# Survey of Evolutionary Models and Methods of Incorporating Space

A. Papula
*MIT Department of Physics*
(Dated: May 22, 2017)

Study of population dynamics finds valuable applications including fighting disease (the evolution of cancer cells as they outpace one's own immune system; the co-evolution of HIV and the immune system of the infected) and attempting to protect ecosystems as imbalances are accelerated due to climate change, in addition to academic interest. Quantitatively, evolution and the relative importance of various evolutionary processes is as yet poorly understood, and at this point there does not exist a realistic model that enjoys widespread success. This review begins by summarizing classical models and well-mixed (effectively zero-dimensional) populations and then discusses several models which attempt to incorporate the spatial environments in which evolution occurs. The models discussed are neutral, meaning that there is assumed to be no significant differences in the fitness associated with the different alleles. All models in this paper unless otherwise stated will be implicitly for non-mutating *haploid, asexual* organisms (which carry only a single copy of their genome) without recombination, though some of the models described can be extended to diploid populations and many of the papers referenced have introduced mutation and/or recombination as extensions to their models.

Evolution is driven by several distinct processes including mutation, recombination, interaction with spatial structure, natural selection, and genetic drift. In natural selection, the relative fitness of a particular genotype (with higher fitness defined by relatively higher birth rate and/or lower death rate) deterministically drives the population to higher fitness; frequencies of alleles change over time in accordance to the relative fitness advantages or disadvantages they convey. Genetic drift introduces stochasticity into population dynamics; it is the variation in relative frequencies of different genotypes in the population due purely to chance. Because populations comprise of discrete individuals, though births and deaths may be characterized by an average rate they will be inherently subject to randomness, which will contribute to genetic drift. The discreteness of individuals also allows for fixation, when the frequency of a certain allele reaches 1; for each variety of a given allele, frequency $f = 1$ and $f = 0$ are absorbing boundary conditions which, once reached, are featured forever by the population in the assumption of no mutations. In a population distributed over a spatial continuum, an example of genetic drift is a sudden localized illness killing only the individuals living (by chance) in a particular region during a particular window of time. A genotype which developed resistance to the illness is an example of natural selection.

## I. MODEL REQUIREMENTS

Useful models must make predictions that can be compared to the DNA sequence data currently available. These data sets typically consist of the sequences of non-coding regions of a gene, drawn from contemporaries sampled from a given population. Sampled sequences are used to construct an estimate of the genealogical tree describing the ancestry of the gene backwards in time. A non-coding region of the gene is preferable for constructing such relations because it is unlikely to affect fitness, and thus mutations in non-coding regions accumulate in a population at a more constant rate over time; they do not change the functionality of a protein or the frequency at which it is transcribed (unless they introduce an extraneous start or stop codon, in which case the mutation should not accumulate). Thus, the likelihood of establishment of such a non-coding mutation in a population has no fitness-based basis for variation from one mutation to the next [1]. If there is a discrepancy in such a non-coding region between two individuals, the minimum length of time to their most recent common ancestor can be estimated.

In order to compare these data sets to a model describing the propagation of allele frequencies in an evolving population forwards in time, the model must have a backwards-in-time analog characterizing the form of the genetic tree relating contemporaries sampled without replacement from the current population. This backwards-in-time model (termed a *coalescent*) must predict the time scales on which lineages merge (as viewed backwards in time; a merging event in a coalescent corresponds to a branching event (such as multiple offspring) in the forwards-in-time model) and the nature of such mergers (how frequently more than two branches merge at the same node and how frequently multiple independent mergers occur at the same time step).

Other forms of data to which models can be compared include the "probability of identity" of alleles of contemporary individuals sampled at a known spatial separation. This is the likelihood that the two individuals share the same allele at the given genetic locus. Models of population dynamics that incorporate a spatial structure should be able to predict the correlation functions of alleles of two individuals as a function of their spatial separation.

## II. KINGMAN AND CLASSICAL APPROACHES

The Kingman coalescent is one of the greatest successes of coalescent theory. It is derived from a very simple population model with assumptions that make in unsuitable for any population in nature or in a laboratory. These assumptions are:

- all labels that can be assigned to members of the population (such as allele, geographic location, time since most recent branching, etc) have no effect on the number of offspring that individual produces (the probability distributions for number of offspring are identical for the members of a population)
- the population size $N$ is constant in time

The number of offspring of contemporaries in a population are exchangeable random variables; identical but not independent.

