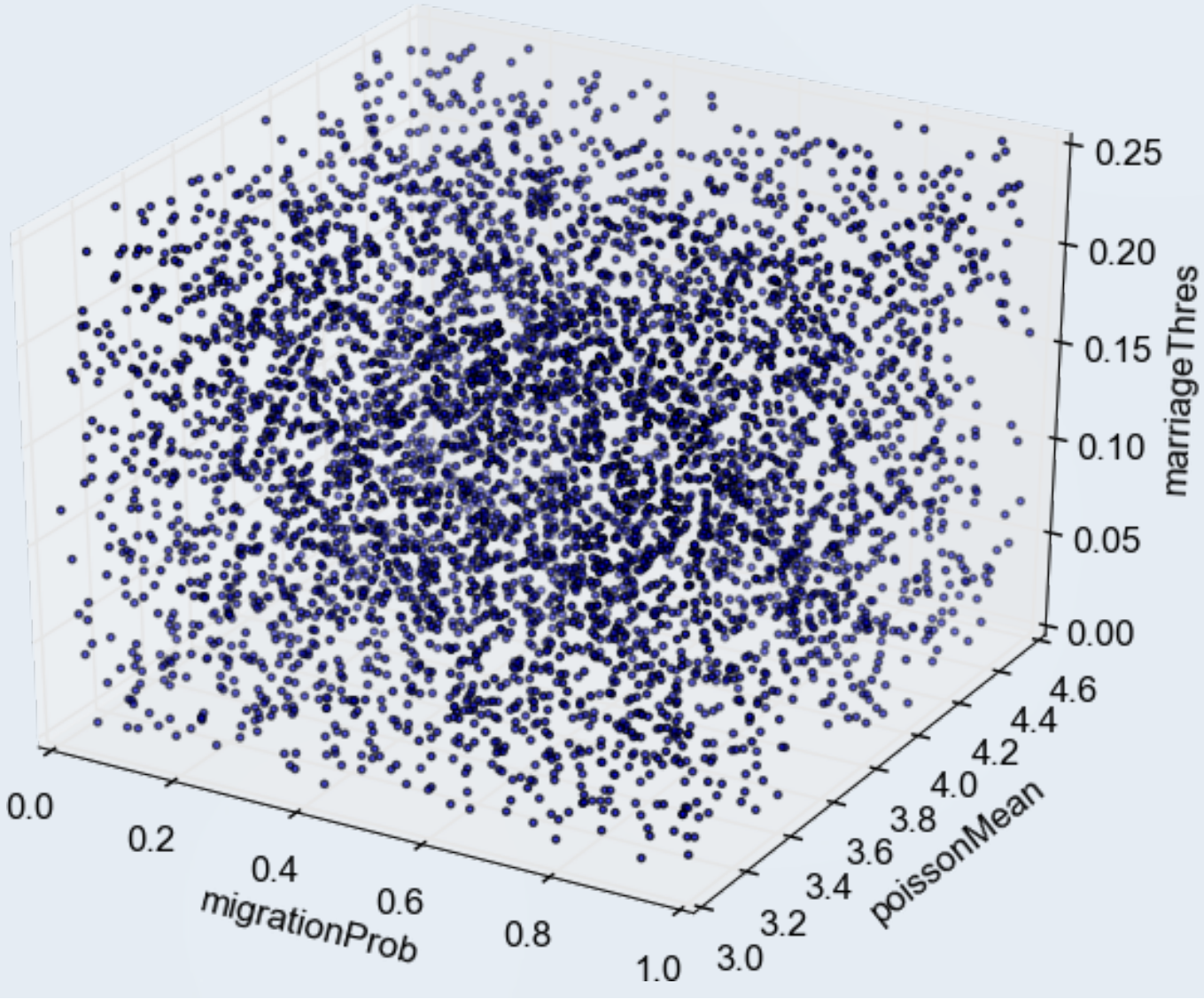
GRID SEARCH



STATISTICAL ANALYSIS



STATISTICAL ANALYSIS



STATISTICAL ANALYSIS

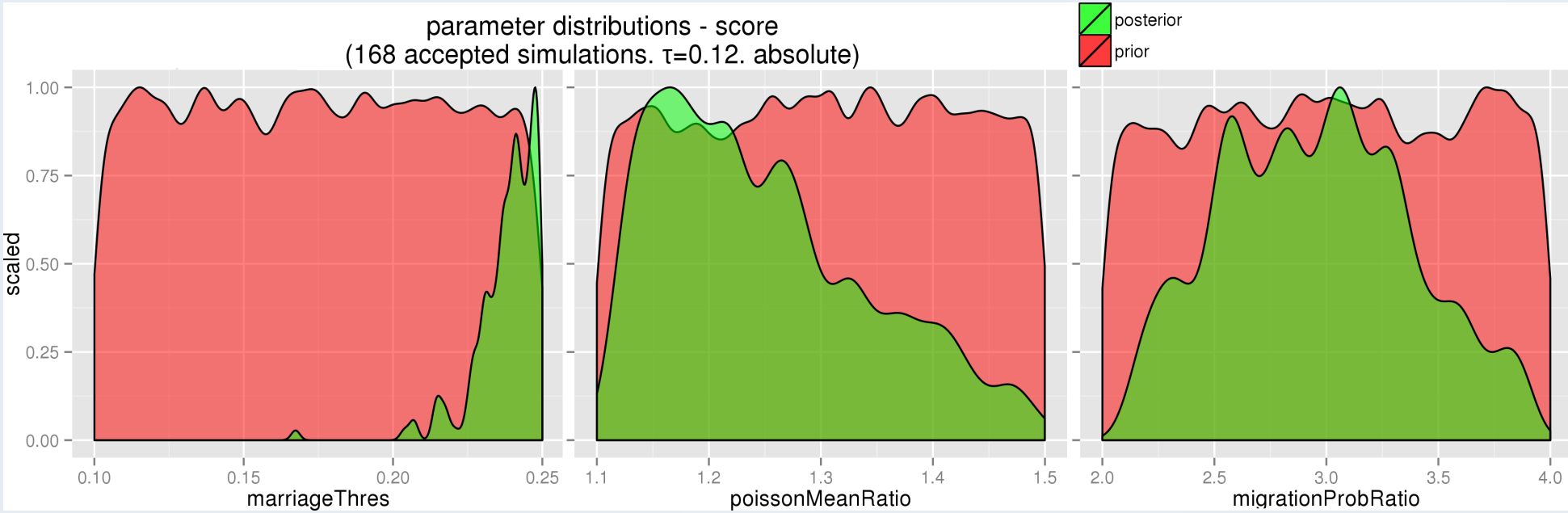


Figure: Example output from an ABC analysis

CONCLUSION

- Pipeline:
  - Modular pipeline
  - Stream processing
  - Database use
- Statistically powerful approach:
  - Approximate Bayesian Computation
  - Accept good results
  - Discard bad results
- “Austronesian Expansion”:
  - Higher fecundity
  - Higher migration rates
  - Melanesian ♂ — Asian ♀ marriages favoured

