Dispersal Routes Reconstruction and the Minimum Cost Arborescence Problem

Wim Hordijk[a,*], Olivier Brönnimann[a]

[a] *University of Lausanne, Department of Ecology and Evolution, 1015 Lausanne, Switzerland*

Abstract

We show that the dispersal routes reconstruction problem can be stated as an instance of a graph theoretical problem known as the minimum cost arborescence problem, for which there exist efficient algorithms. Furthermore, we derive some theoretical results, in a simplified setting, on the possible optimal values that can be obtained for this problem. With this, we place the dispersal routes reconstruction problem on solid theoretical grounds, establishing it as a tractable problem that also lends itself to formal mathematical and computational analysis. Finally, we present an insightful example of how this framework can be applied to real data. We propose that our computational method can be used to define the most parsimonious dispersal (or invasion) scenarios, which can then be tested using complementary methods such as genetic analysis.

Keywords: Dispersal routes, minimum cost arborescence, invasive species

1. Introduction

Consider the problem of reconstructing the dispersal routes of a given species based on historical data of the first occurrences in various locations. This problem is commonplace, for instance, in the reconstruction of dispersal routes by invasive exotic species [20, 10, 1, 11]. Traditionally, the spatial expansion of exotic species is investigated using a sequence of maps representing the distribution of herbarium records at different time periods (e.g. [20]). However, this graphical approach does not allow for a quantitative understanding of the dispersal process. Consequently, quantitative studies of dispersal have mostly been performed by using analytical models of dispersal of individuals in a population according to random-walk and diffusion processes (see e.g. [29, 13]). In other words, these models concentrate on "mechanistic" bottom-up instead of "empirical" top-down approaches.

The first versions of these models were already realized during the first half of the 20^{th} century [15, 28]. More recently, different approaches based on spatial

*Corresponding author. Email: `wim@WorldWideWanderings.net`, phone: +41216924268.

statistics [5, 2] or on the use of genetic data (see [18] for an overview) have been proposed. However, these techniques often rely on overly simplistic assumptions (such as a constant invasion velocity), are unsuitable for small sample sizes, or do not allow for testing multiple invasion scenarios against each other. Here, we propose a versatile method to reconstruct dispersal routes, using actual observed historical data (e.g., herbarium records), that overcomes most of the above shortcomings, has a firm theoretical foundation, and can easily incorporate various additional model assumptions and features.

The dispersal routes reconstruction problem can generally be formulated as follows. Given a number of geographic locations with (X, Y) (or lat&long) coordinates, and for each location a "first occurrence time" t (when the particular species of interest was first observed in that location), reconstruct possible dispersal routes such that the total sum of the distances of the dispersal events is minimized and there is a unique dispersal path from the earliest occupied location to each later occupied location. In other words, for each location where the species has been observed, choose a candidate "seed" location (i.e., a location that was already occupied at least as early as the current one) from where the species could have dispersed to the current location, in such a way that the total length of the dispersal routes is as small as possible and each location has exactly one "seed" location.

As we will show, this problem can be viewed as an instance of a graph theoretical problem known as the minimum cost arborescence (MCA) problem, for which there exist polynomial-time algorithms. In this paper, we prove formally that the dispersal routes reconstruction problem can indeed be converted to and solved efficiently as an instance of the MCA problem. We also show that if the first occurrence times t_i are all unique (i.e., no two first occurrence times are the same), the dispersal routes reconstruction problem reduces to the minimum cost spanning tree problem for which a simple and efficient "greedy" algorithm suffices. Next, we derive some additional theoretical results, in particular analytical expressions (in a spatially restricted setting) for the minimum, average, and maximum possible values of the total length of optimally reconstructed dispersal routes. These theoretical results are then verified with corresponding empirical results that are obtained from performing computer simulations of (random) instances of this restricted problem version. Finally, we present an example of how the theoretical framework and analysis can be applied to real data from an invasive plant species, and can generate insightful results.

Our method clearly is an "empirical" one (i.e., it fits observed historical data without defining specific mechanisms), as opposed to the "mechanistic" methods mentioned earlier. In that sense, it is very similar in spirit to for example phylogenetic tree reconstruction [25, 14]. In phylogenetics, a parsimony assumption is used to reconstruct a phylogenetic tree that is most likely to have generated the observed data. However, it is generally not sufficient, or even desirable, to generate just one tree. In fact, different model assumption can be incorporated (such as varying mutation rates, or a given number of fixed sites) so that different scenarios (i.e., trees) can be reconstructed, which can then be further tested with additional methods (using, e.g., fossil data). Or, uncertainty

in the data is taken into account by adding a stochastic component so that a statistical analysis (such as bootstrap) can be performed on an ensemble of reconstructed trees. We will show here that our proposed dispersal routes reconstruction method allows for all of this as well, thus making it a versatile and theoretically sound method.

This paper is organized as follows. The next section reviews some basic graph theory and describes the minimum cost arborescence problem. Section 3 then formally states the dispersal routes reconstruction problem as an instance of the minimum cost arborescence problem, and also shows that when all first occurrence times are unique, it reduces to the minimum cost spanning tree problem. In section 4, some theoretical expressions are derived for the possible optimal total lengths of the reconstructed dispersal routes in a spatially restricted version of the problem. These results are then compared to empirical results obtained from computer simulations. Next, section 5 present an insightful example with real data. Finally, section 6 summarizes the main conclusions, and discusses the relevance of our work in a more general context.

2. The minimum cost arborescence problem

We start by reviewing some relevant definitions from graph theory (see e.g. [32], or any other standard text book on graph theory).

