Received: 16 November 2021 | Accepted: 16 October 2022

DOI: 10.1111/1365-2664.14356

Journal of Applied Ecology BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Boosting propagule transport models with individual-specific data from mobile apps

Samuel M. Fischer[1,2] | Pouria Ramazi[3] | Sean Simmons[4] | Mark S. Poesch[5] | Mark A. Lewis[1,6]

[1]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada; [2]Department of Ecological Modelling, Helmholtz-Centre for Environmental Research—UFZ, Leipzig, Germany; [3]Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada; [4]Angler's Atlas, Goldstream Publishing, Prince George, British Columbia, Canada; [5]Department of Renewable Resources, University of Alberta, Edmonton, Alberta, Canada and [6]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

Correspondence
Samuel M. Fischer
Email: samuel.fischer@ualberta.ca

Funding information
Alberta Environment and Parks; Canada Research Chairs; Natural Sciences and Engineering Research Council of Canada

Handling Editor: Joseph Bennett

Abstract

1. Management of invasive species and pathogens requires information about the traffic of potential vectors. Such information is often taken from vector traffic models fitted to survey data. Here, user-specific data collected via mobile apps offer new opportunities to obtain more accurate estimates and to analyse how vectors' individual preferences affect propagule flows. However, data voluntarily reported via apps may lack some trip records, adding a significant layer of uncertainty. We show how the benefits of app-based data can be exploited despite this drawback.
2. Based on data collected via an angler app, we built a stochastic model for angler traffic in the Canadian province Alberta. There, anglers facilitate the spread of whirling disease, a parasite-induced fish disease. The model is temporally and spatially explicit and accounts for individual preferences and repeating behaviour of anglers, helping to address the problem of missing trip records.
3. We obtained estimates of angler traffic between all subbasins in Alberta. The model's accuracy exceeds that of direct empirical estimates even when fewer data were used to fit the model. The results indicate that anglers' local preferences and their tendency to revisit previous destinations reduce the number of long inter-waterbody trips potentially dispersing whirling disease. According to our model, anglers revisit their previous destination in 64% of their trips, making these trips irrelevant for the spread of whirling disease. Furthermore, 54% of fishing trips end in individual-specific spatially contained areas with mean radius of 54.7 km. Finally, although the fraction of trips that anglers report was unknown, we were able to estimate the total yearly number of fishing trips in Alberta, matching an independent empirical estimate.

4. *Policy implications*. We make two major contributions: (1) we provide a model that uses mobile app data to boost the mechanistic accuracy of classic propagule transport models, and (2) we demonstrate the importance of individual-specific behaviour of vectors for propagule transport. Ignoring vectors' local preferences and their tendency to revisit previous destinations can lead to significant overestimates of vector traffic and biased estimates of propagule flows. This has clear implications for the management of invasive species and animal diseases.

KEYWORDS
angler, gravity model, Invasives (applied ecology), modelling (disease ecology), smartphone apps, survey method, vector, whirling disease

1 | INTRODUCTION

Recreational overland traffic is a major vector for several invasive species and pathogens (Hulme, 2009; Karesh et al., 2005). Examples include invasive plants and pathogens carried via the soil attached to gear and vehicles of tourists (Cushman & Meentemeyer, 2008; Von der Lippe & Kowarik, 2007), invasive insects introduced along with campers' firewood (Koch et al., 2012), or invasive mussels, nonindigenous bait fish, and water-borne diseases spread by recreational boaters (Johnson et al., 2001) and anglers (Gates et al., 2007; Kilian et al., 2012; Litvak & Mandrak, 1993; Nalepa & Schloesser, 2013). Given the difficulties and costs associated with eradicating invasive species and pathogens once they have established at a site, it is key that any management strategy prevents propagule transport and detects new infestations early (Leung & Mandrak, 2007; Pluess et al., 2012). This requires a detailed understanding of transport pathways and vector's movement patterns.

Data collected via smartphone apps have become a valuable resource to study human mobility (Wang et al., 2019) and offer new opportunities to understand and predict the dispersal of invasive species and pathogens (Papenfuss et al., 2015; Venturelli et al., 2017). Assuming a sufficiently large user base, mobile app data can be collected at relatively low cost over large spatial and temporal scales (Papenfuss et al., 2015; Venturelli et al., 2017). However, even if many trip records are available, the datasets collected via apps are typically far from complete: often, only a small fraction of the population of interest, for example, hikers or anglers, use any particular app, and app users do not record all their trips (Papenfuss et al., 2015). Even if an app records thousands of trips, this number remains small in comparison to the vast number of origin–destination pairs for which traffic estimates may be desired. For example, if we seek to estimate the vector traffic between 100 origins and 100 destinations, the number of origin–destination *pairs* is 10,000. Hence, direct empirical estimates of traffic flows can be prone to significant statistical error.

A common approach to bridge such data gaps is to use models, such as gravity models (Bossenbroek et al., 2001; Ferrari et al., 2006; Li et al., 2011; Muirhead & MacIsaac, 2011; Potapov et al., 2010). By combining empirical observations with additional covariates, for example, geographical and socioeconomic data, models can provide detailed estimates of vector traffic on broad scales and may even allow insights into the mechanisms behind traffic patterns. In the past, vector traffic models have been fitted to data collected via mail-out surveys (Chivers & Leung, 2012; Drake & Mandrak, 2014; Muirhead & MacIsaac, 2011; Potapov et al., 2010), roadside traffic surveys (Fischer et al., 2020), on-site surveys at origins and destinations (Bossenbroek et al., 2007; Leung et al., 2004) or registration records from origins and destinations (Bossenbroek et al., 2001; Prasad et al., 2010). However, since gathering data via these methods is often costly and the data may represent specific locations and time frames only, app data are a promising alternative resource for fitting vector traffic models. In this study, we show how this can be done.

A drawback of app data is that app users may report their trips sparsely, making the temporal sequence of their trips incomplete. The data may still yield insight into how often each destination is visited, but without knowing the full trip sequence, it is difficult to gauge how far and quickly vectors will spread propagules after being infested. A vector frequently revisiting their previous destination has a much lower risk of spreading a disease than a vector who prefers to alternate between destinations. As the risk of a successful transmission is highest between consecutively visited sites, disregarding the trips not recorded by app users could bias the predictions of propagule dispersion models significantly.

To estimate the number of relevant trips between sites despite missing data, some studies assume that the destinations of consecutive trips are chosen independently from one another (Bossenbroek et al., 2001; Leung et al., 2004). Then, incomplete trip sequences would be representative for all trips. In practice, however, individual travellers may have local preferences and tend to revisit previous destinations, which would lower the risk of propagule transport. Accounting for these individual preferences requires a more intricate modelling approach. We tackle this problem and build a vector movement model that can be fitted to incomplete app-based data.

Although our approach is applicable to studying the spread of various pests in both terrestrial and aquatic systems, we introduce and demonstrate it by considering a particular case study: we model angler movement in the Canadian province Alberta based on data collected via the 'MyCatch' angler app so as to create a risk map for the spread of whirling disease in Alberta. Our focal invader, whirling disease, is a fish disease caused by the aquatic parasite *M. cerebralis* (Hofer, 1903), which can increase the death rate of juvenile salmonid fish up to 90% (Elwell et al., 2010) and may thus entail severe ecological and economic consequences (Ramazi, Fischer, et al., 2021; Turner et al., 2014). As there is currently no known cure for whirling disease in natural ecosystems (Turner et al., 2014), management is limited to reducing the risk of parasite introduction.

Our main objective is to estimate how often each subbasin in Alberta is visited by anglers who have visited an infested area on their previous trip. While estimating angler traffic, we also assess how traffic depends on local preferences of individual anglers and their tendency to revisit previous destinations. Our results indicate that if these factors are not accounted for, local traffic is significantly underestimated while long-distance traffic is overestimated. This, in turn, has general implications for risk assessment and management of invasive species and infectious diseases.

2 | MATERIALS AND METHODS

An overview of the data and submodels used in our approach is displayed in Figure 1. Below we describe the components and their interplay in detail. An overview of the mathematical symbols used in this paper is provided in Appendix S1.

2.1 | Data

We used a dataset collected via the MyCatch angler app, which can be downloaded free of charge for Android and iOS devices and allows anglers to share information regarding waterbodies they visit, for example, their catch success. App users need to provide their home postal codes and may record their fishing destinations either via GPS or select their destination waterbodies on a map. In addition to using the app, registered users can also enter information via a web interface. Although not all anglers in Alberta use the app, the app users have been found to be mostly representative of the province's anglers, with a slight bias towards higher app usage in urban areas (Johnston et al., 2021).