To describe the ancestry of a sample of $n$, $n \leq N$ contemporaries, define $t'$ as the time variable increasing backwards in time (giving positive values to events in the past), with $t' = 0$ when the sample is drawn. The assumption $n \ll N$ is used in allowing only a single, two-lineage merge per generation. That is, no more than two lineages merge at any given node in the tree, and no more than one such merge happens in any single generation. This shall be called a "simple" merger. Define $\mathrm{T}_n$ as the value of $t'$ at with the first merging event occurs among the sample of $n$ lineages; $\mathrm{T}_{n-1} = t' - \mathrm{T}_n$ as the time interval until the second, $\mathrm{T}_{n-2} = t' - \mathrm{T}_{n-1} - \mathrm{T}_n$ as the time between the third and second merger backwards in time, etc, until the sample has merged into a single most recent common ancestor (MRCA) at $t' = \mathrm{T}_2 - \sum_{i=3}^{n} \mathrm{T}_i$. In the limit $N \to \infty$, $n \ll N$, the time intervals $\mathrm{T}_i$ of a population evolving under a model that meets Kingman's criteria are distributed exponentially according to the following probability density functions[2]:

$$p_{\mathrm{T}_i}(t') = \binom{i}{2} \exp\left(-\binom{i}{2} t'\right), \ \langle \mathrm{T}_i \rangle = \frac{2}{i(i-1)}$$

where $t'$ is scaled such that one unit of $t'$ is equal to the time of $N$ generations. The expected amount of time for a sample size $n$ to coalesce to a single common ancestor is

$$\langle T \rangle = \sum_{i=2}^{n} \langle \mathrm{T}_i \rangle = \frac{2(n-1)}{n} = 2 - \frac{2}{n}$$

corresponding to $(2 - \frac{2}{n}) \times N$ generations of a population of total size $N$. A brief description of two models to which the Kingman coalescent can be applied, in the limit $N \to \infty$, follows.

### II.1. The Wright-Fisher Model

The neutral Wright-Fisher model uses a constant population size $N$ of individuals with no relative fitness difference in which generations are discrete and strictly do not overlap. There is no mutation; offspring have the same genotype as their parent, and the parent of each offspring in generation $(t+1)$ is independently and randomly sampled (with replacement) from the individuals living in generation $t$. $N$ offspring are born every generation. Thus, if an allele is present in $m_t$ individuals of generation $t$, the probability that this allele will be shared by exactly $m_{t+1}$ individuals in the next generation is

$$P(m_{t+1}, m_t) = \binom{N}{m_{t+1}} \left(\frac{m_t}{N}\right)^{m_{t+1}} \left(1 - \frac{m_t}{N}\right)^{N - m_{t+1}}$$

The probability that a sample with exactly $n$ distinct lineages in generation $t$ has exactly $n$ distinct lineages in generation $(t-1)$ is used to calculate the coalescent. This probability is denoted $G_{n,n}$ (more generally, $G_{i,j}$ has $i$ lineages at $t$ and $j$ at $(t-1)$) and is equal to

$$\begin{aligned} G_{n,n} &= \left(\frac{N}{N}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)...\left(\frac{N-(n-1)}{N}\right) \\ &= 1 - \frac{n(n-1)}{2N} + \mathcal{O}(\frac{1}{N^2}) \\ P_{sc} &\leq (1 - G_{n,n}) = \frac{n(n-1)}{2N} + \mathcal{O}(\frac{1}{N^2}), \end{aligned}$$

where $P_{sc}$ is the probability of a single coalescence from $t$ to $(t-1)$, with the inequality saturated with the assumption of at most a single two-lineage merger per generation (only simple mergers)[3]. Note that for smaller $N$, the likelihood of multiple mergers at a given node and/or generation becomes an important correction.

Starting from $t' = 0$, the probability that the first merging event occurs at or before generation $x = t'N$ for some scaled-time $t'$ is:

$$\begin{aligned} P(N\mathrm{T}_n \leq x) &= 1 - (G_{n,n})^x \\ = 1 - \left(1 - \frac{n(n-1)}{2N}\right)^{t'N} &+ \mathcal{O}(\frac{1}{N^2}) = P(\mathrm{T}_n \leq t') \\ lim_{N\to\infty} P(\mathrm{T}_n \leq t') &= 1 - exp\{-\frac{n(n-1)}{2} t'\}. \end{aligned}$$

Differentiating gives the PDF

$$p(\mathrm{T}_n) = \frac{n(n-1)}{2N} e^{-\frac{n(n-1)}{2N} t'},$$

indeed recovering the Kingman coalescent.

### II.2. The Moran Model

The Moran model for $N$ haploid individuals retains a constant population size $N$ but instead of having discrete, non-overlapping generations has a single birth/death event per timestep. Thus, $N$ generations is analogous

to $N^2$ Moran timesteps. At each timestep, an individual is selected randomly from the population to reproduce, with probability $\frac{1}{N}$ in the neutral model (in general, if the fitness of individual $i$ is $r_i$, then $p_i = \frac{r_i}{\sum_{j=1}^{N} r_j}$ is the probability that $i$ is chosen to reproduce). With replacement, an individual is drawn randomly to die. The dying individual is then deleted from the population and is replaced with an individual of the same genotype (assuming no mutation) as the reproducing individual. Note that it is very likely for some subset of individuals present at timestep $t$ to still be present in the next "generation" at timestep $(t+N)$[4].