A *graph* $G = (V, E)$ consists of a set of *vertices* (or nodes) V and a set of *edges* E that connect pairs of vertices. For example, if there is a connection between two vertices $v_i, v_j \in V$, then $(v_i, v_j) \in E$. If these edges are *directed*, i.e., the edge "points" from vertex v_i to vertex v_j in a one-way manner, then G is called a *directed graph* (or *digraph* for short). Note that in an undirected graph, the vertex pairs $(v_i, v_j) \in E$ are *unordered*, i.e., (v_i, v_j) and (v_j, v_i) denote the same (undirected) edge. However, in a directed graph, the vertex pairs are *ordered*, i.e., (v_i, v_j) and (v_j, v_i) denote different (directed) edges. The total number of vertices in G is denoted $|V|$ (the size of the vertex set V) and the total number of edges in G is denoted $|E|$ (the size of the edge set E). In the following, we are only concerned with directed graphs, a simple example of which is shown in figure 1.

A (directed) *path* from vertex v_1 to vertex v_k in G is a sequence of vertices $(v_1, v_2, \ldots, v_k)$ such that $(v_i, v_{i+1}) \in E$ for $1 \leq i < k$. An *arborescence* is a directed graph G in which, for a unique vertex v_r called the *root*, and any other vertex v_k, there is exactly one directed path from v_r to v_k. In other words, each vertex v_j in an arborescence has exactly one "predecessor" v_i such that $(v_i, v_j) \in E$, and there is one unique vertex v_r (the root) which has no predecessors (all edges point "away" from the root). Even though an arbitrary directed graph G can (potentially) contain *cycles* (a path that returns to its starting vertex), an arborescence has, by definition, never any cycles.

Now suppose there is a *cost function* $c(e)$ that assigns a "cost" to each edge $e = (v_i, v_j)$ in a directed graph G. One could imagine this as the cost of "traveling" from vertex v_i to v_j. The *minimum cost arborescence* problem, also known as the *minimum branching* problem, is then stated as follows [17]:

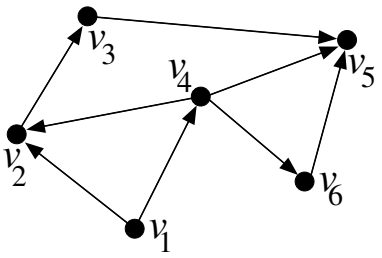

Figure 1: A simple example of a directed graph $G = (V, E)$ with vertex set $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ ($|V| = 6$) and edge set $E = \{(v_1, v_2), (v_1, v_4), (v_2, v_3), (v_3, v_5), (v_4, v_2), (v_4, v_5), (v_4, v_6), (v_6, v_5)\}$ ($|E| = 8$). The sequence $(v_1, v_4, v_2, v_3, v_5)$ is an example of a (directed) path from v_1 to v_5.

> **Minimum cost arborescence (MCA) problem:** Given a directed graph $G = (V, E)$, a unique root vertex $v_r \in V$, and a cost function $c(e), e \in E$, find a subgraph $G_{MCA} = (V, E')$ with $E' \subset E$, that is an arborescence with root v_r and that has a minimum total cost $C = \sum_{e \in E'} c(e)$.

Polynomial-time algorithms for solving this problem were proposed independently first in [8] (calling it "shortest arborescence"), then in [12] (calling it "optimum branching"), and finally in [6] (calling it "minimum spanning tree in a directed network"). These algorithms are based on well-known and efficient *minimum cost spanning tree* algorithms for undirected graphs, with a recursive extension to check for, and resolve, potential cycles [17]. The running time of the basic MCA algorithm is $\mathcal{O}(|V||E|)$. Later, an improvement was made in [31], resulting in a running time of $\mathcal{O}(|E|log|V|)$ for sparse graphs and $\mathcal{O}(|V|^2)$ for dense graphs. So, the MCA problem can clearly be solved efficiently, and is therefore in the problem class **P** [16].

Note that when a given (directed) graph G is acyclic (i.e., contains no cycles), then obviously any subgraph $G' \subset G$ will also be acyclic, and there is no need for checking and resolving potential cycles during the construction of the minimum cost arborescence. In this case, an MCA $G_{MCA} \subset G$ can be found by simply using a "greedy" minimum cost spanning tree algorithm as follows [17]. For each non-root vertex v_j, the (unique) "predecessor" is chosen as vertex v_i such that $e = (v_i, v_j) \in E$ and $c(e)$ is minimal (choosing, e.g., at random in case of a tie). More formally, for each $v_j \neq v_r$, only retain the incoming edge (v_i, v_j) for which $i = \text{argmin}_{k \neq j} \{c(e = (v_k, v_j)) | e = (v_k, v_j) \in E\}$. This algorithm has a running time of $\mathcal{O}(|V|^2)$ in general, or even $\mathcal{O}(|V| + |E|)$ when using an appropriate adjacency list implementation.

3. Dispersal routes reconstruction as an MCA problem

The dispersal routes reconstruction problem, as described in section 1, can now be formalized as an instance of the minimum cost arborescence problem.

Theorem 3.1. *Given an instance of the dispersal routes reconstruction problem with n locations and, for each location $i \in [1, \ldots, n]$, its (X_i, Y_i) coordinates and a first occurrence time t_i, this problem can be solved in polynomial time as an instance of the minimum cost arborescence problem.*

Proof. First we need to show that the dispersal routes reconstruction problem instance can be converted to a minimum cost arborescence problem instance in polynomial time. This will be done by construction.

1. Construct a directed graph $G = (V, E)$ with $|V| = n$ vertices, where each vertex $v_i \in V$ corresponds (one-to-one) to a location i. Label each vertex v_i with the occurrence time t_i of the corresponding location.
2. Include an edge $e = (v_i, v_j)$ (i.e., from vertex v_i to v_j) in E if $t_i \leq t_j$, i.e. if location v_i has an earlier or equal first occurrence time as location v_j.
3. Choose as root vertex v_r the vertex (location) with the smallest t_i (the earliest occurrence time).
4. Finally, define a cost function $c(e)$ as the (euclidean) distance between the two locations corresponding to the vertices $(v_i, v_j) = e$, calculated using their respective coordinates (X_i, Y_i) and (X_j, Y_j).