The data were collected from May 2018 to April 2020 inclusive. We determined the home *locality* (city, town, village, etc.) of each app user who recorded at least one trip within this time frame and collected the sequence of their fishing destinations along with the trip dates. If an angler recorded several trips to the same waterbody on a day, we merged these into a single trip. To keep the number of fishing destinations tractable, we aggregated them over *subbasins* (hydrologic units of level 8) and neglected more detailed information. Subbasins are a natural unit for modelling the spread of aquatic diseases, because they have a unique outflow each. Alberta consists of 422 subbasins with a mean area of 1517 km^2. Our dataset included 575 anglers, who made 2104 trips. For 229 of these trips, we could not determine the destination subbasin, because the anglers did not provide destination coordinates and the reported destination waterbodies spanned multiple subbasins. We disregarded these

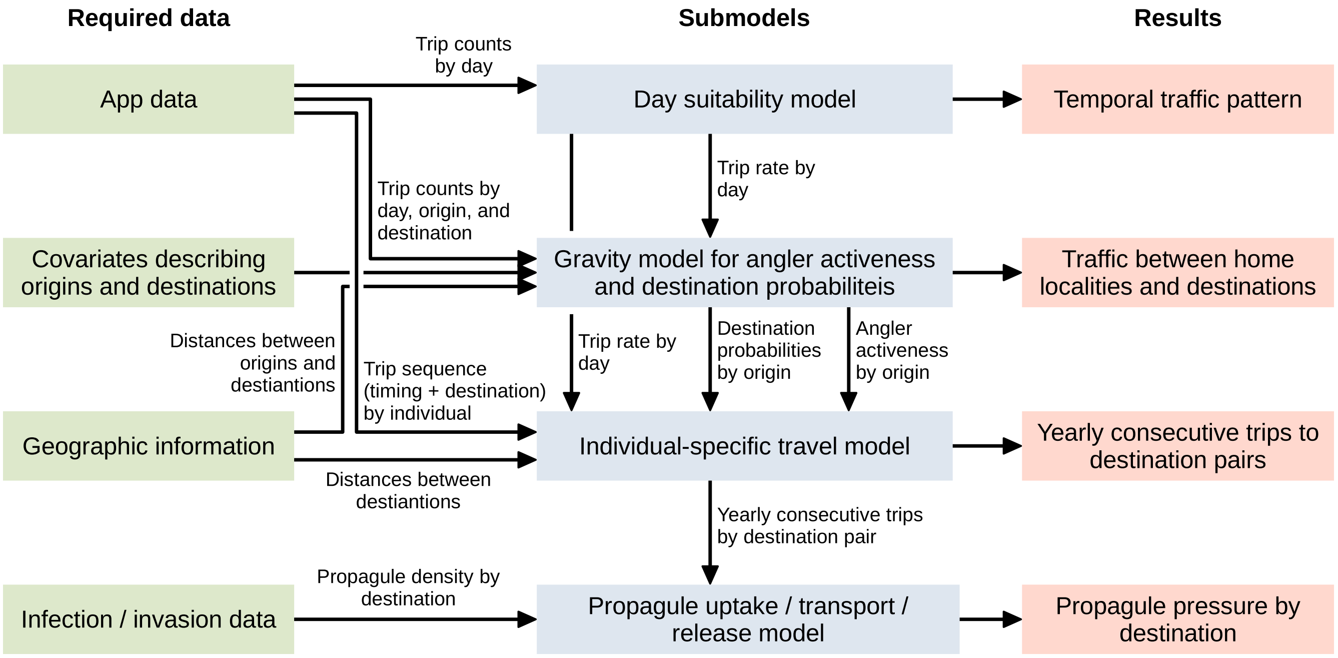
Required data
Submodels
Results
App data
Trip counts by day
Day suitability model
Temporal traffic pattern
Trip rate by day
Trip counts by day, origin, and destination
Covariates describing origins and destinations
Gravity model for angler activeness and destination probabiliteis
Traffic between home localities and destinations
Distances between origins and destiantions
Trip sequence (timing + destination) by individual
Trip rate by day
Destination probabilities by origin
Angler activeness by origin
Geographic information
Individual-specific travel model
Yearly consecutive trips to destination pairs
Distances between destiantions
Yearly consecutive trips by destination pair
Propagule density by destination
Infection / invasion data
Propagule uptake / transport / release model
Propagule pressure by destination

FIGURE 1 Model components. The data required for the analysis are displayed in green, the different submodels in blue and the results in red. The submodels are combined into a stochastic traffic model described in Section 2.2. Although incorporating a sophisticated propagule transport model is possible, we use the number of directly consecutive trips from infested to uninfested areas as a proxy for propagule pressure in this study.

trips. All research was conducted in accordance with the Human Research Ethics Policy of the University of Alberta (approval number Pro00102610).

As predictors for anglers' behaviour, we used data on the localities and the subbasins (Table 1). Besides geographical and socioeconomic data, we compiled data collected on the Angler's Atlas website (https://www.anglersatlas.com). The website contains a page for each major waterbody in Alberta, providing anglers with waterbody-specific information and allowing them to report the species of fish they have caught there. Fish species reports can be upvoted and downvoted by other anglers to confirm or rebut an observation. We computed for each subbasin the area and perimeter of all waterbodies with at least one confirmed fish species. Furthermore, we computed the cumulative number of waterbody webpage visits and species upvotes per subbasin. For waterbodies spanning over multiple subbasins, we distributed the values over all applicable units according to their share of the waterbodies' perimeters. In addition to the listed covariates, we determined the number of registered anglers for each locality. The sources of the individual datasets we used are listed in the Data Sources section.

2.2 | Angler traffic as a stochastic process

We modelled anglers' decision-making as a stochastic process, which determines both the recorded number of fishing trips between each home locality and subbasin (*origin* and *destination*) and the *expected* yearly number $\mu_{j_1j_2}$ of trips anglers make to destination j_1 directly after visiting destination j_2. While we sought to estimate the latter number for all destination pairs, we fitted the model based on the former (Figure 2).

We assumed that anglers start all their fishing trips at their home localities and visit a single destination per trip. They make trips randomly at rates dependent on their origins, the date and random factors not explicitly covered in the model, for example, weather conditions. We modelled the trip rate for an angler from origin i on day t as $\mu_i \epsilon_t$, where the *angler activeness* μ_i is the mean number of trips per day for an angler from origin i, and the *day suitability* ϵ_t is a gamma random variable with mean τ_t, denoting how well day t of the study period is suited for going fishing:

$$\epsilon_t \sim \text{Gamma}\left(\frac{\tau_t}{\alpha}, \alpha\right). \tag{1}$$

The gamma distribution can take on a variety of shapes and is thus suited for diverse modelling applications (Husak et al., 2007; Kleiber & Kotz, 2003). The dispersion parameter α determines the variance of the day suitability; the expected day suitability τ_t is normalized so that its temporal average is 1, that is, $\frac{1}{T}\sum_t \tau_t = 1$ with T being the number of days in the study period. We supposed that the day suitability ϵ_t is the same for all anglers in Alberta, whereas their individual decisions are independent from one another. The values μ_i and τ_t are given by the submodels in Section 2.4.

We assumed that anglers choose the destinations of their trips based on individual local preferences and their previous fishing destinations (Figures 3 and 4). Consider an angler from origin i. With probability ξ_{same}, they decide to revisit the destination of their last trip. Otherwise, they choose a new destination as follows: with probability ξ_{region}, they constrain their destination choice to their *region of preference* $\mathscr{R}$—a spatially contained set of destinations that they personally like best—and choose a destination $j \in \mathscr{R}$ according to probabilities $p_{ij|\mathscr{R}}$. Alternatively, with probability $\xi_{\text{all}} = 1 - \xi_{\text{region}}$, they make an unconstrained choice from all available destinations j according to probabilities p_{ij}.

We supposed that each region of preference consists of destinations intersecting with a buffer of radius ρ around a subbasin

TABLE 1 Considered covariates and their median and maximum values

Group	Covariate	Median	Maximum
Angler activeness in localities	Locality population (2019)	0.34×10^3	1286×10^3
	Mean income (2013)	36, 900 CAD	99, 600 CAD
	Median income (2013)	34, 600 CAD	78, 100 CAD
Fishing opportunities in subbasins	Total perimeter of waterbodies	350 km	1570 km
	Total area of waterbodies	4 km^2	2160 km^2
	Total perimeter with confirmed species	0 km	620 km
	Total area with confirmed species	0 km^2	1390 km^2
Infrastructure in subbasins	Population in 10 km range (2019)	0.4×10^3	1394×10^3
	Public campgrounds in 10 km range	1	19
Social media presence of subbasins	Total species upvotes (2018–2019)	0	196
	Total waterbody web page visits (2018–2019)	140	21,945

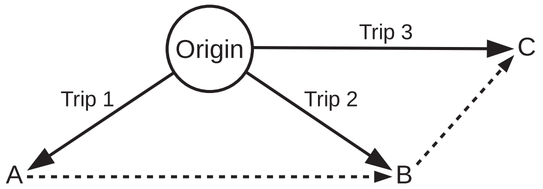

FIGURE 2 Possible sequence of trips to the destinations *A*, *B* and *C* for an angler with home locality 'origin'. The risk that the angler transports propagules or pathogens is highest for consecutively visited fishing locations, that is, for destinations *A* and *B* and for *B* and *C*. Our goal is to estimate the number of such consecutive fishing trips for any pair of destinations. Our dataset contains information on individual trips, but some trips may not have been recorded.

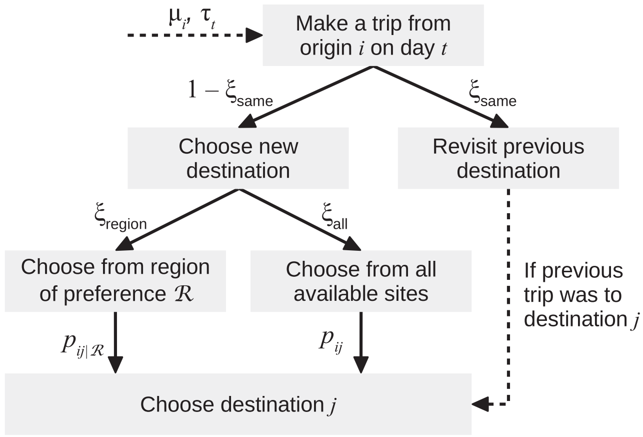
μ_i, τ_t
Make a trip from origin i on day t
$1-\xi_{\text{same}}$
ξ_{same}
Choose new destination
Revisit previous destination
ξ_{region}
ξ_{all}
Choose from region of preference $\mathcal{R}$
Choose from all available sites
If previous trip was to destination j
$p_{ij|\mathcal{R}}$
p_{ij}
Choose destination j

FIGURE 3 Visualization of anglers' decision-making process. The parameters μ_i and τ_t determine the expected rate at which anglers from origin *i* make trips on day *t*. When an angler chooses their destination, they may revisit their previous destination with probability ξ_{same}. Otherwise, they may either constrain their choice to their region of preference (with probability ξ_{region}) or make an unconstrained selection from all available destinations (with probability ξ_{all}). If they decide to constrain their choice to their region of preference $\mathcal{R}$, they choose destination *j* with probability $p_{ij|\mathcal{R}}$. Otherwise, they choose it with probability p_{ij}.