The Kingman coalescent can be applied readily to the Moran model because irrespective of population size there can be by definition only simple mergers in any given Moran timestep. At each timestep $t$, with probability $(1-\frac{1}{N})$, the individual who reproduces at $t$ is *not* the same as that which dies at $t$, and in this case (because there is no spatial structure to the population in the Moran model) the event can be viewed for example as one bacterium splitting into two daughter cells; the lineages of the daughter cells have a simple merging event at $t$. However, with probability $\frac{1}{N}$ the same individual who reproduces simultaneously dies and there is no merging event at $t$[1].

Rescale $t'$ to units of $N^2 t$, the time of $N$ generations in the Moran model. The probability that a sample (drawn without replacement at timestep $t$) of $n$ individuals have a simple merge event at timestep $(t-1)$ is

$$\begin{aligned} G_{n,n-1} &= P(\text{merge occurs}) \times P(\text{both daughters in subset } n) \\ &= \left(1-\frac{1}{n}\right)\left(\frac{n(n-1)}{N(N-1)}\right) = (1-G_{n,n}), \end{aligned}$$

as only simple mergers occur (thus the inequality for Wright-Fisher is always saturated for Moran). The probability that the first merge of the $n$ sample occurs before timestep $\tau$ in the past, $\tau = N^2 t'$, is

$$P(\mathrm{T}_n < \tau) = 1-(1-G_{n,n-1})^{\tau} = 1-\left(1-\frac{n(n-1)}{N^2}\right)^{N^2 t'}$$

$$lim_{N\to\infty} P(\mathrm{T}_n < \tau) = 1-\exp\{-n(n-1)t'\},$$

which does not yield the Kingman coalescent. However, rescaling $t'$ again such that $t' = \frac{\tau}{N^2 \backslash 2}$ recovers

$$p(\mathrm{T}_n) = \frac{n(n-1)}{2N} e^{-\frac{n(n-1)}{2N}t'}.$$

Therefore, the *effective population size* (as opposed to the census population size $N$) for the Moran Model is $N_{eff}^{\text{Moran}} = \frac{N}{\sqrt{2}}$.

#### *II.2.1. Effective Population Size*

The Kingman coalescent is an ancestral tree, with the time $\mathrm{T}_n$ until the first merge of a sample of size $n$ distributed as above. These times depend only on the effective population size $N_{eff}$, which, in well-mixed, neutral, non-mutating populations without recombination, $N_{eff} = \frac{N_{census}}{\sigma^2_{off}}$, where $N_{census}$ is the adult census size of the population and $\sigma^2_{off}$ is the variance in the number of offspring produced by a single individual per generation, in a large population[1]. It may be surprising that real populations with spatial structure, mutation, recombination, selection, and other complications can be accurately described by the Kingman coalescent with the proper $N_{eff}$; this suggests that a great variety of evolutionary processes are somehow incorporated into the single parameter $\frac{N_{eff}}{N}$. Census populations can be orders of magnitude larger than effective populations; observed diversity and times to MRCA are typically much less than expected given the size of current populations. For example, though the human population stands around $7.5\cdot 10^9$ as of 2012, the effective population is of order $10^4$ [5].

## III. CLASSICAL MODELS

### III.1. Diffusion

Classical population dynamics models incorporate genetic drift as diffusion in the space of allele frequencies. To motivate this, consider a neutral, non-mutating population of constant size $N$ and let $f_\alpha$ denote the frequency of allele $\alpha$ in the population, with $M$ allele types total. The normalized vector $\boldsymbol{f} \in \mathbb{R}^M$ contains allele frequencies for each type of allele present and has $(M-1)$ degrees of freedom. The assumptions describing the population imply

$$\frac{d\langle f_\alpha\rangle}{dt} = 0 \text{ and } \sigma^2_{f_\alpha} \propto \frac{1}{T_{gen}} f_\alpha \sum_{\beta\neq\alpha}^{M} f_\beta,$$

where, $\sigma^2$ is the variance over a generation of the population, $T_{gen}$ is the generation time of the population, and the sum over $\beta$ is the sum over all other allele types. This yields a Fokker-Planck equation for $P(\boldsymbol{f},t)$, the probability density function for $\boldsymbol{f}$ at time $t$, of the form[6]

$$\frac{\partial P(\boldsymbol{f},t)}{\partial t} = \frac{\tilde{D}_g}{2} \sum_{\alpha}^{M-1}\sum_{\beta}^{M-1} \frac{\partial^2}{\partial f_\alpha \partial f_\beta}\left[(\delta_{\alpha,\beta} f_\alpha - f_\alpha f_\beta) P(\boldsymbol{f},t)\right]$$

where $\tilde{D}_g$ is the genetic diffusion constant, proportional to $(T_{gen}N)^{-1}$ [7][6][5]. Selective advantages of certain alleles can be incorporated as drift terms.