The time required for step 1 is $\mathcal{O}(|V|)$, $\mathcal{O}(|V|^2)$ for step 2, $\mathcal{O}(|V|)$ for step 3, and $\mathcal{O}(|V|^2)$ for step 4. So, overall this conversion can be done in $\mathcal{O}(|V|^2)$ time.

Next, applying one of the existing MCA algorithms [8, 12, 6, 31] to the directed graph G as constructed above, will result in a subgraph G_{MCA} that has a minimum total cost (i.e., sum of the edge costs or, correspondingly, total length of the dispersal routes) and in which each vertex (location) v_j has exactly one "predecessor" v_i, i.e., the location from which it was (supposedly) first occupied, linking it back to the root vertex v_r via a unique dispersal path. This subgraph G_{MCA} (the minimum cost arborescence of G) thus provides a most parsimonious solution to the original dispersal routes reconstruction problem in polynomial time.

So, both the conversion to an MCA instance and generating the optimal solution can be done in polynomial time, which completes the proof. □

Note that the optimal solution resulting from applying the MCA algorithm is not necessarily unique. There may be several different arborescences (dispersal routes) with the same minimum cost (total dispersal routes length). It is therefore possible that there are multiple most parsimonious dispersal scenarios.

Several further remarks regarding the dispersal routes reconstruction problem as an instance of the MCA problem are in place here. Firstly, cases where there are multiple independent locations of origin or introduction of a species can still be formulated in terms of only one (unique) root vertex v_r by choosing the earliest of the two origin or introduction location as the root vertex, and then simply removing the incoming edge for the other (later) introduction location in the resulting minimum cost arborescence G_{MCA}. Simply removing this incoming edge will (correctly) result in two "disconnected" arborescences, where each occupied location is linked through a unique dispersal path to only

one of the two origin or introduction location. In section 5, we will show an example with real data that contains two known introduction locations.

Secondly, if the (observed) occurrence times t_i for the various locations are all unique, then the problem becomes even simpler, as is stated in the following theorem.

Theorem 3.2. *If, for a given instance of the dispersal routes reconstruction problem, all first occurrence times t_i are unique (i.e., $t_i \neq t_j$ if $i \neq j$), then the problem reduces to the minimum cost spanning tree problem and can be solved using a simple and efficient greedy algorithm.*

Proof. Construct a directed graph G as described in the proof of Theorem 3.1, adding only edges (v_i, v_j) to E if $t_i \leq t_j$. The only way this graph G can contain a cycle such as $(v_1, v_2, v_3, \ldots, v_1)$, is if $t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_1$, which would imply $t_1 = t_2 = t_3 = \ldots = t_1$. Therefore, if all first occurrence times t_i are unique, then G will necessarily be acyclic. And, as was shown at the end of section 2, in case of an acyclic directed graph the minimum cost arborescence problem reduces to the minimum cost spanning tree problem for which a simple and efficient greedy algorithm suffices. □

Thirdly, it is perhaps undesirable to simply apply the MCA algorithm to a given data set and then consider the resulting minimum cost arborescence G_{MCA} as "the" solution. Especially if there is uncertainty in the data (for example, an actual first occurrence obviously happened some time before the first observation was made), this approach may not be adequate. However, this can be dealt with easily in our methodology by adding a stochastic component. In particular, random values (from some appropriate distribution) can be subtracted from the first occurrence times to take the uncertainty into account, before applying the MCA algorithm. This procedure can then be repeated a number of times, creating an *ensemble* of instances (and corresponding minimum cost arborescences) on which a statistical or bootstrap analysis can be performed. This way, a measure of robustness of the solution (or set of solutions) can be obtained. We provide an example of this in section 5. A similar statistical method can be applied to the actual distances between locations as well, as there is always a non-zero probability that an empty site is invaded by an already occupied location that is not the nearest one (although we return to this issue shortly with an alternative approach). Adding or subtracting random values from the distances will cause the algorithm to occasionally choose a non-nearest location as "predecessor". And even if the method is not put in a statistical context this way, it can, for example, be used to generate plausible dispersal scenarios as a null-model, against which other scenarios can be compared and tested.

Finally, many additional features and assumptions can easily be incorporated in the general MCA framework. For example, simply using the euclidean distance between two locations as the cost function $c(e)$ does not take into account that there may be "barriers" to dispersal (such as rivers, roads, or mountains), or that different locations may have different habitat suitabilities. However,

such features can be taken into account by modifying the cost function. A barrier between two locations that cannot be crossed, or only with large difficulty, can be reflected in a very large (or even infinite) cost on the corresponding edge. Or, as another example, actual known dispersal events (e.g., from location v_i to location v_j) can be dealt with by simply leaving out, in the initial graph G, all possible incoming edges to vertex v_j except the one from v_i. This way, the resulting minimum cost arborescence G_{MCA} will always include this known dispersal event (otherwise the algorithm may choose a different, possibly shorter, dispersal event into location v_j). Furthermore, and importantly, the method does not strictly depend on the assumption that a location is always invaded from the nearest previously occupied location, even though it is explained above in that context. An additional term, which is inversely proportional to the difference in first occurrence times t_i between two locations, can be added to the cost function $c(e)$. This can give an "older" location preference over a "younger" location to be chosen as a predecessor (even if it might not be the nearest one), as its added cost term is lower than that of the younger location. This additional time-based term can be weighted (or even completely replace the distance-based term) to give it more, or less, importance as desired by the modeler. As a final example, if genetic data is also available, a cost function based on sequence distance can be used, in which case the algorithm minimizes the total genetic distance. This is perhaps more desirable than using euclidean distances, although much harder to obtain reliable data for.