centre (Figure 4). Each angler's region of preference is fixed over time and chosen randomly. The probability $p_{i\mathcal{R}}$ that an angler from origin *i* has region of preference $\mathcal{R}$ is proportional to how likely they would choose a destination in $\mathcal{R}$ under the unconstrained strategy:

$$p_{i\mathcal{R}} = \frac{1}{\sum_{\tilde{\mathcal{R}}\in\mathfrak{R}} \sum_{j\in\tilde{\mathcal{R}}} p_{ij}} \sum_{j\in\mathcal{R}} p_{ij}. \tag{2}$$

Here, $\mathfrak{R}$ is the set of all potential regions of preference. The probabilities $p_{ij|\mathcal{R}}$ are defined accordingly as

$$p_{ij|\mathcal{R}} = \frac{p_{ij}}{\sum_{\tilde{j}\in\mathcal{R}} p_{i\tilde{j}}}. \tag{3}$$

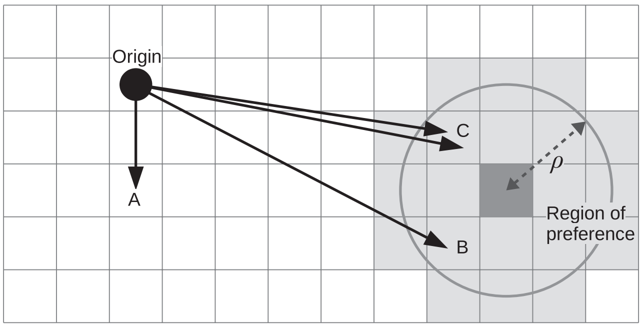

FIGURE 4 An example for a series of trips to destinations *A*, *B*, *C* and again *C* in order (depicted as black arrows) for an angler with the region of preference drawn in grey. Each grid cell represents a destination. The region of preference contains all destinations intersecting with the buffer of radius ρ drawn around the centre of the destination coloured dark grey. The angler may choose any of the available destinations but often selects destinations within their region of preference. Furthermore, anglers may tend to revisit destinations on consecutive trips (e.g. destination C). Note that subbasins are not square grid cells in practice but can take any shape.

The assumptions above lead to a simplified model on an aggregate level. A random angler from origin *i* will choose destination *j* with probability

$$\xi_{\text{region}} \sum_{\mathcal{R}\in\mathfrak{R}} p_{i\mathcal{R}} p_{ij|\mathcal{R}} + \xi_{\text{all}} p_{ij} = p_{ij} \tag{4}$$

unless they revisit their previous destination. Note that they cannot revisit a destination on their first trip. The probability that they choose destination *j* on their second trip is therefore $(1-\xi_{\text{same}})p_{ij} + \xi_{\text{same}}p_{ij} = p_{ij}$. By induction, the probability that a random angler from origin *i* chooses destination *j* is p_{ij}.

Since not all anglers in Alberta used the MyCatch app and app users may not have recorded all their trips, an additional submodel for the sampling process is needed to incorporate the data recorded via the app. We assumed that each angler decided randomly to install and use the app with probability ν_{app}, and that app users record a trip with probability ν_{record}.

2.3 | Computing expected trip counts

Based on the model introduced above, the expected number of consecutive angler trips to destinations j_1 and j_2 can be computed as follows. Let n_i be the number of anglers residing at origin *i*. Then, the expected number of trips that anglers from origin *i* make during the study period is $n_i\mu_iT$. Now consider the probability that, for any pair of consecutive trips, j_1 is the destination of the first trip and j_2 is the destination of the second trip. Recall that anglers may either revisit their previous location, constrain their destination choice to their region of preference or choose their destination freely. Hence, the mean number of consecutive trips by anglers from origin i to j_1 and j_2 during the study period is

$$\mu_{ij_1j_2} = \underbrace{n_i\mu_i T}_{\substack{\text{expected}\\ \text{trip count}}} \left(\underbrace{\xi_{\text{same}}\delta_{j_1j_2}p_{ij_1}}_{\substack{\text{prob. to travel } j_1 \to j_2 = j_1\\ \text{by choosing } j_1 \text{and revisiting it}\\ \text{without considering alternatives}}} + \underbrace{\left(1-\xi_{\text{same}}\right)}_{\substack{\text{prob. to consider}\\ \text{alternatives to}\\ \text{previous dest. } j_1}} \left(\xi_{\text{all}}^2 \underbrace{p_{ij_1}p_{ij_2}}_{\substack{\text{prob. to travel } j_1 \to j_2\\ \text{if both are chosen from}\\ \text{all available sites}}} \right.\right. \tag{5}$$

$$\left.\left. + \xi_{\text{all}}\xi_{\text{region}} \left(\underbrace{p_{ij_1}\sum_{\mathcal{R}:j_2\in\mathcal{R}} p_{i\mathcal{R}}p_{ij_2|\mathcal{R}}}_{\substack{\text{prob. to travel } j_1 \to j_2 \text{ if } j_1 \text{ is}\\ \text{chosen from all available sites}\\ \text{and } j_2 \text{ from a region of preference}}} + \underbrace{p_{ij_2}\sum_{\mathcal{R}:j_1\in\mathcal{R}} p_{i\mathcal{R}}p_{ij_1|\mathcal{R}}}_{\substack{\text{prob. to travel } j_1 \to j_2 \text{if } j_1 \text{ is}\\ \text{chosen from a region of preference}\\ \text{and } j_2 \text{ from all available sites}}} \right) + \xi_{\text{region}}^2 \underbrace{\sum_{\mathcal{R}:j_1,j_2\in\mathcal{R}} p_{i\mathcal{R}}p_{ij_1|\mathcal{R}}p_{ij_2|\mathcal{R}}}_{\substack{\text{prob. to travel } j_1 \to j_2 \text{ if both are}\\ \text{chosen from a region of preference}}} \right)\right).$$

Here, $\delta_{j_1j_2}$ is 1 if $j_1 = j_2$ and 0 otherwise. The right hand side of Equation (5) can be simplified to speed up computations, as we show in Appendix S2.

We computed the expected number $\mu_{j_1j_2}$ of consecutive trips to destinations j_1 and j_2 by summing $\mu_{ij_1j_2}$ over all origins i. To determine how many anglers access a destination j_2 after having visited a whirling-disease infested site, we furthermore summed the $\mu_{ij_1j_2}$ over all subbasins j_1 where the disease is present already. Based on these results, we also computed the number of these trips for each origin i.

2.4 | Submodels for day suitability, angler activeness and destination probabilities

The expected day suitability τ_t may change in weekly and seasonal cycles. We modelled these variations using the probability density function $f_{\text{vM}}(\cdot;\kappa,\theta)$ of the von Mises distribution, which is a cyclic distribution resembling the normal distribution (Lee, 2010). The shape of the function is controlled via the two parameters θ, determining the location of the mode, and κ, determining how sharp the maximum is. We defined the expected suitability τ_t of day t as follows:

$$\tau_t = c_{\text{norm}} \underbrace{\left(c_{\text{week}} + f_{\text{vM}}\left(2\pi\frac{t \bmod 7}{7}; \theta_{\text{week}}, \kappa_{\text{week}}\right)\right)}_{\text{weekly variations}} \underbrace{\left(c_{\text{year}} + f_{\text{vM}}\left(2\pi\frac{t \bmod 365}{365}; \theta_{\text{year}}, \kappa_{\text{year}}\right)\right)}_{\text{yearly variations}}. \tag{6}$$

The constants c_{year} and c_{week} are parameters controlling the amplitude and vertical shift of the weekly and seasonal cycles; the constant c_{norm} is chosen so that $\frac{1}{T}\sum_t \tau_t = 1$. If the study period includes leap years, τ_t must be adjusted accordingly.

As the 2104 trips in our dataset did not suffice to estimate the choice probabilities p_{ij} for all 180,000 pairs of localities and subbasins directly, we estimated μ_i and p_{ij} based on covariates on the origins and destinations. To that end, we applied the framework of gravity models. Gravity models estimate the mean number of trips between each origin and destination as a product of (1) the *repulsiveness* of the origin, proportional to the number of outbound trips; (2) the *attractiveness* of the destination, proportional to the number of inbound trips; and (3) a decaying function of the distance between origin and destination. Repulsiveness and attractiveness are typically functions of covariates characterizing the origins and destinations. In our model, the expected outbound traffic of origin i is given by the product $n_i\mu_i$ of angler count and activeness, and the expected traffic between origin i and destination j is $n_i\mu_i p_{ij}$. Therefore, the repulsiveness of origin i corresponds to the product $n_i\mu_i$, whereas the product of distance decay function and attractiveness defines the choice probabilities p_{ij}. Let d_{ij} be the linear distance between origin i and destination j, and let a_j be the attractiveness of destination j, measuring both the quantity and quality of fishing opportunities. Then,

$$p_{ij} = \frac{a_j D(d_{ij})}{\sum_{\tilde{j}} a_{\tilde{j}} D(d_{i\tilde{j}})}, \tag{7}$$

with the distance decay function D, which we define as

$$D\left(d_{ij}\right) = \frac{d_0^{\gamma_{\text{distance}}}}{d_0^{\gamma_{\text{distance}}} + d_{ij}^{\gamma_{\text{distance}}}}. \tag{8}$$

The parameter d_0 is the half saturation constant, given as the distance at which $D\left(d_{ij}\right) = \frac{1}{2}$.