THANK YOU

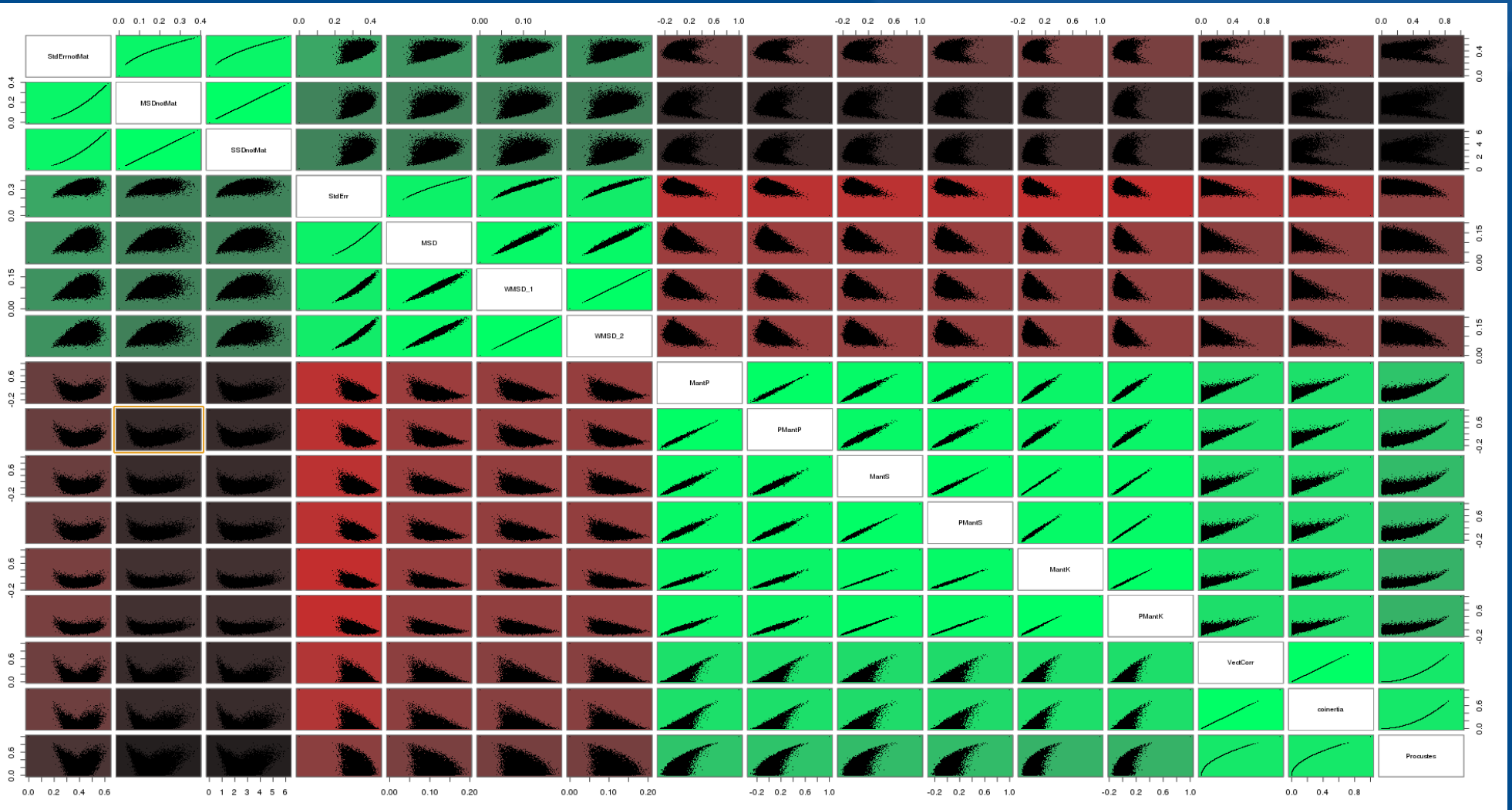
- Context
  - Geological and anthropological context
  - Previous papers
  - Model
- Measures and comparisons
  - Observed data
  - Parameters
  - Comparison functions
- Pipeline
  - Overview
- Statistical analysis
  - Grid search
  - Approximate Bayesian Computation

Questions?