These are just some examples of how additional features and assumptions can be built into the method. As indicated, this is mainly done by adjusting the cost function $c(e)$ appropriately, or adding a stochastic component to deal with uncertainty in the given data. Once the cost of each edge is calculated, the MCA algorithm then simply finds the arborescence with the minimum overall cost, whether that reflects distance, time, barriers, habitat suitability, or any combination of these or other desired features. Thus, many extensions and modifications can be made to the method, also allowing for a statistical analysis if so desired, making it a versatile tool, not directly depending on any particular (set of) assumption(s).

4. Theoretical results on optimal dispersal routes

In this section, we derive analytical expressions for the expected value and for the possible minimum and maximum values of the total length of optimal dispersal routes in a simplified setting. Consider a spatially restricted, randomized version of the dispersal routes reconstruction problem with n locations that are equally spaced, at distance one from each other, along a one-dimensional (1D) line. Each location i is assigned a first occurrence time $t_i \in [1, \ldots, n]$ at random and without repetition (i.e., the numbers 1 to n are randomly distributed over the n locations). Using the integers from 1 to n as occurrence times t_i is mostly for convenience in the mathematical derivations below, but in principle they can be any values, integer or real. In practice the only requirement is that

they are unique (i.e., mutually distinct), so they can always be ranked and then re-numbered from 1 to n.

Converting this spatially restricted version of the problem to the corresponding minimum cost arborescence problem (as described in the proof of Theorem 3.1), results in a graph $G = (V, E)$ with the $|V| = n$ vertices regularly placed in a linear arrangement at equal distances of one, and with $(v_i, v_j) \in E$ if $t_i < t_j$. The root vertex v_r is chosen as the vertex with occurrence time $t_i = 1$, and the cost function $c(e)$ is simply the (linear) distance between the two vertices $(v_i, v_j) = e$. In the remainder of this section, for notational convenience we simply denote vertex v_i by its assigned label, i.e., the occurrence time $t_i \in [1, \ldots, n]$. For example, vertex k is that vertex which has the label (occurrence time) $t_i = k$ assigned to it.

Since all first occurrence times t_i are unique, we can apply the simple greedy minimum cost spanning tree algorithm to find G_{MCA}. So, for each vertex $j \neq 1$ we choose as "predecessor" the nearest vertex i such that $i < j$. Let $d(j) = c(e = (v_i, v_j))$ be the distance of vertex j to this nearest vertex i with $i < j$. The total cost C of (the optimal) G_{MCA}, which is equivalent to the total length of the optimal dispersal routes, is then:

$$C = \sum_{i=2}^{n} d(i). \tag{1}$$

We can now ask, for example, what the minimum and maximum possible values of C are, or what its expected value is, over all possible instances of this 1D randomized dispersal routes reconstruction problem for various values of n.

The minimum possible value of C is easy to derive. This minimum value is obtained when each vertex i has at least one direct neighbor with a number $j < i$. This is achieved, for example, when the vertices are labeled either in strictly increasing or strictly decreasing order (although there are several other possibilities). The total cost, or length, is then simply

$$C_{min} = \sum_{i=2}^{n} 1 = n - 1. \tag{2}$$

Figure 2 shows an example of such a case.

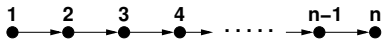

Figure 2: A 1D optimal dispersal routes example with the minimum possible total length $C_{min} = n - 1$.

The maximum possible value of C is somewhat more complicated. The main idea is to assign the numbers 1 to n to the vertices one by one, in such a way that each next one is as far away from the nearest lower number as possible. This can be done in the following way (assume for now that $n - 1$ is a power of 2). First, assign the numbers 1 and 2 to the two outermost vertices, so $d(2) = n-1$. The next number, 3, is then put halfway in between, so $d(3) = (n - 1)/2$. This

divides the linear array of vertices into two halves, and the next two numbers, 4 and 5, are each placed halfway in between either one of these halves, so $d(4) = d(5) = (n-1)/4$. This then divides each half into two halves, and the next four numbers (6 to 9) are each placed halfway in between one of these four parts, so $d(6) = d(7) = d(8) = d(9) = (n-1)/8$, and so on until all numbers are assigned to the vertices, which is achieved after $s = \log_2(n-1)$ such steps. Generalizing this sequence, there is one vertex with distance $d(i) = n-1$, and then 2^{k-1} vertices with distance $d(i) = (n-1)/2^k$ for $k = 1, \ldots, s$. The sum of all these distances is then

$$\sum_{i=2}^{n} d(i) = n - 1 + \sum_{k=1}^{s} 2^{k-1} \frac{n-1}{2^k} = n - 1 + \sum_{k=1}^{s} \frac{n-1}{2} = n - 1 + s\frac{n-1}{2}.$$

So, in general, the maximum possible value of C can be calculated as

$$C_{max} = n - 1 + \left\lfloor \frac{n-1}{2} \log_2(n-1) \right\rfloor. \tag{3}$$

If $(n-1) = 2^k$ for some integer value k, this theoretical value of C_{max} is accurate. In other cases, it may be a small overestimate. For example, for $n = 9$ $(= 2^3+1$; an example of which is shown in Fig. 3), both the theoretical and the actual maximum values for C are 20. However, for $n = 10$, the theoretical value is 23, whereas the actual maximum is only 22, and for $n = 11$, the theoretical value is 26 and the actual value is 25, and so on until they are equal again for $n = 17$ $(= 2^4 + 1)$.

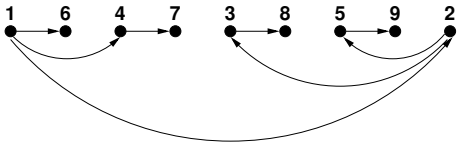

Figure 3: A 1D optimal dispersal routes example with the maximum possible total length $C_{max} = 20$ for $n = 9$.

The minimum and maximum possible values C_{min} and C_{max} are only obtained in a few very specific instances of the 1D randomized dispersal routes problem (such as the ones shown in Figs. 2 and 3). What is perhaps more interesting is the average (or expected) value of C for a random instance. For this, we need to know what the expected distance $E[d(i)]$ is for an arbitrary vertex i, for which, in turn, we need to know the distance probabilities.