To define an appropriate function to compute μ_i and a_j based on the covariates, we categorized the covariates into groups $\mathcal{X}$ (Table 1), each accounting for a different component that is necessary for high angler traffic between an origin and a destination (cf. Fischer et al., 2020). For each origin or destination k, we assigned the score $\left(\beta_{\boldsymbol{x}} x_k\right)^{\gamma_x}$ to each covariate $\boldsymbol{x} \in \mathcal{X}$, where x_k is the component of $\boldsymbol{x}$ corresponding to k and the parameters $\beta_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{x}}$ describe the impact of $\boldsymbol{x}$. We then added these individual scores to obtain a score for each group $\mathcal{X}$, so that a high score for *one covariate* suffices to make the group's score large. Finally, we multiplied the scores for the different groups, making a high score for all components necessary to boost the number of angler trips. With scaling constant c, we set

$$\mu_i = c \prod_{\substack{\text{origin covariate} \\ \text{groups}\,\mathcal{X}}} \left(1 + \sum_{\boldsymbol{x} \in \mathcal{X}} \left(\beta_{\boldsymbol{x}} x_i\right)^{\gamma_x}\right), \tag{9}$$

$$a_j = \prod_{\substack{\text{destination covariate} \\ \text{groups}\,\mathcal{X}}} \left(1 + \sum_{\boldsymbol{x} \in \mathcal{X}} \left(\beta_{\boldsymbol{x}} x_j\right)^{\gamma_x}\right), \tag{10}$$

where the *origin* and *destination covariate groups* are the locality and subbasin groups in Table 1.

2.5 | Fitting the model

We fitted the model via maximizing the likelihood associated with the recorded app data. However, fitting the complete model all at once is computationally costly due to the complicated form of the likelihood function and the large number of parameters. Therefore, we eliminated parameters by summing over certain quantities to obtain submodels with simpler likelihood functions. Furthermore, we made approximations via independence assumptions, disregarding the identity of anglers in some fitting stages (see below). Since most trips are made by independent anglers, our parameter estimates remain valid despite these simplifications (Varin, 2008).

We fitted the model in three steps: first, we considered the submodel for the day suitability ε_t; second, we estimated the angler activeness μ_i and the destination choice probabilities p_{ij}; and third, we estimated the parameters ξ_{same}, ξ_{region}, ξ_{all}, ν_{app}, ν_{record} and ρ modelling anglers' tendencies to constrain their trip choices and to record trips. Below, we briefly explain each of these steps; more details can be found in Appendix S3. In each of the steps, we exploited that (1) a Poisson random variable with a gamma distributed mean is negative binomially distributed and that (2) the mixture of a negative binomial and a binomial distribution remains negative binomially distributed (Villa & Escobar, 2006).

2.5.1 | Day suitability

We estimated the expected day suitability τ_t by fitting the distribution of the total number N_t of recorded angler trips on day t to the data. According to our model, N_t is negative binomially distributed with dispersion parameter $\frac{\alpha}{\tau_t}$ and mean $\nu_{\text{record}}\tau_t \sum_i \tilde{n}_i \mu_i$, where $\tilde{n}_i$ is the number of app users in locality i. As $\tilde{n}_i$ is a random variable itself and constant over the study period, it is not straightforward to derive the exact distribution of N_t. However, since N_t describes the aggregate trip counts of many anglers, who rarely make more than one trip per day, it is reasonable to consider trips as mutually independent on each day. Then, N_t is negative binomially distributed with dispersion $\frac{\alpha}{\tau_t}$ and mean $\tau_t \overline{\mu}$, where $\overline{\mu} = \nu_{\text{app}} \nu_{\text{record}} \sum_i n_i \mu_i$. Hence, by fitting the distribution of N_t, we obtained estimates for the parameters α, $\overline{\mu}$ and those controlling the shape of τ_t. See Appendix S3.1 for further details.

2.5.2 | Angler activeness and destination choice probabilities

To estimate the angler activeness values μ_i and the destination choice probabilities p_{ij}, we considered the trip counts N_{ijt} for origin–destination pairs (i,j) and days t. We fitted the joint distribution of the N_{ijt} to our data for all origin–destination pairs and days of the study period. With the independence approximation from the previous section, each N_{ijt} follows a negative binomial distribution with dispersion parameter $\frac{\alpha}{\tau_t}$ and mean $\nu_{\text{app}} \nu_{\text{record}} \tau_t n_i \mu_i p_{ij}$. To improve the computational performance, we also considered the trips from different localities as mutually independent. The values τ_t were known from the previous fitting stage. By fitting N_{ijt} to the observed values, we obtained estimates for the parameters of μ_i and p_{ij}. The scaling constant c and the probabilities ν_{app} and ν_{record} are not identifiable in this fitting stage, and we replaced them with a parameter $C = \nu_{\text{app}} \nu_{\text{record}} c$ here. Refer to Appendix S3.2 for a method to compute the likelihood efficiently.

2.5.3 | Remaining parameters

To fit the choice parameters ξ_{same}, ξ_{region}, ξ_{all}, ν_{app} and ν_{record}, we considered each angler and their trips individually. First, we determined the likelihood for the temporal sequence of their trips. Then we computed the likelihood for their destination choices given the timing of the trips. Because the destination choices for consecutive trips are not independent and we need to consider unknown numbers of intermediate unrecorded trips, the likelihood function has a complicated form involving convolutions and special functions. Nonetheless, it can be computed numerically with reasonable effort if partial intermediate results are reused where possible. We refer the reader to Appendix S3.3 for details. Dependencies between trips of *different* anglers were disregarded at this stage so as to facilitate efficient computation.

To fit the radius ρ of the inscribed circle of anglers' regions of preferences, we conducted a grid search with steps of 1 km in the interval between 10 km and 80 km. For each considered value of ρ, we maximized the likelihood with respect to the remaining parameters; finally, we chose the radius leading to the maximal likelihood. We conducted a grid search, because the regions of preference are discrete entities, making gradient descent methods inapplicable to fit ρ.

2.5.4 | Optimization methods and model selection

We used a combination of multiple optimization algorithms to maximize the likelihood. We applied the differential evolution algorithm (Storn & Price, 1997) for a global search of the parameter space, and improved upon the results via gradient-based search algorithms (Byrd et al., 1995; Kraft, 1988; Nocedal & Wright, 2006). Details can be found in Appendix S4. We implemented the model in the programming language Python (version 3.7) along with the Scipy libraries (Jones et al., 2001).

To decide which covariates and parameters should be included in the model without overfitting, we used the information criterion by Akaike (1974) (AIC). This metric is particularly suitable if the modelling goal is prediction (Ghosh & Samanta, 2001). When fitting the day suitability function τ_t, we considered simplified models with the parameters c_{week}, c_{year}, κ_{week} and κ_{year}, (Equation (6)) set to 0 respectively. For the angler activeness μ_i and destination choice probabilities p_{ij}, we considered models with any combination of the parameters $\beta_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{x}}$ (Equations (9) and (10)) set to zero. Only for covariates $\boldsymbol{x}$ that had 0-values for some origins or destinations, we tested models with $\gamma_{\boldsymbol{x}} = 1$ instead. We furthermore tested models without the parameter d_0 (Equation (8)). We searched for the model with the minimal AIC value by using a branch and bound algorithm (Appendix S4.2). This allowed us to find the optimal model without having to consider all potential candidates.

Note that we made approximations via independence assumptions, which violate the underlying assumptions of AIC. Hence, the metric may tend to favour overfitting models. We therefore chose the simplest model among those with AIC values less than 10 units higher than the minimal AIC. Models whose AIC value is more than 10 units higher than the minimal AIC may have little empirical support (Burnham & Anderson, 2004).

2.5.5 | Model evaluation

We evaluated the trustworthiness of our parameter estimates by computing confidence intervals using a method based on the profile likelihood (Fischer & Lewis, 2021). Note that since we did not evaluate the joint model all at once and made approximations via independence assumptions, the true confidence intervals may be larger. Nonetheless, the approximate confidence intervals are suited to detect estimability issues and problems arising from multicollinearity of covariates.

To validate the submodel for angler traffic between localities and subbasins, we randomly split the app data into a training (fitting) and a testing (validation) dataset, each containing observations for half of the anglers respectively. Note that a random split is in line with our purpose of evaluating the model accuracy in predicting unreported trips, and hence, a temporal split used for evaluating the model accuracy in making future predictions is not needed (Ramazi, Kunegel-Lion, et al., 2021). We fitted our model to the training data and computed the mean yearly number of recorded angler trips for each origin, destination and origin–destination pair. Then we plotted the predicted values against the observed values.

The purpose of the submodel for locality-to-subbasin traffic was to fill data gaps stemming from the limited number of angler trips in our dataset. To ensure that the model is suited to fill these gaps without introducing additional error, we computed the mean absolute errors between the submodel's results and the observations from the testing data. We then compared the resulting values with those obtained by using direct estimates from the training data without an additional model.

Our model yields absolute estimates of angler traffic based on voluntarily reported trip data without using a priori information on how many trips anglers actually made. To assess the accuracy of our estimates of the trip frequency, we compared the number of days anglers go fishing per year as per our model (see Appendix S5) with empirical data collected in a mail-out survey by the Department of Fisheries and Oceans Canada in 2016 (DFO, 2019).

To facilitate a qualitative comparison of our model's accuracy with predictions from similar studies, we computed Nagelkerke's pseudo-R^2 (Nagelkerke, 1991). This measure indicates how well the model performs in comparison to a noninformative null model. In contrast to the classical R^2, Nagelkerke's pseudo-R^2 can be applied even if the data are not assumed to be normally distributed with identical standard deviations. We computed pseudo-R^2 values for each model stage: the day suitability model; the joint day suitability, angler activeness and destination choice model; and the complete model with all submodels. As null models, we used negative binomial distributions treating all days, localities and subbasins similarly. The parameters ξ_{same} and ξ_{region} capturing anglers' local preferences were set to zero.