A *well-mixed* population is one in which individuals sample all locations of the environment within their lifetime; it is therefore completely isotropic in terms of allele frequencies and is effectively zero-dimensional. An example of such a system is a vigorously-shaken test tube containing a bacterial colony; few realistic applications of this model exist but it is a starting point for theoretical modeling. The Wright-Fisher and Moran models

discussed above assume a well-mixed population. To create models that incorporate spatial structure of the environment in which the population exists, it may not be irrelevant to ask: is it abuse to refuse to diffuse the use of diffusion to spatial inclusion?

Indeed, diffusion in space is the limiting behavior reached by some early models that incorporate a finite-dimensional environment.

### III.2. Kimura Stepping Stone Model

In any population that is not well-mixed (this living in a finite-dimensional environment), reproduction events affect the allele frequencies only locally. Individual migration occurs nontrivially only over finite distances that are usually much less than the spatial extent of the population. Thus, allele frequencies will not in general be isotropic across the environment, and different localized subdivisions of the population, known as *demes*, will evolve differently. The existence of deme subdivisions is observed in many real populations.

*Island Models* take the extreme limit of demes and discretize space into a finite number of 'island' patches which interact with neighboring islands. Islands are arranged on the vertices of a graph in dimension of the spatial environment, each contains a constant population of effective population size $N_{eff}$, and every generation a small proportion randomly drawn from the members of each island migrate to neighboring islands along the edges of the graph. The number of individuals in each island is kept constant and apart from migration each island evolves independently according to a well-mixed model, so if island $i$ has a set $\{n_i\}$ of neighboring islands and the number of individuals moving *from* island $i$ *to* island $j$ is $m_{ij}$, then $\sum_{j\in\{n_i\}} m_{ji} = \sum_{j\in\{n_i\}} m_{ji}$. In the Kimura stepping stone model, the population $N_{eff}$ of each island evolves with the Wright-Fisher model with migration included after each new generation is chosen. For an example with two alleles, let $f_i$ denote the frequency of allele 1 on island $i$, and again let $\{n_i\}$ denote the set of nearest-neighbor islands to $i$. The genetic diffusion equation gives Brownian motion in allele-frequency space and can be solved with the addition of a white noise term $\eta(t,i)$ [6]. The addition of migration yields:

$$\frac{df_i(t)}{dt} = \sqrt{\tilde{D}_g f_i(t)(1-f_i(t))}\eta(t,i) + \sum_{j\in\{n_i\}} \frac{m_{ji}}{N}(f_j(t) - f_i(t))$$

$$\text{where } \langle \eta(t,i)\rangle = 0 \forall i,t \quad \text{and } \langle \eta(t_1,i)\eta(t_2,j)\rangle = \delta(t_1-t_2)\delta_{i,j}$$

where again time is scaled to the unit of $N$ generations, which corresponds to $N$ Wright-Fisher timesteps. If the islands are arranged on a regular two-dimensional lattice, then for any island $i$ two indices are required, let $i \to (x_a, y_b)$ for integers $a, b$ labeling points on the lattice. The set $\{n_i\}$ contains $(x_{a+1}, y_b), (x_{b-1}, y_b), (x_a, y_{b+1}), (x_a, y_{b-1})$ (neglecting boundary conditions). Assume $m_{ij} = N\tilde{m} \quad \forall i, j \in n_i$.

$$\sum_{j\in\{n_i\}} \frac{m_{ji}}{N}(f_j(t) - f_i(t)) \to$$

$$\tilde{m}\left(f_{a+1,b} + f_{a-1,b} + f_{a,b+1} + f_{a,b-1} - 4f_{a,b}\right)$$

Denote the $x$-axis lattice spacing $l_x$ and the $y$-axis lattice spacing $l_y$. In the limit that these spacings vanish, this results in

$$\frac{df_i(t)}{dt} = \sqrt{\tilde{D}_g f_i(t)(1-f_i(t))}\eta(t,i) + \tilde{m}\left(l_x^2\frac{\partial^2 f}{\partial x^2} + l_y^2\frac{\partial^2 f}{\partial y^2}\right),$$

giving a diffusion equation in space, here for two dimensions. Depending on the dimensionality of the space, interesting stationary solutions may or may not exist; [6] focuses on one dimension while [5] holds that no solutions exist in two dimensions. The full model proposed in [6] is one-dimensional spatial diffusion, together with classical genetic drift, mutation, and weak selection in the case for two alleles but no selection for a Potts-like model of $q$ alleles. This model is meant to explain experimental results of the growth of a population of microbes, with individuals differing only by the allele coding for either red or green fluorescent protein. The initial population is deposited on a Petri dish with a razor blade and subsequently expands on a one-dimensional front; the more-or-less constant length of the front is meant to enforce constant (boundary) population size.