For the distance $d(i)$ to be larger than k, vertex i needs to be surrounded by at least $2k$ higher-numbered vertices. Given that there are $n - i$ such vertices, the probability of this is

$$\Pr[d(i) > k] = \frac{\binom{n-i}{2k}}{\binom{n-1}{2k}},$$

i.e., the number of ways to choose $2k$ numbers larger than i out of $n-i$ available ones, divided by the number of ways to choose any $2k$ numbers out of the $n-1$ total ones (we cannot choose the number i itself, hence $n-1$). Note that since there are at most $n-i$ numbers larger than i available, the distance $d(i)$ cannot be larger than $\lfloor \frac{n-i}{2} \rfloor + 1$.

The probability that $d(i)$ is exactly equal to k is now simply

$$\Pr[d(i) = k] = \Pr[d(i) > k-1] - \Pr[d(i) > k],$$

and the expected value of $d(i)$ is thus

$$E[d(i)] = \sum_{k=1}^{\lfloor \frac{n-i}{2} \rfloor + 1} k \cdot \Pr[d(i) = k].$$

Finally, combining all of this together, we get the expected value of C:

$$E[C] = \sum_{i=2}^{n} E[d(i)] = \sum_{i=2}^{n} \sum_{k=1}^{\lfloor \frac{n-i}{2} \rfloor + 1} k \left[\frac{\binom{n-i}{2(k-1)}}{\binom{n-1}{2(k-1)}} - \frac{\binom{n-i}{2k}}{\binom{n-1}{2k}} \right]. \quad (4)$$

Of course this assumes independence of all the vertices, and no influence of boundary effects, which is not entirely true in practice. However, one can expect $E[C]$ to be increasingly accurate for increasing values of n.

To verify these theoretical results, we wrote a small computer simulation that generates random instances of the 1D dispersal routes reconstruction problem, and then applies the simple greedy algorithm (section 2) to obtain the minimum cost arborescence G_{MCA} and the corresponding value of C (or, equivalently, the total length of the optimal dispersal routes). Figure 4 shows a comparison of the theoretical values $E[C]$ from equation 4 and the empirical values C_{avg} obtained from the computer simulations (averaged over 10^5 random instances), for various values of n.

As the figure shows, the theoretical estimate is close, but slightly too low. However, as expected, in a relative sense the theoretical estimates get better for larger n. Figure 5 shows the theoretical values $E[C]$ as a fraction of the empirical values C_{avg}. Clearly, this fraction increases with n, and continues to increase even for large n. The fractions are about 0.866 for $n = 10$, increasing to 0.909 for $n = 100$, and to 0.935 (and still slowly increasing) for $n = 1000$.

Finally, it is interesting to compare the theoretically derived minimum and maximum possible values C_{min} and C_{max} with the minimum and maximum values for C actually obtained in the random samples of 10^5 instances. Figure 6 compares these theoretical values (solid lines) with the empirical ones (dashed lines). The single +s are the empirical averages (as in Fig. 4), as a reference. As the figure shows, for very small values of n the obtained minimum and maximum C values are close to the theoretical values. However, the total number of possible assignments of the numbers 1 to n to the vertices is $n!$, and thus grows exponentially with n, whereas the number of specific assignments that result in

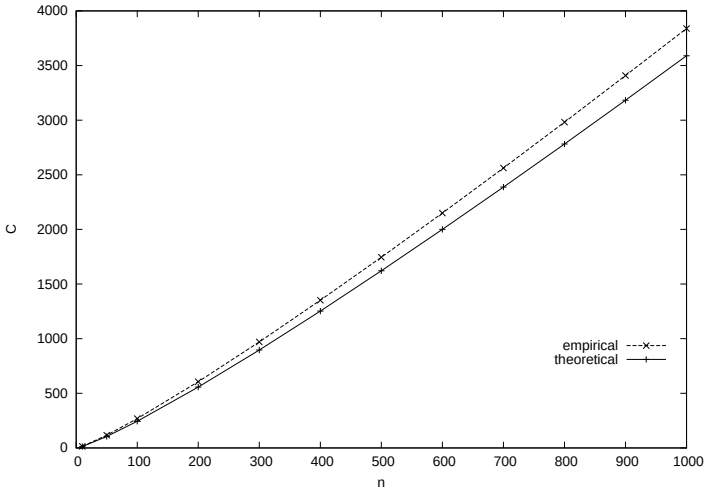

Figure 4: The empirical and theoretical average total length C for 1D optimal dispersal routes for various values of n.

the theoretical minimum or maximum value of C only increases very slowly. So, the probability of achieving the theoretical minimum or maximum decreases quickly with larger n, and the observed minimum and maximum values in a random sample will remain closer to the average.

5. An example: Invasion of *Centaurea stoebe L.*

We now provide an example of how the theoretical framework and analysis as described above can be applied to real data. We consider an invasive plant species that was introduced from Europe to North America at the end of the 19^{th} century: *Centaurea stoebe L.*, a member of the daisy family. There are (at least) two known introduction points for this species, one on the east coast (near Westford, MA, USA), and one on the west coast (near Victoria, BC, Canada). Dispersal data for this species is available in the form of the location and year this species was first observed in various counties throughout the US and Canada.