3 | RESULTS

The selected day suitability model contained all considered parameters (Table 2). The traffic was estimated highest on Saturdays and to peak on 14 July. The rates of fishing trips during the weekly peak were estimated to be 2.27 times higher than on the weekly low; yearly cycles changed the expected traffic rate by up to factor 6.6. The confidence intervals were relatively narrow for most of the parameters; only the shape parameter for the weekly cycles had a large upper confidence interval bound. The temporal distribution of trips and the expected trip rates estimated by the model are displayed in Figure 5.

The angler activeness and destination choice model with the least number of parameters and ΔAIC ≤ 10 included 15 parameters (Table 3; ΔAIC = 1.5). The model with minimal AIC included the number of website visits as an additional covariate. The selected model uses localities' population counts and the mean income numbers to estimate angler activeness. The angler activeness varied by up to factor 4.5 among localities. For a locality with median characteristics, a population increase of 1000 increases angler activeness by 6%, and a mean income increase by CAD 1000 increases activeness by 2%.

The localities from which anglers make the most consecutive trips between infested and uninfested subbasins were the population centres Calgary and Edmonton. Setting the trip count into relation with the number of registered anglers, Calgary, located in direct proximity to the infested area, had higher relevant traffic with 2.86 high-risk trips per registered angler and year as compared to 0.65 for Edmonton. Considering all inhabitants, rural municipalities had higher relative trip counts: the 100 localities with the most high-risk trips per inhabitant had less than 8000 inhabitants.

TABLE 2 Parameters and estimates along with approximate 95% confidence intervals after fitting the model to the daily trip counts

Parameter	Explanation	Estimate	Confidence interval	
α	Dispersion parameter	0.33	0.25	0.42
$\overline{\mu}$	Mean total recorded trips per day	2.56	1.52	4.34
c_{week}	Week addition constant	0.45	0.24	0.7
θ_{week}	Week location constant	5.64	5.5	5.8
κ_{week}	Week shape constant	2.87	1.36	∞
c_{year}	Year addition constant	0.12	0.1	0.15
θ_{year}	Year location constant	3.35	3.29	3.42
κ_{year}	Year shape constant	7.4	6.53	8.29

The subbasin attractiveness was estimated based on the perimeter of the waterbodies in the subbasins, the surface area of the waterbodies with confirmed species, the number of campgrounds and the website species upvotes. The attractiveness values varied greatly among subbasins, by up to factor 1179. For a subbasin with median characteristics, an increase in the total waterbody perimeter by 500 km increased attractiveness by 8.7%, whereas an increase of 1000 km increased attractiveness by factor 3. If the total area of waterbodies with confirmed species were increased by 10 km^2, attractiveness would rise by 69%. An additional campground increased attractiveness by 20%, and an additional positive species vote by 180%. The traffic between localities and subbasins decreased in square order of their distance; that is a subbasin twice as far as a similar subbasin was only 25% as likely to be chosen as fishing destination.

The estimates for the remaining choice parameter are displayed in Table 4. The fraction of anglers using the app was estimated to be 0.22%, and the estimated probability that app users report a trip was 0.05. The dispersion parameter, modelling the impact of stochastic events on anglers' daily trip rates was estimated 11.8. The model predicts that on 64% of their trips, anglers revisit their last destination. They choose a destination in their region of preference in 54% of their trips and choose the destinations for the remaining 46% trips from all over Alberta. The estimated radius ρ for the inscribed circle of regions of preference was 31 km. This translates to a mean radius of 54.7 km for the regions of preference. The parameter confidence intervals were relatively narrow except for the probability that app users report a trip (Table 4).

Angler trips were estimated to be strongest between subbasins located close to metropolitan areas, with estimates up to 22.3 thousand (95% confidence interval $\left[14.5 \times 10^3, 32.6 \times 10^3\right]$) directly consecutive angler trips per year (Figure 6a). The subbasin most at risk of receiving anglers infested with whirling disease propagules were those located close to larger cities and in proximity to the already infested area (Figure 6b). The subbasin with the highest inflow of high-risk trips was estimated to receive 27.7 thousand ($\left[18.1 \times 10^3, 40.3 \times 10^3\right]$) such trips per year. The estimated mean number of fishing days per angler and year was 20.5 ($[12.5, 34.1]$) as per our model. For comparison, the corresponding estimate from a 2016 mail-out survey was 18 (standard deviation between 0.9 and 2.7) (DFO, 2019).

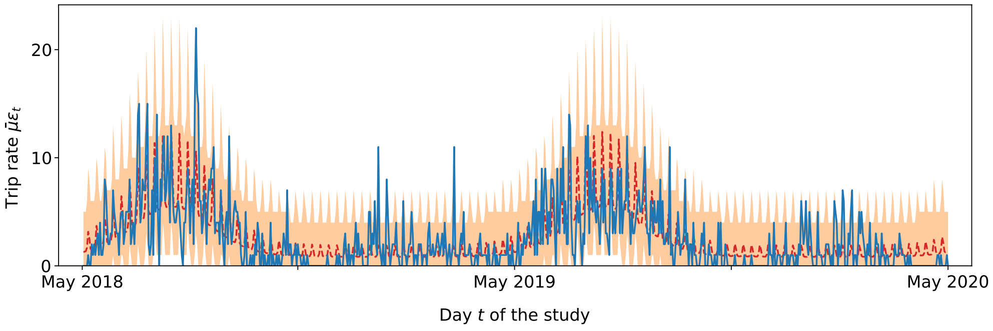

FIGURE 5 Observed and modelled trip rates for each day of the study period. The observed number of trips N_t is drawn as solid blue line, the predicted mean of N_t as dashed red line and the predicted 95% confidence range as light red area.

TABLE 3 Parameters and estimates along with approximate 95% confidence intervals after fitting the model for angler activeness and destination choice. Note that although we report all parameters on the original scale, we worked with log-transformed parameters internally to avoid numerical errors due to extreme parameter values.

Parameter	Explanation	Estimate	Confidence interval	
α	Dispersion parameter	6.99	4.91	9.59
C	Scaling constant for the mean daily number of recorded trips	1×10^{-9}	3.51×10^{-11}	5.12×10^{-8}
δ_0	Distance of half choice-probability decay $[\text{km}]$	28.74	22.02	35.91
γ_{distance}	Distance exponent	2.09	1.97	2.22
$\beta_{\text{population}}$	City population factor $\left[1/10^3\right]$	6.95×10^{29}	4.29×10^{29}	1.23×10^{30}
$\gamma_{\text{population}}$	City population exponent	0.14	0.09	0.16
$\beta_{\text{mean income}}$	Mean income factor $\left[1/\left(10^3\text{CAD}\right)\right]$	909.1	11.61	2.3×10^6
$\gamma_{\text{mean income}}$	Mean income exponent (not included in the model)	1	—	—
$\beta_{\text{perimeter}}$	Water perimeter factor $\left[1/\left(10^3\text{km}\right)\right]$	0.82	0.76	0.94
$\gamma_{\text{perimeter}}$	Water perimeter exponent	6.76	3.89	9.62
$\beta_{\text{area confirmed}}$	Water area confirmed factor $\left[1/\left(10^3\text{km}^2\right)\right]$	38.95	9.25	305.8
$\gamma_{\text{area confirmed}}$	Water area confirmed exponent	0.4	0.24	0.67
$\beta_{\text{campground}}$	Campground factor $[1]$	0.25	0.12	0.62
$\gamma_{\text{campground}}$	Campground exponent	1	0.65	1.46
$\beta_{\text{species vote}}$	Species vote factor $[1]$	2.8	1	8.4
$\gamma_{\text{species vote}}$	Species vote exponent	0.57	0.5	0.65

TABLE 4 Parameters and estimates along with approximate 95% confidence intervals after fitting the model for the individual angler choices

Parameter	Explanation	Estimate	Confidence interval	
α	Dispersion parameter	11.81	8.65	15.6
ξ_{same}	Probability to revisit the previous destination	0.64	0.53	0.79
ξ_{region}	Probability to constrain the destination choice to the region of preference	0.54	0.5	0.59
ν_{app}	Probability to use the app	0.0022	0.0021	0.0023
ν_{record}	Probability to record a trip	0.052	0.022	0.104

The pseudo-R^2 values of the model components decreased as more complexity was added. The submodel for the day suitability achieved a pseudo-R^2 of 0.47. Adding the component for the origin-dependent trip rates and the destination choice probabilities yielded a value 0.35. Predicted and observed values for this model component are depicted in Figure 7. The joint model including the remaining choice parameters had a pseudo-R^2 of 0.26.

The submodel for the traffic between origins and destinations estimated the outflow from origins with a mean error of 1.48, the inflow to destinations with a mean error of 1.6 and the traffic between individual pairs with a mean error of 0.0073. The corresponding values obtained by using the data directly were 1.83, 1.81 and 0.0074. That is, applying the model did not increase the error.

4 | DISCUSSION

Mobile apps exist for a variety of outdoor activities (e.g. birding, hiking and fishing) that could be related to the spread of animal diseases and invasive species. These apps can yield highly detailed individual-specific, spatially and temporally representative data and provide valuable insights into the traffic of anthropogenic vectors of invasive species and pathogens (Papenfuss et al., 2015; Venturelli et al., 2017). However, although the datasets collected via mobile apps can be large, they often cover only a small fraction of all trips of potential vectors, making direct estimates via the data's empirical distribution error-prone. Our results indicate that modelling approaches can reduce this issue and provide additional insights into the mechanisms behind vector movement.