### III.3. Wright-Malécot Model

Sewall Wright and Gustave Malécot independently developed two similar models for populations evolving in a spatial continuum. To begin, individuals of the population are distributed on a two-plane (or, in some of the preliminary papers, on the real line) according to a constant-intensity Poisson point process. Evolution proceeds in discrete generations as in the Wright-Fisher model but with the addition of the assignment of a spatial position, drawn independently from a Gaussian distribution centered on the location of the parent, for each offspring. There is mutation at rate $\mu$. The outcome of both models is termed *Isolation by Distance*, though the meaning in each is slightly different. Both Malécot and Wright's models were developed before Kingman's coalescent and predict the probability of identity in state (defined above) instead.

In Wright's model[8], mating is sexual and biparental. The effective population in the area from which parents may be drawn is the *neighborhood size*, and the variance of the distribution of grandparents is twice that of parents, etc. Spatial restrictions on available mates leads to inbreeding and spatial isolation of distance subpopulations, such that correlations in allele frequencies between two neighborhoods decays with their separation. Thus, long-range correlations between individuals should become negligible. In [9], Ishida stresses that Wright's model of isolation by distance is statistical, stemming from the idea that limited spatial neighborhoods exist in populations; the 'isolation' is spatial. That the size of these neighborhoods determine the scale of allele correlations is a consequence.

Malécot's isolation by distance is a genetic isolation; it is probabilistic, based on the assumption that genetic patterns exist in populations (namely, probability of identity in allelic state decreases with separation) because it is less likely that two distant individuals share a recent common ancestor[9]. Both models yield similar results. Malécot finds the probability of identity in state $F(\vec{x})$ of two individuals separated by $\vec{x}$ to be:

$$F(\vec{x}) \approx \frac{1}{\mathcal{N} + \ln(l \backslash \kappa)} K_0 \left( \frac{|\vec{x}|}{l} \right), \ |\vec{x}| \gg \kappa \qquad (1)$$

with $K_0$ the modified Bessel function of the second kind, order zero; $\mathcal{N}$ is the population of a 'neighborhood,' approximately the number of potential parents of each offspring; $\kappa$ a local length scale; $l$ a characteristic length scale of the population related to neighborhood size, $l = \frac{\sigma}{\sqrt{2\mu}}$, with mutation rate $\mu$ and $\sigma$ the standard deviation of the neighborhood size as chosen for the spatial Gaussian distribution for offspring[5].

The assumptions in the Wright-Malécot model were found to be mathematically inconsistent by Felsenstein[10], a criticism to which Malécot conceded. In the model, the population is, at all times, distributed according to a Poisson point process with uniform intensity, and individuals reproduce independently with number of offspring drawn from a Poisson distribution of mean 1, and the offspring are assigned spatial points according to a normal distribution. Reproduction is thus a branching process/random walk in space. Felsenstein showed that a population evolving this way could *not* have a stationary distribution, violating the first assumption of the model. He then attempted to restrict the environment to a two-torus rather than $\mathbb{R}^2$ but found that "large clumps separated by empty spaces" still formed; he deemed the model "biologically irrelevant"[10].

The island model and Wright-Malécot model give surprisingly similar results for identity in state. Plots in [5] demonstrating this are included in Fig.1a).

## IV. SPATIAL Λ-FLEMING-VIOT MODEL

The classical gene flow models described above yield results that are not consistent with biological data on the following counts:

- they predict much more genetic diversity than is observed (this is analogous to the fact that the effective population size used in the Kingman coalescent are often dramatically smaller than the census population size [13]).
- the functional form of probability of identity over long separations (particularly, the persistence of such long-range correlations) is not explained by these models
- the evolution of non-linked genetic loci is treated by these models as independent, while studies have found that spatial patterns of alleles at multiple loci can be correlated [14].

The paper by Barton et al. in 2010 [5] identifies these discrepancies and proposes a model to address them by incorporating large-scale extinction and repopulation events. The space of the model is a two-torus over which an initial population is distributed according to a Poisson distribution with a spatially-constant density $\rho$. This population is then subject to *events* occurring at a time $t$ and centered at a point $z$ in the two-dimensional space, where for each event, $t$ and $\vec{z}$ are determined by Poisson distribution with rate $\Lambda(r, u, \alpha)$, which is constant in space and time (with units $\frac{1}{m^2 s}$). The event is characterized by a *size* $r$, an *impact* $u$, and a third parameter $\alpha$, such that the rates at which events of different sizes, impacts, and $\alpha$s may be distinct. The extent to which an event effects an individual at point $\vec{x}$ increases with $u$ and decreases with the individual's distance $|\vec{x} - \vec{z}|$ from the event center.