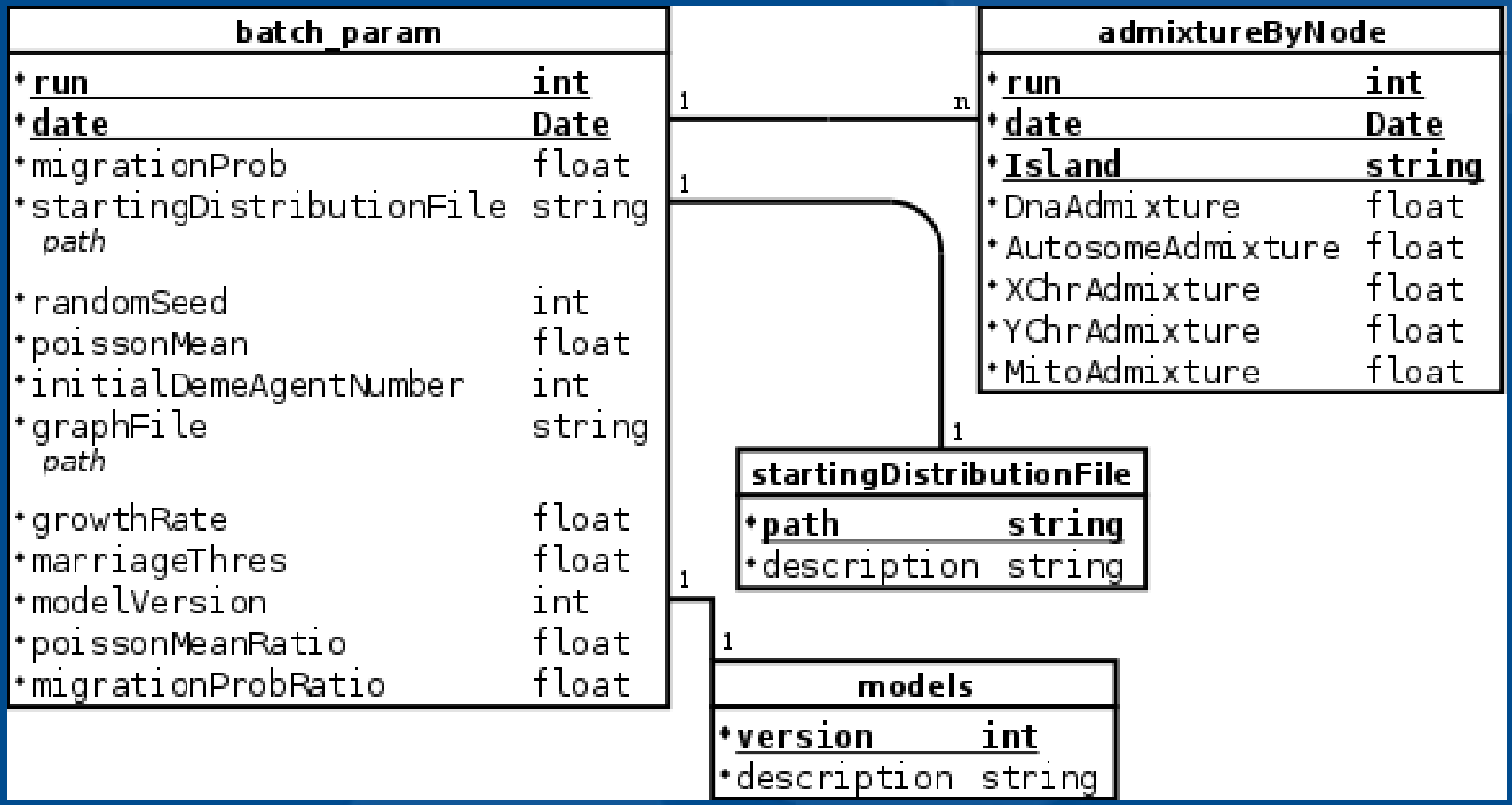
map backgrounds from “HERE Satellite”