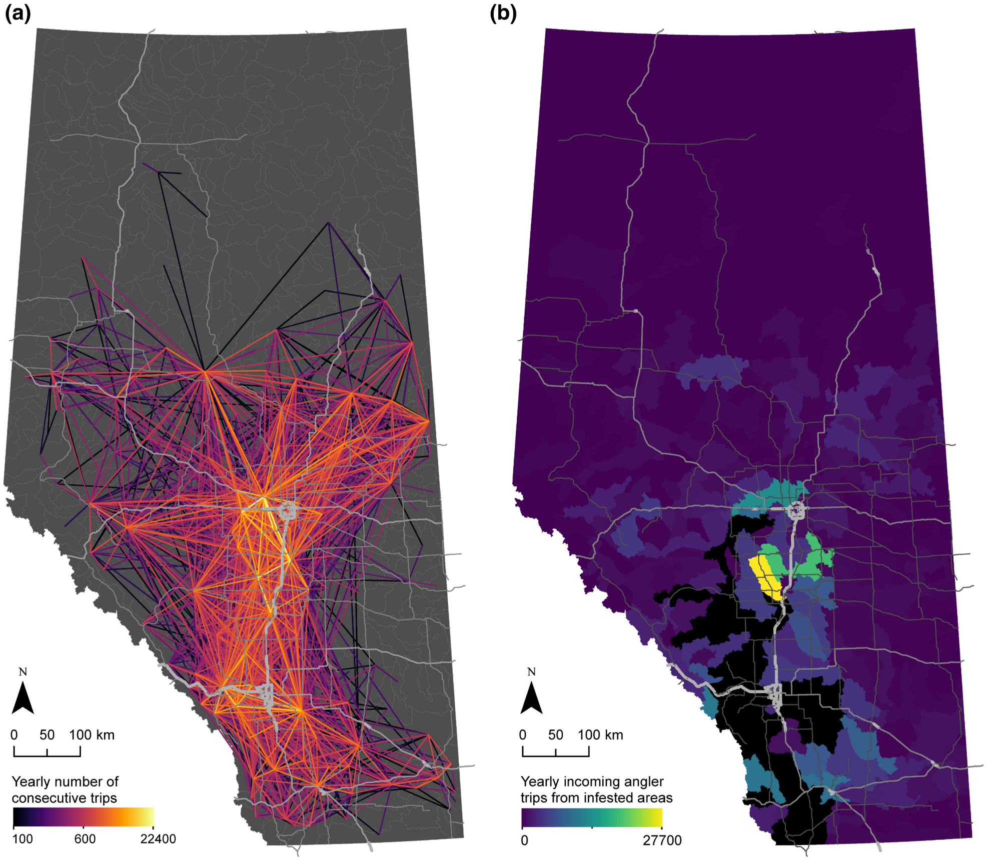
(a)
(b)
N
0 50 100 km
Yearly number of consecutive trips
100
600
22400
N
0 50 100 km
Yearly incoming angler trips from infested areas
0
27700

FIGURE 6 (a) Number of consecutive trips to subbasin pairs and (b) total number of incoming trips by potentially infested anglers. In (a) only subbasin pairs with more than 100 trips per year are shown. In (b) black colours depict subbasins that are already infested (March 2020).

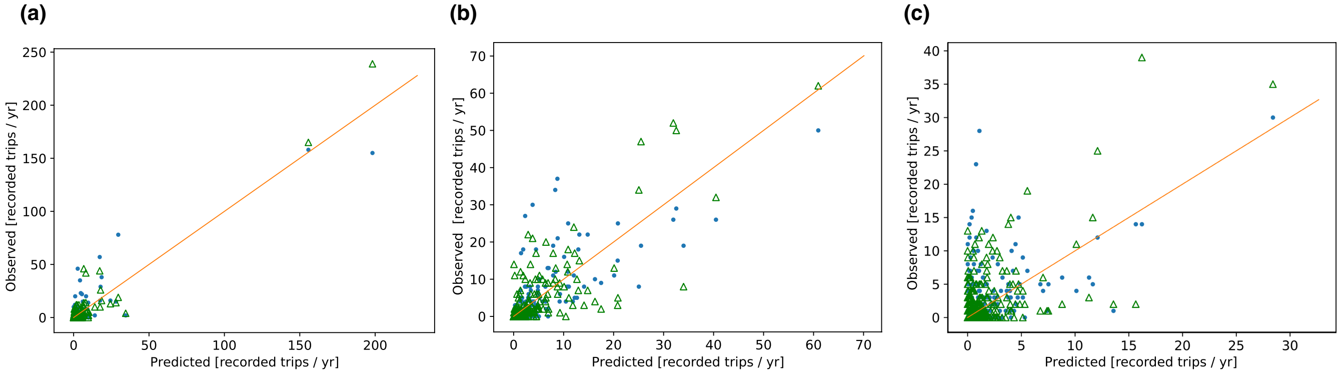

FIGURE 7 Predicted and observed mean values of yearly recorded angler trips (a) by origin locality, (b) by destination subbasin and (c) by locality–subbasin pair. The values used to fit the model are depicted as solid blue circles; the values computed from the independent fitting dataset are drawn as hollow green triangles. The orange line indicates where predictions and observations would coincide.

Conversely, models can profit strongly from the new data source. Although mail-out or online surveys could collect the same data as apps in principle, our modelling approach based on app data has the following advantages:

(1) *Increased accuracy*. Apps can be downloaded by users from different geographical areas, and data may be collected over extended time periods. Therefore, inference from app data is generally less sensitive to local and temporal peculiarities, and modellers can identify and account for the sources of spatial and temporal heterogeneity. This makes the estimates more accurate, especially when results are extrapolated into the future or to larger geographical scales. Furthermore, the temporal fingerprint of app data records permits a longitudinal study design. By considering the day-to-day variations of the data, the unexplained recurring stochasticity in individual decisions can be distinguished from systemic errors due to misspecified models. Without this distinction (e.g. Drake & Mandrak, 2010; Muirhead et al., 2011; Muirhead & MacIsaac, 2011), residuals would be solely attributed to the stochasticity in the

individuals' decisions, and the dispersion parameter would be overestimated (Fischer et al., 2020). Then, low-traffic angler flows gain an inordinate weight when the model is fitted to observations, resulting in decreased model accuracy (see Appendix S6).

(2) *Estimates account for reduced vector mobility due to individual preferences*. Accounting for anglers' individual preferences allowed us to estimate the frequency at which they switch destinations, potentially transporting propagules. We found that in 64% of their trips, anglers revisit their previous destination and hence do not spread invasive species and pathogens to new areas. Furthermore, anglers tend to choose half of their fishing destinations from spatially contained areas. This suggests that models disregarding the correlations within anglers' destination choices (e.g. Bossenbroek et al., 2001; Leung et al., 2004) are prone to overestimating traffic between distant destinations.

(3) *Absolute estimates of vector traffic can be obtained without additional survey data*. It is difficult to obtain absolute traffic estimates from survey or app data without knowing which fraction of trips surveyed individuals or app users report. However, by considering anglers' tendency to revisit previous destinations, we were able to infer this missing information. If an angler does not record all their trips, the probability that the next trip they record has the same destination as their previously recorded trip decays with time, because they may make additional unrecorded trips to other destinations in the mean time. The slope at which the fraction of consecutively recorded trips with same destinations decays with the intermediate time depends on the trip recording probability. This makes it possible to infer this information from the data.

Absolute estimates of traffic enable modellers to link traffic to invasion or infection success. This link is needed to predict the spread of a disease or invasive species (Lewis et al., 2016). Although invasion success can be estimated based on relative traffic estimates if historical invasion data are available for the studied area (Leung et al., 2004; Muirhead et al., 2006; Potapov et al., 2011), these estimates will remain site specific unless the scaling of the traffic is known or the same traffic model is at the other site. Transferring a traffic model to a new site requires that similar data are available at the new site and that vectors behave the same. Absolute traffic estimates allow modellers to estimate the establishment success per individual vector and to transfer such information to or from other study areas.

4.1 | Validity of the estimates

Exploiting the connection between reported destinations and the completeness of the data, we estimated how many days an average angler goes fishing in Alberta per year. Our estimate for this quantity agreed with an independent estimate, differing only by about one standard deviation of the independent estimate. This suggests that our approach estimates the overall angler trip frequency accurately and hence can be used to obtain absolute traffic estimates even if the fraction of reported trips is unknown. Note, however, that the confidence interval for our estimate was relatively large and had an upper bound 60% higher than the estimated value. Hence, additional surveys determining the total trip count—and thereby the fraction of reported trips—remain worthwhile to reduce model uncertainty. We refrained from incorporating this additional information, because our estimate of the total trip counts was already in agreement with the independent estimate.

The predicted versus observed analysis of the submodel for angler activeness and destination choice probabilities indicated a reasonable estimation accuracy. Comparison with the raw data showed that the model extrapolates the limited available data to all origin–destination pairs without introducing additional error—the model did even slightly better than the direct estimates. This is due to the stochasticity inherent in the system and the limited number of available trip records.

The pseudo-R^2 values we computed for different model components decreased as more complexity was added. This is expected, because the level of detail of the validation data increased as additional model components were considered. The differing level of detail makes it difficult to compare our pseudo-R^2 values to similar metrics obtained in studies with less rich data sources (e.g. Chivers & Leung, 2012; Drake & Mandrak, 2010). Nonetheless, the pseudo-R^2 value we reported may be a helpful benchmark for future studies on a similar system.

4.2 | Management implications

Our approach facilitates management in three ways: (1) it provides location-specific risk proxies; (2) it shows which locations are best connected and might thus be potential hubs for secondary infections; and (3) it helps to identify the origins of the agents most at risk of spreading the disease. While risk estimates facilitate early detection and rapid response to new infections, the latter two points help targeting management actions to the subbasins and localities where they are most effective.

(1) *Risk proxies*. The subbasins with the highest propagule inflow were those encompassing extended water areas with confirmed fish presence and located close by a population centre in proximity of an infected subbasin. Our results indicate that anglers show a strong preference for fishing destinations near their homes, which also agrees with earlier studies (Drake & Mandrak, 2010; Fischer et al., 2020; Papenfuss et al., 2015). Both the inscribed radius of the regions of preference and the distance at which traffic decays by half were at about 30km, suggesting this as main spatial scale of angler traffic. Besides water area and fish species confirmations, the number of campgrounds was another useful indicator of attractiveness, probably because they are typically built in scenic areas attractive to outdoor tourists.