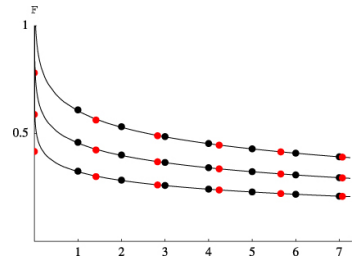
An event kills an individual at $\vec{x}$ with a probability $p_d(|\vec{x} - \vec{z}|, r, u)$, after which the species repopulates. The population is maintained as a uniform Poisson distribution by placing offspring according to a Poisson point process with intensity that varies spatially as $\rho \times p_d(|\vec{x} - \vec{z}|, r, u)$ at a given point $\vec{x}$. The parent of each offspring is selected from the population present before the event (i.e. including those individuals who perished during the event), with a probability of parenthood $p_p(|\vec{x} - \vec{z}|, r, u, \alpha)$. Here, $\alpha$ is used to differentiate the characteristic size of the event, $r$, from that of the pool of parents, $\alpha r$; that is, for $\alpha > 1$, parents are drawn from a larger circle than the event size, while conversely $\alpha < 1$ has parents confined to a smaller region. Barton et. al. make the simplifying assumption that all offspring after a given event derive from the same parent.

This defines a path forward in time for the lineages of a population. The analogous backwards-in-time description, describing the rate at which the lineages of individuals at $\vec{x}$ and $\vec{y}$ coalesce to a common ancestor, follows: the rate at which an individual at $\vec{x}$ is born is the rate at which events occur multiplied by the probability that the individual was born in a given event; because the density of distribution for offspring is $\rho p_d(|\vec{x} - \vec{z}|, r, u)$, this is

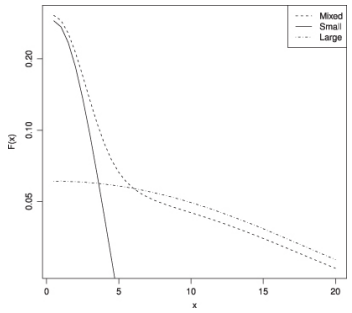
$$\int_0^1 du \int_0^\infty d\alpha \int_0^\infty dr \Lambda(r, u, \alpha) \int d^2\vec{z} p_d(|\vec{x} - \vec{z}|, r, u). \qquad (2)$$

An individual born in a given event derives from a parent whose location is distributed according to $p_l(\vec{x}_{parent}) = \int d^2\vec{z} p_d(|\vec{x} - \vec{z}|, r, u) p_p(|\vec{x}_{parent} - \vec{z}|, r, u)$ (where the integrations over $r$, $u$, and $\alpha$ have been suppressed but are implied). In the single-parent model, if more than one individual is born in a given event, they must share a common parent; thus, the rate at which two lineages at $\vec{x}$ and $\vec{y}$ coalesce to a common parent is

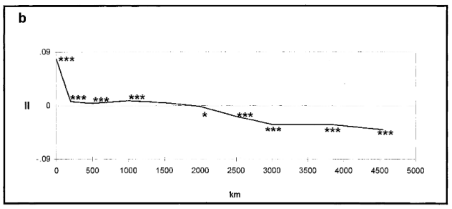
$$\int_0^1 du \int_0^\infty d\alpha \int_0^\infty dr \Lambda(r, u, \alpha) \int d^2\vec{z} p_d(|\vec{x} - \vec{z}|, r, u) p_d(|\vec{y} - \vec{z}|, r, u). \qquad (3)$$



(a) dots: stepping stone model on $\mathbb{Z}^2$ with various $N_{eff}$; lines: Wright-Malécot model with varying parameters, from [5]




(b) Spatial-Λ model [11]. The line has only small events, the dash has only large, and the dots have both types of events. Parameters given in paper.




(c) Data on human identity in state, from [12]

FIG. 1: Identity in State as a function of Separation, for the Island Model, Wright-Malécot model, Spatial-Λ-Fleming-Viot, and comparison to some data.

This information allows the tree of lineages to be propagated backwards in time.

### IV.1. Identity in State

Barton et. al. compute the identity in state for a high-density population in which the allele under study is in either state 0 or 1, distributed with the allele frequency $f(\vec{x}, t)$. Mutations occur at a rate $\mu$ but do not occur during events (such that the offspring repopulating after an event all have genotypes present in the population before the event). Let $p_{t'}(t', \vec{x})$ be the probability distribution of the $t'$ for the lineages of two individuals separated by $\vec{x}$ to coalesce. The probability of identity in state is then:

$$P(\vec{x}, \mu) = \int_{t_0}^{0} p_{t'}(t', \vec{x}) e^{-2\mu t'} dt', \tag{4}$$

where $t_0$ is the $t'$ value at which the population began (i.e. how long the population existed before the sample was drawn), and the decay goes as $2\mu$ because each of the two lineages has probability $(1 - e^{-\mu t'})$ of having mutated in the time $t'$ before the two present individuals are sampled.