Computational Biology Research Group:
`massey.genomicus.com`

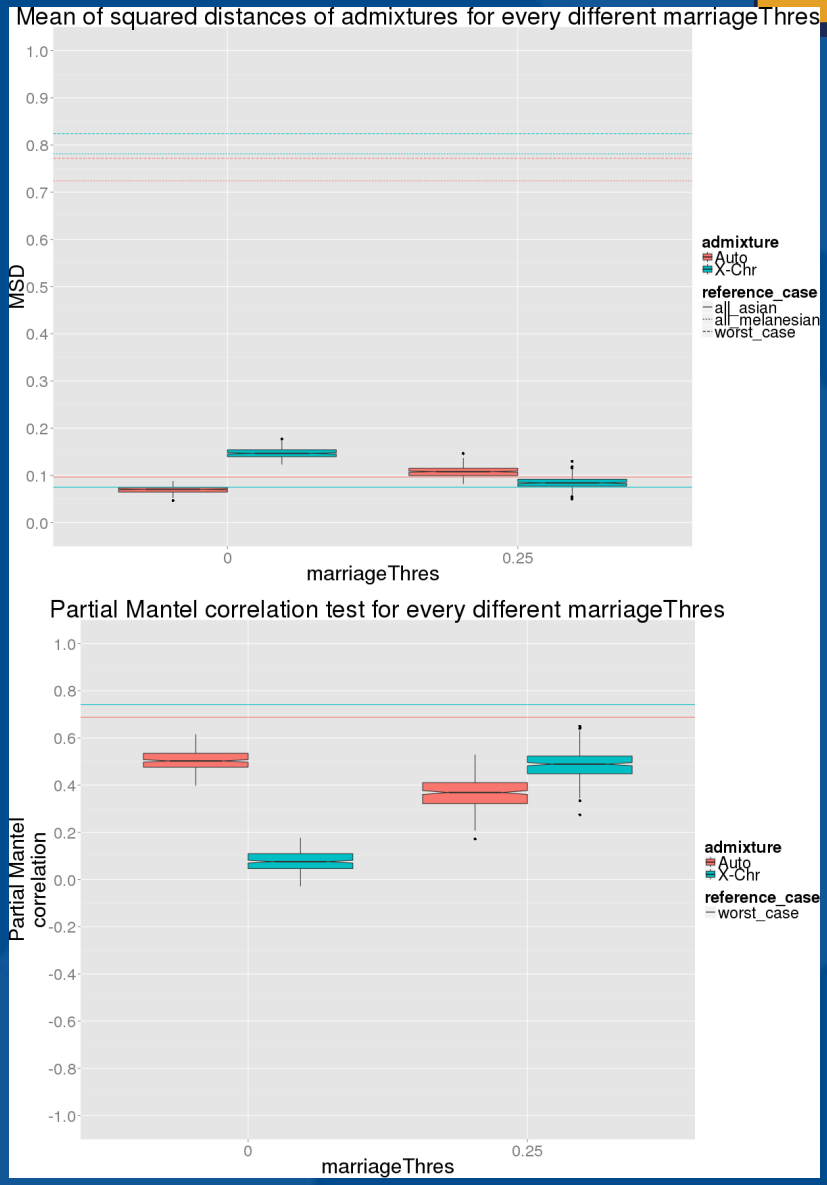
COMPARISON FUNCTIONS



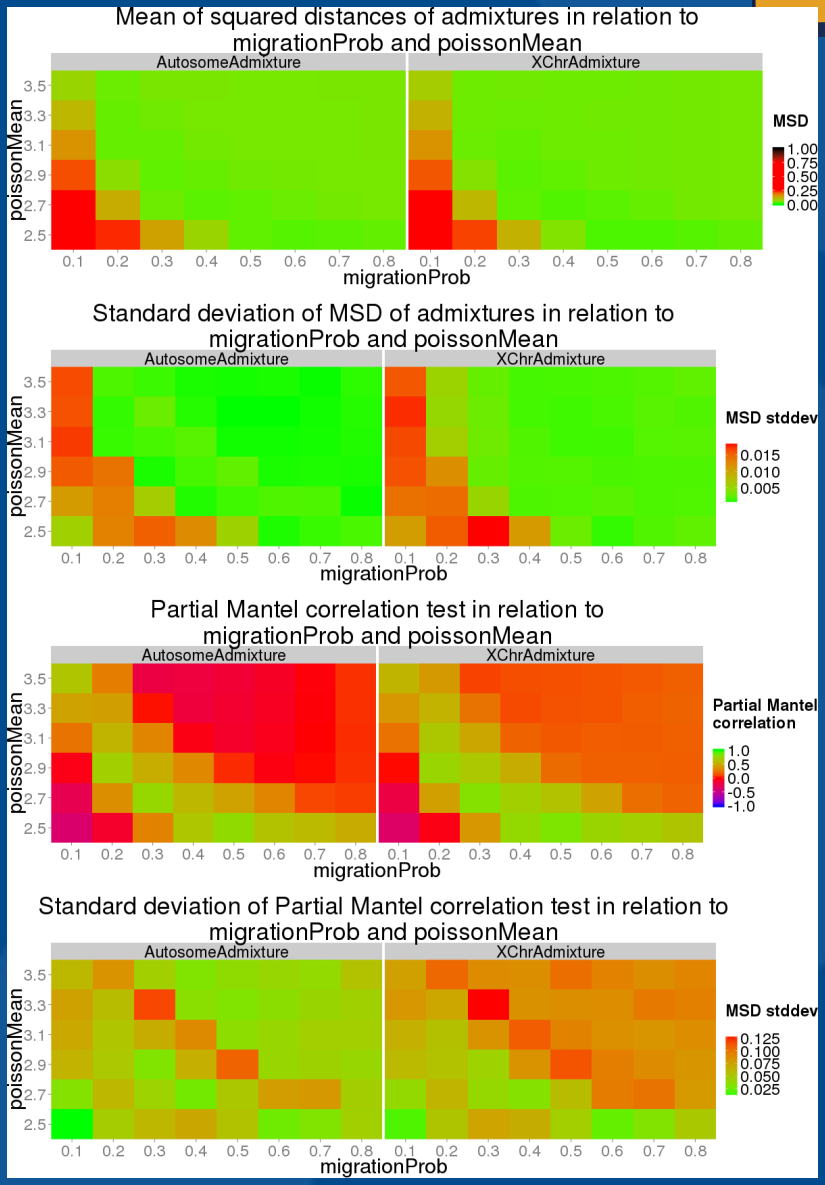