The total waterbody circumference (i.e. shoreline) was positively connected with angler traffic as well, but the best-AIC model did not incorporate this covariate in conjunction with fish presence data, potentially due to incomplete species data in smaller rivers. Interestingly, the number of visits of waterbody-specific websites was not a worthwhile additional predictor for attractiveness—perhaps because this number is also affected by the waterbodies' proximity to angler origins and their repulsiveness.

(2) *Potential hubs*. Uninfected subbasins with strong connections to other uninfected subbasins might become significant hubs for secondary infections. This makes these well-connected subbasins a primary target for surveillance and rapid response measures. In our model, these subbasins have similar characteristics to those receiving most high-risk trips (see above) except for being located farther away from present infections.

(3) *Most relevant angler origins*. Education and outreach measures may be applied with different intensity in different localities. According to our results, maximal high-risk trip counts per inhabitant are found in rural areas, where the density of anglers per inhabitant is higher there than in cities. Hence, general outreach may be most effective in rural areas close to the edge of the infested area. In contrast, the number of high-risk trips per angler was also high in cities, making cities close to the infected area good soil for outreach measures specifically targeting anglers. Outreach may be cheaper and hence more cost-effective in cities than in rural areas.

4.3 | Limitations and potential extensions

Voluntarily reported data such as app data can be prone to a variety of biases ranging from demographic bias to avidity bias (Venturelli et al., 2017). For example, app usage may be higher among the young population from urbanized areas, and active sport fishers may use the app more frequently. As a result, certain demographic groups may be underrepresented, and highly active app users may influence the estimates disproportionately. This can lead to biased and overconfident results.

For the data from the MyCatch app used in this study, no spatial bias was detected (Johnston et al., 2021). This increases the credibility of our results. Nonetheless, some potential sources of bias remain, and other datasets may suffer from stronger bias. This issue could be addressed with additional data. If demographic and socioeconomic data are collected along with the other app data, these covariates could be directly incorporated into the model. If an independent sample with these data is available, it can be used to weigh the observations differently (Chen et al., 2020).

Our model uses socioeconomic covariates to estimate the activeness of anglers at different localities. Although incorporating population counts and the mean income in localities improved the model fit significantly, angler activeness is unlikely to be uniform within localities. Since there is no obvious mechanistic justification for the dependency of angler activeness on the covariates we used, we emphasize the phenomenological nature of the model and refrain from further analysis of the underlying mechanisms.

We also assumed that the individual preferences of individuals do not change with time. This assumption may be inaccurate if extended time periods are considered, and individuals may change both their region of preference and their home place. As a result, the connectivity of close-by locations in directly consecutive trips might be underestimated. This issue could be addressed by splitting the dataset into subsets for different time periods and to treat them as independent replicates of one another.

By fitting model components in separate steps, we obtained multiple estimates for the dispersion parameter α. This indicates that some sources of stochasticity, such as weather, act on small scales only and have a reduced impact on an aggregate level. This is a common effect seen in ecological models (Dungan et al., 2002) and difficult to resolve without ignoring spatial correlations completely or strongly increasing the model's complexity. However, since the estimates for the other parameters were insensitive to moderate variations of the dispersion parameter (Appendix S7), our estimates for the mean angler traffic remain valid despite the scale dependency of the dispersion parameter.

Noting that propagules may be washed out at any site visited by anglers, we concentrated on estimating the number of consecutive fishing trips to distinct subbasins. This is in line with existing studies on invasive species transport (Bossenbroek et al., 2001; Leung et al., 2004; Muirhead & MacIsaac, 2011; Potapov et al., 2011). An alternative approach is to consider all trips that anglers make within the time frame propagules may survive (Papenfuss et al., 2015). Note, however, that we assumed that anglers choose their destinations independently of the past unless they revisit their previous destination. Therefore, the number of higher-order trips between two subbasins equals the count of directly consecutive trips unless a constrained time frame is considered. Thus, incorporating higher-order trips comes down to determining a sensible time-scale for propagule decay.

We focused on intraprovincial angler traffic, but our model could be extended to also consider nonresident anglers visiting the province on vacation trips. Within our model framework, such anglers could be added to the populations of the localities where they reside temporally. However, in Alberta, nonresident anglers are responsible for only 2.6% of the angler traffic (measured in yearly fishing days) (DFO, 2019), which is below the error margin of our estimates.

4.4 | Alternative approaches

Our model's complexity is driven by the challenges stemming from missing trip records in app data. If complete trip records were available, a phenomenological gravity model for the connectedness of destinations (e.g. Chivers & Leung, 2012; Muirhead & MacIsaac, 2011; Potapov et al., 2010) could be constructed directly. However, complete trip records are often difficult to obtain without strong simplifying assumptions, such as that vectors have visited all destinations they reported with the same frequency (Potapov et al., 2010).

If data on the temporal progression of the disease or invasion are available, the connectedness between destinations could also be inferred based on the observed infestation dynamics (Bossenbroek et al., 2001; Leung et al., 2004). However, the resulting traffic estimates can typically not be validated due to the lack of data. Besides, this approach makes it necessary that an establishment model exists that accounts for on-site conditions affecting establishment success.

5 | CONCLUSIONS

The increasingly widespread use of mobile apps by anglers, hikers, campers and other potential vectors of invasive species and pathogens opens new opportunities for research and management. To exploit the full potential of this new data source, models accounting for spatial, temporal and individual heterogeneity are needed. The presented model demonstrates the wealth of information that can be gained from app data, including (1) temporally explicit estimates of vector traffic between cities and waterbodies, (2) estimates of how often vectors choose new trip destinations, potentially carrying propagules to other places and (3) the spatial scale at which individual local preferences play a major role in vectors' decisions. Our results suggest that ignoring individual-specific components in vectors' decision-making can bias estimates by underestimating local traffic and overestimating long-distance traffic. We furthermore showed that incorporating vectors' tendency to revisit past locations can bridge the common data gap arising from incomplete trip records reported by app users.

ACKNOWLEDGEMENTS

All authors are thankful for the funding received from Alberta Environment and Parks. In addition, P.R. and M.A.L. thankfully acknowledge an NSERC Discovery Grant and M.A.L. also gratefully acknowledges a Canada Research Chair. All authors thank the anglers who voluntarily provided data via the MyCatch app and the Anglers' Atlas website. Furthermore, the authors thank the Goldstream Publishing staff, who helped prepare the data for the analysis as well as the members of the Lewis Research Group, who contributed through helpful feedback and discussions. Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

Pouria Ramazi, Sean Simmons, Mark S. Poesch and Mark A. Lewis conceived the project. Sean Simmons collected and prepared the app data and the website data. All authors contributed to the methods; Samuel M. Fischer finalized and implemented the model. Samuel M. Fischer and Pouria Ramazi led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

CONFLICT OF INTEREST

Sean Simmons is founder and president at Angler's Atlas and MyCatch. The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The sources of the data used in this study are listed below. The original code used in this research is available via Zenodo https://doi.org/10.5281/zenodo.7499230 (Fischer et al., 2023). Please refer to the project's GitHub repository for the most up-to-date version of the code: https://github.com/vemomoto/indspecvemo. The compiled dataset used in this study is available via the Dryad Digital Repository https://doi.org/10.5061/dryad.6m905qg3j (Fischer et al., 2022).

DATA SOURCES

Data	Source	URL
Angler App Data Website Visit Data Species Vote Data Waterbody GIS Data	Goldstream Publishing	The raw data are not available online. A compilation of the data used as input for the model will be made available along with the article
Subbasin GIS Data	Government of Alberta (Alberta Environment and Parks)	Original URL (not available anymore): maps.alberta.ca/genesis/rest/services/Hydrologic Unit Code Watersheds of Alberta Potential alternative data source: geospatial.alberta.ca
Locality GIS Data	Open Street Map	www.openstreetmap.org
Campground GIS Data	USCAmpgrounds	www.uscampgrounds.info
Angler Licence Count Data	Government of Alberta (Alberta Environment and Parks)	Retrieved through a standard data request from the Fish and Wildlife Management Information System www.alberta.ca/access-fwmis-data.aspx
Population Count Data	Government of Alberta	open.alberta.ca/opendata/alberta-municipal-affairs-population-list
Population Income Data	Government of Alberta	open.alberta.ca/dataset/labour-income-profile-for-all-forward-station-areas-city-totals-and-rural-postal-codes-canada-2013

ORCID

Samuel M. Fischer https://orcid.org/0000-0001-8913-9575
Pouria Ramazi https://orcid.org/0000-0003-4906-0090
Mark S. Poesch https://orcid.org/0000-0001-7452-8180
Mark A. Lewis https://orcid.org/0000-0002-7155-7426

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bossenbroek, J. M., Johnson, L. E., Peters, B., & Lodge, D. M. (2007). Forecasting the expansion of zebra mussels in the United States. *Conservation Biology*, *21*, 800–810. https://doi.org/10.1111/j.1523-1739.2006.00614.x

Bossenbroek, J. M., Kraft, C. E., & Nekola, J. C. (2001). Prediction of long-distance dispersal using gravity models: Zebra mussel invasion of inland lakes. *Ecological Applications*, *11*, 1778–1788. https://doi.org/10.1890/1051-0761(2001)011[1778:POLDDU]2.0.CO;2

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. https://doi.org/10.1177/0049124104268644

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208. https://doi.org/10.1137/0916069

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, *115*, 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Chivers, C., & Leung, B. (2012). Predicting invasions: Alternative models of human-mediated dispersal and interactions between dispersal network structure and Allee effects. *Journal of Applied Ecology*, *49*, 1113–1123. https://doi.org/10.1111/j.1365-2664.2012.02183.x

Cushman, J. H., & Meentemeyer, R. K. (2008). Multi-scale patterns of human activity and the incidence of an exotic forest pathogen. *Journal of Ecology*, *96*, 766–776. https://doi.org/10.1111/j.1365-2745.2008.01376.x

DFO, D.o.F.a.O.C. (2019). *Survey of recreational fishing in Canada, 2015*. OCLC: 1199132468.