In [11], the PDF $p_d(|\vec{x} - \vec{z}|, r, u)$ is a Gaussian of norm $u$ and variance $r^2$, such that the total rate at which an individual at $\vec{x}$ is killed is $\Lambda \int u e^{-\frac{(\vec{x}-\vec{z})^2}{2r^2}} dx = \Lambda 2\pi u r^2$. PDF $p_p(|\vec{x} - \vec{z}|)$ is a normalized Gaussian of variance $(\alpha r)^2$. The solution for $P(\vec{x}, \mu)$ is computed as an implicit function and is solved numerically, and then compared to simulations. Simulations begin with 1000 individuals and trace their lineages back in time until a single initial lineage is reached, recording the time until coalescence for each pairwise merger. The distribution of times until coalescence, gathered over 100 realizations of identical simulations, is the quantity to which $P(\vec{x}, \mu)$ is compared.

An interesting simulation on a torus of sidelength 50 allows for large, rare events and smaller, more frequent events; Λ takes two possible values: $\Lambda(r = 10, u = 1, \alpha = .5) = 4 \cdot 10^{-7}$ and $\Lambda(r = 1, u = 1, \alpha = .5) = 4 \cdot 10^{-4}$. This simulation is shown in Fig. 1b). The distinct effects of the small and large-scale events can be seen clearly on the plot; indeed, incorporating large-scale extinction and recolonization events has resulted in allele correlations that persist to greater individual separations than those found by the models of Kimura, Wright, and Malécot (Fig. 1 a)). One sample of data, referenced in [5], is shown as well.

Models in which $p_d$ and $p_p$ are uniform within disks of sizes $r$ and $\alpha r$, respectively, and zero elsewhere, as well as where $p_d$ and $p_p$ decay exponentially with $|\vec{x} - \vec{z}|$, have also been tried by the authors [11]. In [15], a subset of the Barton et. al. group extend their model to allow for multiple parents during the repopulation after a single event and also begin to incorporate recombination.

The described Spatial-Λ-Fleming-Viot model may be a good description of populations such as trees, where assumptions that all inhabitable areas are completely filled by trees and that such inhabitable areas are distributed spatially as a Poisson distribution, are valid. The density of the population prevents any new seeds from successfully germinating unless the death of an older tree (e.g. forest fires, falling trees) creates an available spot. At this point, whichever seeds have landed in the area may germinate; thus, there is only reproduction after events which kill pre-existing members of the population.

### IV.2. Traveling Islands

It may not seem completely credible that repeated large scale extinction events characterize the history of a species; while forest fires and other natural disasters may significantly diminish a population of trees, it is worth arguing that trees are much more densely-populated than many animals. Though far from the haploid, neutral, non-mutating examples considered, the human is often a

modeling goal for many scientists. While plagues and natural disasters doubtless took large tolls on human populations historically, the spatial range over which these events have significant effect would perhaps *not* have contained of order one-eighth of the population (in [5], the 'large events' have radius 10 on a two-torus of sidelength 50). The toll on a population inflicted by an inherently-localized natural disaster would scale with the density of the population, and therefore different populations would be affected to greatly-varying extents.

It is these large-scale events that give promising results to Barton et. al. Though the model seems very insightful and applicable to numerous species, this author is somewhat skeptical about the scope of its suitability. Perhaps an interesting ground for thought is that of a "traveling island" model. Here, discretized islands are mobile and, upon encountering other islands, interact in a specified way. These islands may be thought of as demes. For reference, deme sizes for some familiar species are: mice: 2-100 [16]; polar bear: 200-3000 [17]; the weta insect (similar to cricket): depending on size of rocky outcrop, 1-6 or 15-40 [18]. These numbers are not necessarily equivalent to effective deme sizes.

Two examples, motivated by specific species, are given. In each, $N_{eff}$ is the constant effective population size of each island. The islands evolve independently between interactions, according to the well-mixed Wright-Fisher model or some preferred alternative.