DATABASE STRUCTURE



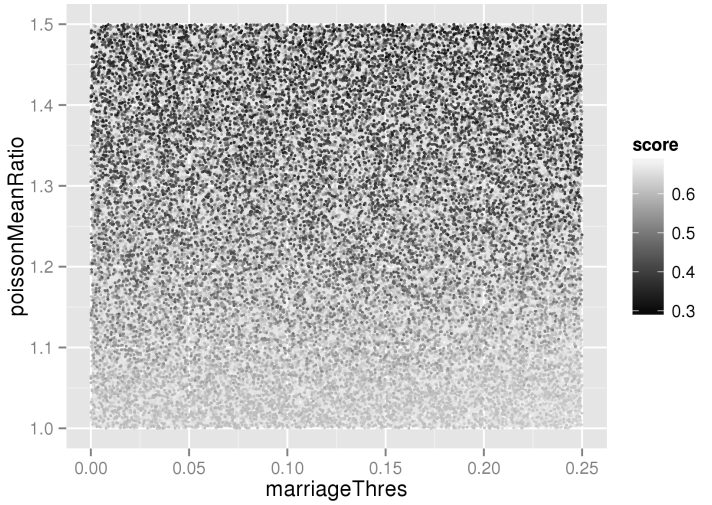
SENSITIVITY 1D - COMPARISONS

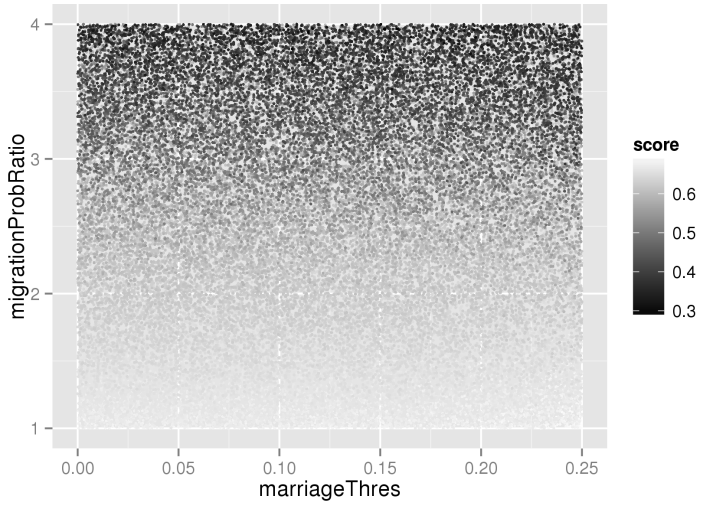


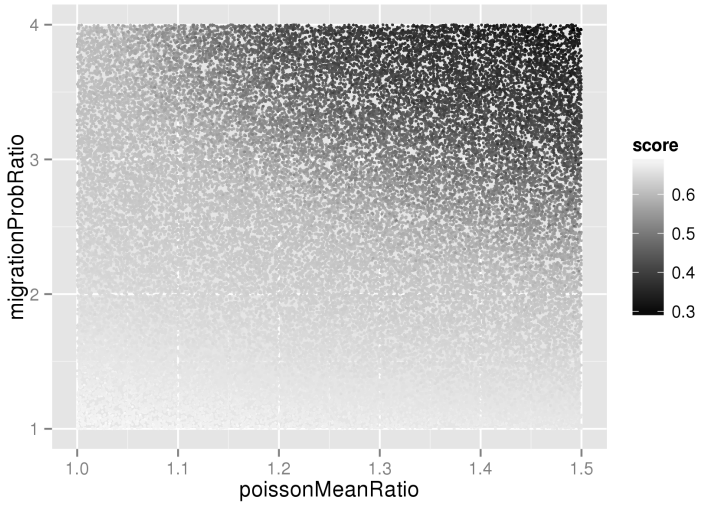