The given occurrence times t_i are not necessarily unique, as they are only provided as the first year in which the species was observed in each location (no day or month specified). However, given that there is always some uncertainty in the observed occurrence times (as already mentioned earlier), we subtract independent and identically distributed (i.i.d.) random values from the reported times t_i, thereby making them unique, and also taking uncertainty in the observation data into account. The random values are drawn from a negative

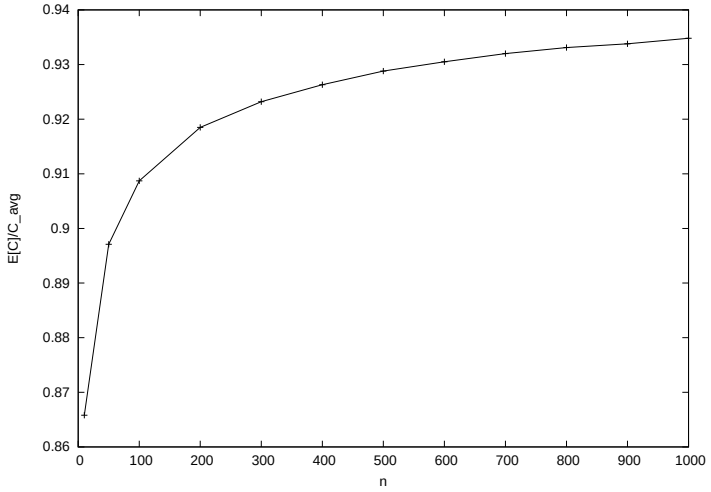

Figure 5: The theoretical expected total cost $E[C]$ relative to the corresponding empirical averages C_{avg} for various values of n.

exponential distribution, which more or less simulates the time it would take for an observer who "walks around randomly" in a given area to discover the plant by chance. We tried different values for the mean μ of this distribution, such as $\mu = 1, 5, 10$ years, although this does not appear to result in any qualitative difference in the reconstructed dispersal routes. Of course if μ is taken too large ($\mu > 20$ in our case), then the first occurrence times t_i become practically completely randomized.

Making the occurrence times t_i unique this way, we then convert the data to an instance of the minimum cost arborescence problem (simply using the euclidean distance between two points as the cost function $c(e)$). From theorem 3.2 it follows that we can now apply the simple greedy (minimum cost spanning tree) algorithm to obtain the most parsimonious (i.e., minimum total length) dispersal routes. Finally, we repeat this procedure 100 times, and calculate bootstrap support values for the reconstructed dispersal routes. In section 3 we already indicated how to deal with multiple introduction points.

Figure 7 shows the result for the *Centaurea stoebe* data (using $\mu = 5$ years for the negative exponential distribution). The gray-scale on the edges indicates the bootstrap support values: lighter colors indicate lower support while darker colors indicate higher support. In a forthcoming paper [7] we will investigate this particular case study and the results of the dispersal routes reconstruction in detail, but one striking conclusion (which, again, does not seem to depend on the chosen value for μ) is that the two separate and independent invasions

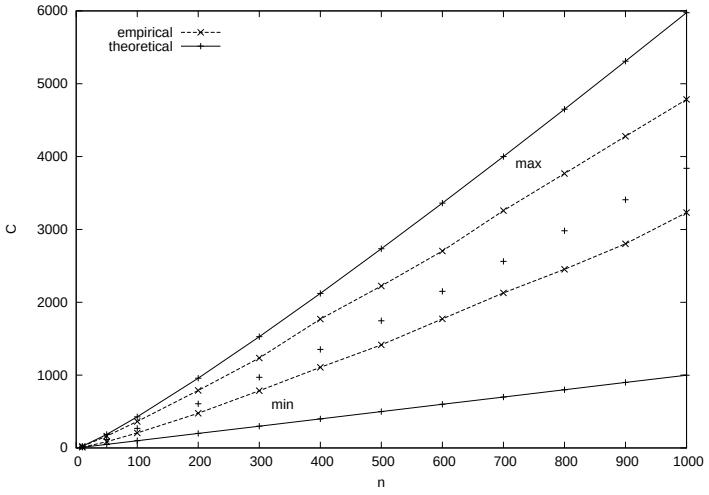

Figure 6: The theoretical (solid lines) and empirically observed (dashes lines) minimum and maximum total length C for 1D optimal dispersal routes for various values of n.

(east and west) show very different spatio-temporal dynamics, in part driven by different climatic conditions.

In section 4 we derived analytical expressions for the minimum, average, and maximum optimal values of the total dispersal routes length in a spatially restricted version of the problem. With real data, such as for *Centaurea stoebe*, it is more difficult to derive similar analytical formulas. However, we can still obtain the empirical values, just as with the computer simulations used in the previous section.

The minimum possible optimal value is obtained when for each location (occupied county) simply the nearest neighboring location is chosen as "predecessor", regardless of the actual occurrence times t_i. The average, or expected, optimal value (and standard deviation) is obtained as follows. Given the locations and occurrence times t_i, randomly re-assign the occurrence times to the available locations. Repeat this a number of times, and for each such instance, apply the minimum cost arborescence algorithm to get the optimal value, and calculate the average and standard deviation over all instances.

Figure 8 presents the results of this. The vertical red line (labeled "**obs**") shows the actual (optimal) value of the total length of the reconstructed dispersal routes. The blue line (labeled "**min**") shows the minimum possible optimal value. The histogram (labeled "**random**") shows the optimal values of 100 instances of randomly re-assigning the t_i values, with the vertical dashed lines representing two standard deviations away from the average.

As the plot shows, the observed value (red line) is closer to the random

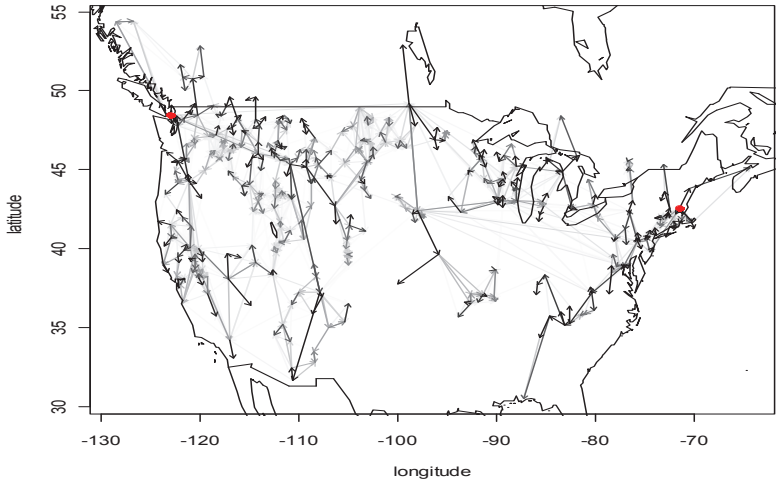

Figure 7: Reconstruction of the dispersal routes for *Centaurea stoebe*, using the minimum cost arborescence method, with $\mu = 5$. The two red dots show the known introduction points. The gray-scale on the edges indicates bootstrap support values (light=low, dark=high).