- Migratory: the movement of islands follows deterministic paths, such as curves connecting two static, specified regions. All islands travel on curves between the two regions, but can follow different curves and end and begin at different points in the two regions. Upon meeting, the interaction between two islands would be cooperative and could proceed in the following way: let $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ denote the average allele frequencies at a particular locus of the members of the meeting islands 1 and 2, respectively. For $n$ possibly alleles at the locus, $\boldsymbol{f}_i \in \mathbb{R}^n$. For each allele $a$, if $f_{1,a} > f_{2,a}$, then perform the scaling $f_{1,a} \to f_{1,a} - c\frac{(f_{1,a}-f_{2,a})}{2}$ and $f_{2,a} \to f_{2,a} + c\frac{(f_{1,a}-f_{2,a})}{2}$, where $c$ is a positive proportion less than 1. The average allele frequencies of the two islands are then closer to one another by an increment depending on $c$. The islands then separate. This is meant to represent chance meeting and subsequent interbreeding between groups like migratory birds or insects.
- Competitive: the movement of the islands follow, for example, a random walk in the space of their environment. This presents the problem that two random walks in two dimensions will never meet; however, they will come arbitrary close to one another [5], so we can define the meeting of two islands as when they are within some predetermined radius from one another. Upon meeting, the interaction is antagonistic; an example of such could be that one of the two islands is selected randomly to be the winner (or, in a non-neutral model, selected in a manner contingent upon member fitness). The losing island then has some (significant) fraction of its population destroyed and replaced with an equal-size subset of copies randomly drawn from the winning island. The two islands then part and continue on their walks. This is meant to be analogous to species with dominance-asserting behavior, such as canines, lions, or even epochs of the human race.

Both types of interactions proposed above would tend to decrease genetic diversity in the population as a whole and also lead to longer-range correlations, as islands that have interacted with one another proceed to separate to potentially long distances.

There are currently many proposed models of population dynamics. Some are famous. None meet the objective of delivering accurate predictions; success compared to data is due to fitting, not knowledge of the population under study. Thus, the dynamics of evolution, quantitatively, remain an open question. Many new models seem to focus on a linear environment such as a shoreline, river bank, or the experiment modeled in [6]. Two-dimensional models seem more scarce and many seem to fall to either mathematical or biological shortcomings. Nonetheless, it seems good progress is being made quickly on all fronts as quantitative evolution becomes a hotter topic in the physics community.


[1] J. Wakeley, *Coalescent Theory: An Introduction*, ch. Gene Genealogies and Genetic Data & The Coalescent, pp. 1–4 & 41–68. W. H. Freeman, 1 ed., 2008.

[2] J. F. C. Kingman, "On the genealogy of large populations," *Journ. of Applied Probability*, vol. 19, 1982.

[3] H. Cui and W. Zhang, "Population genetics: Wright fisher model and coalescent process." University Southern California, 2014.

[4] P. Moran, "Random processes in genetics," *Proc. Camb. Phil. Soc.*, vol. 54, pp. 60–71, 1958.

[5] N. H. Barton, A. M. Etheridge, and A. Veber, "Modelling evolution in a spatial continuum," *Journ. of Statistical Mechanics: Theory and Experiment*, vol. 2013, January 2013.

[6] K. S. Korolev, M. Avlund, O. Hallatschek, and D. R. Nelson, "Genetic demixing and evolution in linear stepping stone models," *Reviews of Modern Physics*, vol. 82, April 2010.

[7] G. J. B. et al, "Exact solution of the multi-allelic diffusion model," *Mathematical Biosciences*, vol. 209, January 2007.

[8] S. Wright, "Isolation by distance," *Genetics*, vol. 28,

no. 2, pp. 114–138, 1943.

[9] Y. Ishida, "Sewall wright and gustave malecot on isolation by distance," *Philosophy of Science*, vol. 76, no. 5, pp. 784–796, 2009.

[10] J. Felsenstein, "A pain in the torus: Some difficulties with models of isolation by distance," *American Naturalist*, vol. 109, no. 967, pp. 359–369, 1975.

[11] N. H. Barton, A. M. Etheridge, and J. Kelleher, "A new model for extinction and recolonization in two dimensions: Quantifying phylogeography," *Evolution*, vol. 64, April 2010.

[12] L. S. et. al., "Geographic patterns of mtdna diversity in europe," *Am. J. Hum. Gen.*, vol. 66, pp. 262–278, 2000.

[13] R. Frankham, "Effective population size v.adult population size ratios in wildlife: a review," *Genetics Research*, vol. 66, October 1995.

[14] G. Barbujani and R. Sokal, "Zones of sharp genetic change in europe are also linguistic boundaries," in *Proc. Natl. Acad. Sci.*, pp. 1816–1819, 1990.

[15] A. M. Etheridge and A. Veber, "The spatial lambda-fleming-viot process on a large torus: Genealogies in the presence of recombination," *The Annals of Applied Probability*, vol. 22, no. 6, 2012.

[16] L. M. Silver, *Mouse Genetics*. Oxford University Press, 1995.

[17] "Trends in polar bear subpopulations." World Wildlife Fund, January 2015.

[18] T. D. Schowalter, *Insect Ecology: An Ecosystem Approach*, ch. Population Structure, p. 145. Academic Press, 4 ed., 2016.