SENSITIVITY 2D - COMPARISONS



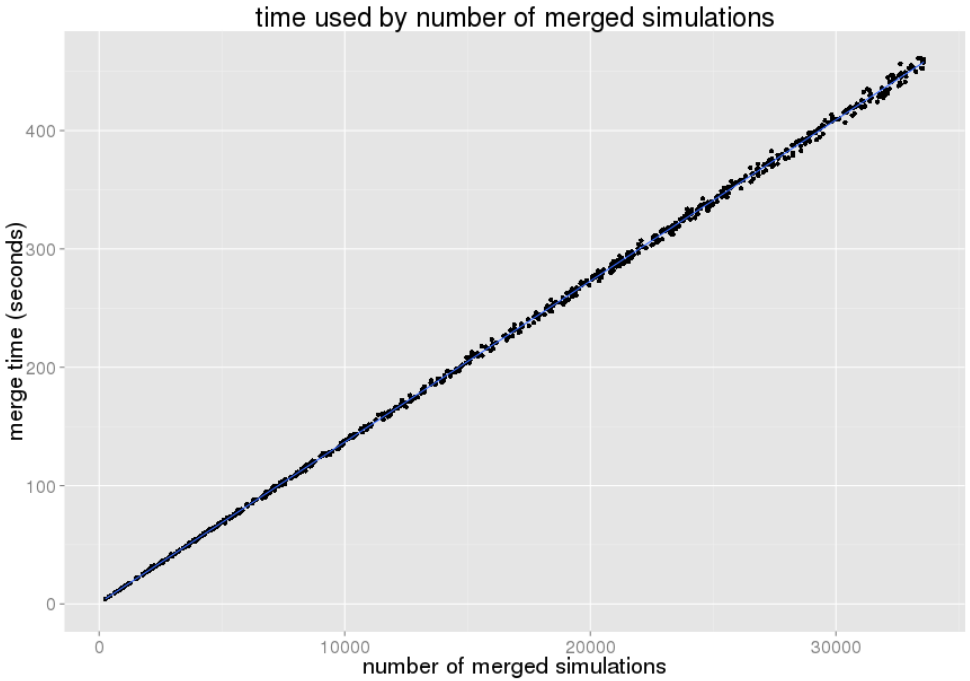
ABC SCATTER SCORE

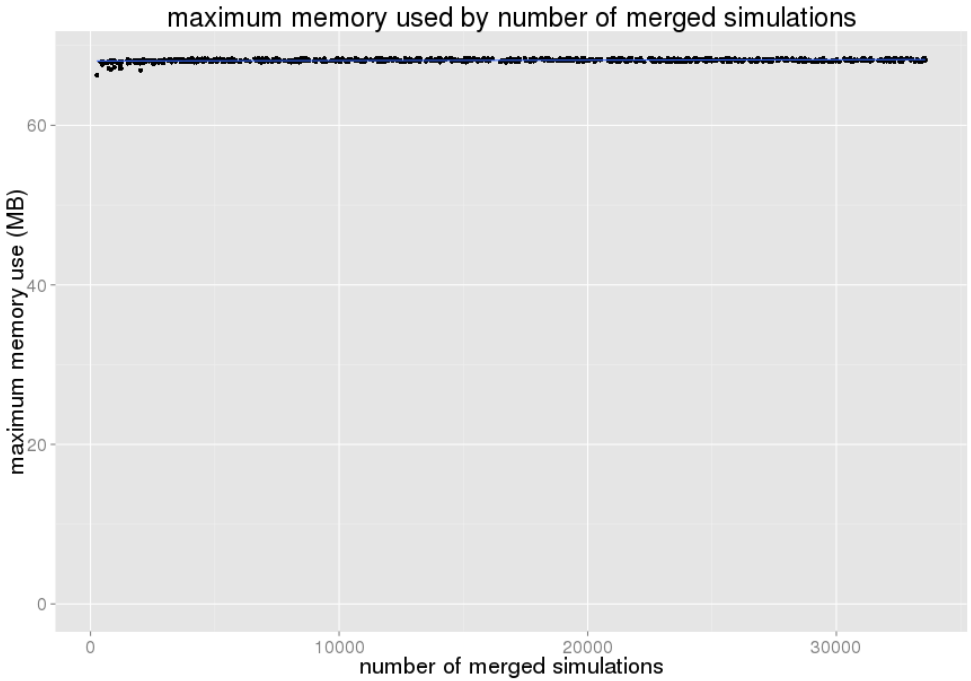






BENCHMARK MERGE





BENCHMARK ANALYSIS & ABC