instances (histogram) than the minimum possible value (blue line). One conclusion that can be drawn from this observation is that there is a fair amount of "long distance" dispersal in the overall dispersal pattern of *Centaurea stoebe*. This can, in part, be explained by the fact that the construction of long distance railroad lines at around the same time as the species was first introduced in North America, has enabled it to disperse over longer distances in one single dispersal event as would have been possible naturally. However, the observed (optimal) value is still well outside of the random sample, indicating that there is indeed a structured pattern in the dispersal dynamics of *Centaurea stoebe*.

To conclude, we have shown how the minimum cost arborescence framework can be applied directly to real data and generate useful and insightful results. Next to efficiently reconstructing invasion scenarios, including bootstrap support values, a (rough) measure of how much "short distance" and "long distance" dispersal has occurred can be obtained by comparing the actual optimal value with the minimum possible one and a sample of instances with randomly reassigned occurrence times. These are just some examples of the possibilities the introduced framework offers.

6. Conclusions and discussion

We have shown formally that the dispersal routes reconstruction problem can be stated as an instance of the minimum cost arborescence problem, for

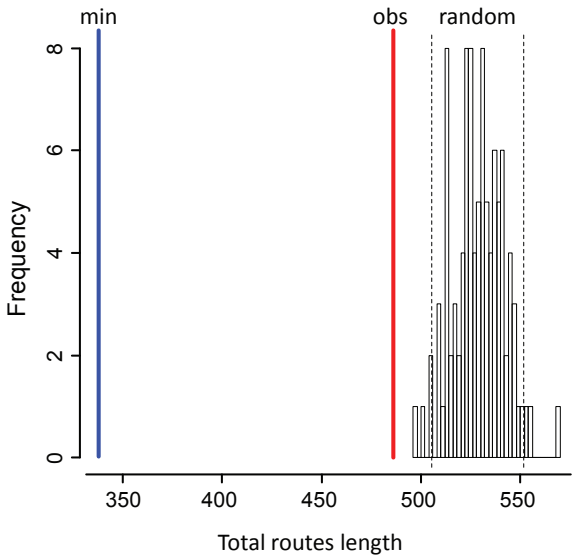

Figure 8: Comparison of the actual optimal dispersal routes length (**obs**), the minimum possible value (**min**), and that of 100 random instances (**random**) for the *Centaurea stoebe* data.

which there exist efficient algorithms. Thus, the dispersal routes problem can be solved in polynomial time (with a small exponent), placing it well within the problem class **P**. Furthermore, we have derived some theoretical results on the possible minimum and maximum, and also the expected values of the total length of optimal dispersal routes in a spatially restricted version of the problem, which were verified by empirical results from computer simulations. With this, we have placed the dispersal routes reconstruction problem on solid theoretical grounds, establishing it as a tractable problem that can also be analyzed mathematically and computationally. In addition, we have provided an insightful example of how this theoretical framework can be applied to real data and generate useful results such as the most parsimonious dispersal routes (plus, if desired, bootstrap support values) and a measure of the amount of "short distance" versus "long distance" dispersal. And, as argued, the framework as introduced here is versatile and can easily be extended and modified, or placed within a statistical framework, depending on needs or circumstances.

Understanding the characteristics and determinants of invasion routes has important practical applications, such as the design and implementation of quarantine strategies, and anticipation of conservation actions such as preventing establishment of new focal populations or eliminating them, rather than focusing on established invasion fronts [24, 30].

So far, several approaches have been used to infer invasion routes. The first quantitative method that was used is based on neighborhood diffusion, and is

often referred to as Fisher's wave of advance model, in which the range of an invading species is predicted to increase linearly with time [15, 28]. A constant velocity is frequently observed in many organisms, especially at local scale [4]. However, recent data at larger scale [3] show that long-distance dispersal could occur simultaneously with neighborhood diffusion within the same species, accelerating the range expansion [26]. This type of stratified diffusion can be successfully modeled using dispersal kernels that incorporate various step lengths of dispersal [23]. This approach has the potential to relate the rate of spread of populations to their behavioral and demographic properties and thus to provide valuable ecological insight into invasion processes. However, the acquisition of proper data to feed the models requires extensive effort. Moreover, the approach relies on the hypothesis that species have a constant dispersal kernel throughout the invasion range, an assumption that might not be valid in many species undergoing range expansion [27]. We believe that our computational approach, requiring only dated observed occurrences and not being based on any dispersal assumptions, has the potential to efficiently document invasion routes with minimal effort. It can, for instance, be used as a null model against which to validate diffusion and long-distance dispersal approaches.

Another approach to infer invasion routes that has emerged recently in the literature is based on the use of genetic data (see [18] for a review). Molecular genetic data can offer unique insights into the sources, routes, and mechanisms of spread [19, 21]. The main insight from molecular genetic studies of invasion routes is that multiple introductions from the native range might occur but remain unnoticed because they cannot be distinguished from spread from existing populations in the invaded range [18]. This bridgehead effect can thus not be taken into account explicitly by kernel diffusion approaches. Molecular genetic studies thus have the power to reveal separate invasion outbreaks resulting from repeated (trans-continental) introductions [22, 9, 21]. However, inferring routes using molecular genetic methods do not replace the analysis of observational and historical records. Having historical data is a necessary requirement for defining a limited set of invasion scenarios that can be tested against each other statistically (e.g., by using an Approximate Bayesian Computational (ABC) approach [21]). We propose that the minimum cost arborescence approach as introduced in this paper can be used to define the most parsimonious invasion scenarios (by, for example, reconstructing several different dispersal routes using different model assumptions), which can subsequently be tested using a population genetic analysis. As such, the two approaches can act in a complementary way as a powerful invasion analysis tool.

Acknowledgments

This work was funded by the 6^{th} European Framework Program grant GOCE-CT-2007-036866 ECOCHANGE and by the National Centre of Competence in Research (NCCR) Plant Survival, a research program of the Swiss National Science Foundation. The work was performed in the ECOSPAT lab of Prof. Antoine Guisan, who we thank for support and helpful discussions.

References

[1] R. G. Ahern, D. A. Landis, A. A. Reznicek, and D. W. Schemske. Spread of exotic plants in the landscape: the role of time, growth habit, and history of invasiveness. *Biological Invasions*, 12(9):3157–3169, 2010.

[2] S. Aikio, R. P. Duncan, and P. E. Hulme. Lag-phases in alien plant invasions: separating the facts from the artefacts. *Oikos*, 119:370–378, 2010.

[3] D. Andow. Spread of invading organisms: patterns of spread. In K. C. Kim, editor, *Evolution of insect pests: the pattern of variations*, pages 219–242. Wiley, 1993.

[4] D. Andow, P. Kareiva, S. A. Levin, and A. Okubo. Spread of invading organisms. *Landscape Ecology*, 4:177–188, 1990.

[5] J. N. Barney, T. H. Whitlow, and A. J. Lembo. Revealing historic invasion patterns and potential invasion sites for two non-native plant species. *PLoS ONE*, 3:e1635, 2008.

[6] F. Bock. An algorithm to construct a minimum spanning tree in a directed network. *Developments in Operations Research*, pages 29–44, 1971.

[7] O. Brönnimann, P. Mráz, W. Hordijk, B. Petitpierre, A. Guisan, and H. Müller-Schärer. Climatic niche dynamics through space and time in the invasive Spotted Knapweed (Centaurea stoebe L.). 2012. In preparation.

[8] Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

[9] M. Ciosi, N. J. Miller, K. S. Kim, R. Giordano, A. Estoup, and T. Guillemaud. Invasion of Europe by the western corn rootworm, Diabrotica virgifera virgifera: multiple transatlantic introductions with various reductions of genetic diversity. *Molecular Ecology*, 17:3614–3627, 2008.

[10] P. H. C. Crawford and B. W. Hoagland. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? *Journal of Biogeography*, 36(4):651–661, 2009.

[11] P. Csontos, M. Vitalos, Z. Barina, and L. Kiss. Early distribution and spread of Ambrosia artemisiifolia in central and eastern Europe. *Botanica Helvatica*, 120(1):75–78, 2010.

[12] J. Edmonds. Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B:233–240, 1967.

[13] R. Engler, W. Hordijk, and A. Guisan. The MigClim R package – seamless integration of dispersal constraints into projections of species distribution models. *Ecography*, 2012. In press.

[14] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004.

[15] R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenetics*, 7:353–369, 1937.

[16] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[17] A. Gibbons. *Algorithmic Graph Theory*. Cambridge University Press, 1985.

[18] L. J. L. Handley, A. Estoup, D. Evans, C. Thomas, E. Lombaert, B. Facon, A. Aebi, and H. Roy. Ecological genetics of invasive alien species. *BioControl*, 56(4):409–428, 2011.

[19] P. M. Hoos, A. Whitman-Miller, G. M. Ruiz, R. C. Vrijenhoek, and J. B. Geller. Genetic and historical evidence disagree on likely sources of the Atlantic amethyst gem clam Gemma gemma (Totten, 1834) in California. *Divers Distrib*, 16:582–592, 2010.

[20] C. Lavoie, Y. Jodoin, and A. G. Merlis. How did common ragweed (Ambrosia artemisiifolia L.) spread in Quebec? A historical analysis using herbarium records. *Journal of Biogeography*, 34(10):1751–1761, 2007.

[21] E. Lombaert, T. Guillemaud, J. M. Cornuet, T. Malausa, B. Facon, and A. Estoup. Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *PLoS One*, 5:e9743, 2010.

[22] N. Miller, A. Estoup, S. Toepfer, D. Bourguet, L. Lapchin, S. Derridj, K. S. Kim, P. Reynaud, L. Furlan, and T. Guillemaud. Multiple transatlantic introductions of the western corn rootworm. *Science*, 310:992, 2005.

[23] D. Mollison. Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society B*, 39:283–326, 1977.

[24] M. E. Moody and R. N. Mack. Controlling the spread of plant invasions: The importance of nascent foci. *Journal of Applied Ecology*, 25:1009–1021, 1988.

[25] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2003.

[26] N. Shigesada, K. Kawasaki, and Y. Takeda. Modeling stratified diffusion in biological invasions. *American Naturalist*, 146:229–251, 1995.

[27] A. D. Simmons and C. D. Thomas. Changes in dispersal during species' range expansions. *American Naturalist*, 164:378–395, 2004.

[28] J. G. Skellam. Random dispersal in theoretical populations. *Biometrika*, 38:196–218, 1951.

[29] M. G. Smolik, S. Dullinger, F. Essl, I. Kleinbauer, M. Leitner, J. Peterseil, L.-M. Stadler, and G. Vogl. Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. *Journal of Biogeography*, 37(3):411–422, 2010.

[30] A. V. Suarez, D. A. Holway, and T. J. Case. Patterns of spread in biological invasions dominated by long-distance jump dispersal: insights from Argentine ants. *Proceedings of the National Academy of Science USA*, 98:1095–1100, 2001.

[31] R. E. Tarjan. Finding optimum branchings. *Networks*, 7:25–35, 1977.

[32] W. T. Tutte. *Graph Theory*. Addison-Wesley, 1984.