Drake, D. A. R., & Mandrak, N. E. (2010). Least-cost transportation networks predict spatial interaction of invasion vectors. *Ecological Applications*, *20*, 2286–2299. https://doi.org/10.1890/09-2005.1

Drake, D. A. R., & Mandrak, N. E. (2014). Bycatch, bait, anglers, and roads: Quantifying vector activity and propagule introduction risk across lake ecosystems. *Ecological Applications*, *24*, 877–894. https://doi.org/10.1890/13-0541.1

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M. J., Jakomulska, A., Miriti, M., & Rosenberg, M. S. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, *25*, 626–640. https://doi.org/10.1034/j.1600-0587.2002.250510.x

Elwell, L. C. S., Stromberg, K. E., Ryce, E. K., & Bartholomew, J. L. (2010). *Whirling disease in the United States: A summary of progress in research and management*. Proceedings of the Wild Trout X Symposium, West Yellowstone, Montana.

Ferrari, M. J., Bjørnstad, O. N., Partain, J. L., & Antonovics, J. (2006). A gravity model for the spread of a pollinatorborne plant pathogen. *The American Naturalist*, *168*, 294–303. https://doi.org/10.1086/506917

Fischer, S. M., Beck, M., Herborg, L. M., & Lewis, M. A. (2020). A hybrid gravity and route choice model to assess vector traffic in large-scale road networks. *Royal Society Open Science*, *7*, 191858. https://doi.org/10.1098/rsos.191858

Fischer, S. M., & Lewis, M. A. (2021). A robust and efficient algorithm to find profile likelihood confidence intervals. *Statistics and Computing*, *31*, 38. https://doi.org/10.1007/s11222-021-10012-y

Fischer, S. M., Ramazi, P., Simmons, S., Poesch, M. S., & Lewis, M. A. (2022). Data for 'Boosting propagule transport models with individual-specific data from mobile apps'. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.6m905qg3j

Fischer, S. M., Ramazi, P., Simmons, S., Poesch, M. S., & Lewis, M. A. (2023). IndSpecVeMo—An individual-specific vector model (0.1). *Zenodo*. https://doi.org/10.5281/zenodo.7499230

Gates, K. K. (2007). *Myxospore detection in soil and angler movement in southwestern Montana: Implications for whirling disease transport* (Ph.D. thesis). Montana State University-Bozeman, College of Letters & Science.

Ghosh, J., & Samanta, T. (2001). Model selection—An overview. *Current Science*, *80*, 1135.

Hofer, B. (1903). Über die drehkrankheit der regenbogenforelle. *Allgemeine Fischerei-Zeitung*, *28*, 7–8.

Hulme, P. E. (2009). Trade, transport and trouble: Managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, *46*, 10–18. https://doi.org/10.1111/j.1365-2664.2008.01600.x

Husak, G. J., Michaelsen, J., & Funk, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, *27*, 935–944. https://doi.org/10.1002/joc.1441

Johnson, L. E., Ricciardi, A., & Carlton, J. T. (2001). Overland dispersal of aquatic invasive species: A risk assessment of transient recreational boating. *Ecological Applications*, *11*, 1789–1799. https://doi.org/10.1890/1051-0761(2001)011[1789:ODOAIS]2.0.CO;2

Johnston, F. D., Simmons, S., van Poorten, B. T., & Venturelli, P. A. (2021). Comparative analyses with conventional surveys reveal the potential for an angler app to contribute to recreational fisheries monitoring. *Canadian Journal of Fisheries and Aquatic Sciences*, *79*, cjfas-2021-0026. https://doi.org/10.1139/cjfas-2021-0026

Jones, E., Oliphant, T., & Peterson, P. (2001). *SciPy: Open source scientific tools for Python*. https://scipy.org/

Karesh, W. B., Cook, R. A., Bennett, E. L., & Newcomb, J. (2005). Wildlife trade and global disease emergence. *Emerging Infectious Diseases*, *11*, 1000–1002. https://doi.org/10.3201/eid1107.050194

Kilian, J. V., Klauda, R. J., Widman, S., Kashiwagi, M., Bourquin, R., Weglein, S., & Schuster, J. (2012). An assessment of a bait industry and angler behavior as a vector of invasive species. *Biological Invasions*, *14*, 1469–1481.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Wiley series in probability and statistics. Wiley.

Koch, F. H., Yemshanov, D., Magarey, R. D., & Smith, W. D. (2012). Dispersal of invasive forest insects via recreational firewood: A quantitative analysis. *Journal of Economic Entomology*, *105*, 438–450. https://doi.org/10.1603/EC11270

Kraft, D. (1988). *A software package for sequential quadratic programming*. Technical Report DFVLR-FB 88-28, DLR German Aerospace Center—Institute for Flight Mechanics.

Lee, A. (2010). Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 477–486. https://doi.org/10.1002/wics.98

Leung, B., Drake, J. M., & Lodge, D. M. (2004). Predicting invasions: Propagule pressure and the gravity of Allee effects. *Ecology*, *85*, 1651–1660. https://doi.org/10.1890/02-0571

Leung, B., & Mandrak, N. E. (2007). The risk of establishment of aquatic invasive species: Joining invasibility and propagule pressure. *Proceedings of the Royal Society B: Biological Sciences*, *274*, 2603–2609.

Lewis, M., Petrovskii, S. V., & Potts, J. R. (2016). *The mathematics behind biological invasions. Number 44 in interdisciplinary applied mathematics*. Springer OCLC: 957633921.

Li, X., Tian, H., Lai, D., & Zhang, Z. (2011). Validation of the gravity model in predicting the global spread of influenza. *International Journal of Environmental Research and Public Health*, *8*, 3134–3143. https://doi.org/10.3390/ijerph8083134

Litvak, M. K., & Mandrak, N. E. (1993). Ecology of freshwater baitfish use in Canada and the United States. *Fisheries*, *18*, 6–13.

Muirhead, J. R., Leung, B., Overdijk, C., Kelly, D. W., Nandakumar, K., Marchant, K. R., & MacIsaac, H. J. (2006). Modelling local and long-distance dispersal of invasive emerald ash borer Agrilus planipennis (coleoptera) in North America. *Diversity and Distributions*, *12*, 71–79. https://doi.org/10.1111/j.1366-9516.2006.00218.x

Muirhead, J. R., Lewis, M. A., & MacIsaac, H. J. (2011). Prediction and error in multi-stage models for spread of aquatic non-indigenous species: Prediction and error in multi-stage models. *Diversity and Distributions*, *17*, 323–337. https://doi.org/10.1111/j.1472-4642.2011.00745.x

Muirhead, J. R., & MacIsaac, H. J. (2011). Evaluation of stochastic gravity model selection for use in estimating nonindigenous species dispersal and establishment. *Biological Invasions*, *13*, 2445–2458. https://doi.org/10.1007/s10530-011-0070-3

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692. https://doi.org/10.1093/biomet/78.3.691

Nalepa, T. F., & Schloesser, D. W. (2013). *Quagga and zebra mussels: Biology, impacts, and control*. CRC Press.

Nocedal, J., & Wright, S. J. (2006). Trust-region methods. In T. V. Mikosch, S. I. Resnick, & S. M. Robinson (Eds.), *Numerical optimization* (2nd ed., pp. 66–100). Springer.

Papenfuss, J. T., Phelps, N., Fulton, D., & Venturelli, P. A. (2015). Smartphones reveal angler behavior: A case study of a popular mobile fishing application in Alberta, Canada. *Fisheries*, *40*, 318–327. https://doi.org/10.1080/03632415.2015.1049693

Pluess, T., Cannon, R., Jarošík, V., Pergl, J., Pyšek, P., & Bacher, S. (2012). When are eradication campaigns successful? A test of common assumptions. *Biological Invasions*, *14*, 1365–1378.

Potapov, A., Muirhead, J., Yan, N., Lele, S., & Lewis, M. (2011). Models of lake invasibility by Bythotrephes longimanus, a non-indigenous zooplankton. *Biological Invasions*, *13*, 2459–2476. https://doi.org/10.1007/s10530-011-0075-y

Potapov, A., Muirhead, J. R., Lele, S. R., & Lewis, M. A. (2010). Stochastic gravity models for modeling lake invasions. *Ecological Modelling*, *222*, 964–972. https://doi.org/10.1016/j.ecolmodel.2010.07.024

Prasad, A. M., Iverson, L. R., Peters, M. P., Bossenbroek, J. M., Matthews, S. N., Davis Sydnor, T., & Schwartz, M. W. (2010). Modeling the invasive emerald ash borer risk of spread using a spatially explicit cellular model. *Landscape Ecology*, *25*, 353–369. https://doi.org/10.1007/s10980-009-9434-9

Ramazi, P., Fischer, S. M., Alexander, J., James, C., Paul, A. J., Greiner, R., & Lewis, M. A. (2021). *M. cerebralis* establishment and spread: A graphical synthesis. *Canadian Journal of Fisheries and Aquatic Sciences*, *79*, cjfas-2020-0352. https://doi.org/10.1139/cjfas-2020-0352

Ramazi, P., Kunegel-Lion, M., Greiner, R., & Lewis, M. A. (2021). Predicting insect outbreaks using machine learning: A mountain pine beetle case study. *Ecology and Evolution*, *11*, 13014–13028. https://doi.org/10.1002/ece3.7921

Storn, R., & Price, K. (1997). Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*, 341–359. https://doi.org/10.1023/A:1008202821328

Turner, K. G., Smith, M. J., & Ridenhour, B. J. (2014). Whirling disease dynamics: An analysis of intervention strategies. *Preventive Veterinary Medicine*, *113*, 457–468.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*, 1–28. https://doi.org/10.1007/s10182-008-0060-7

Venturelli, P. A., Hyder, K., & Skov, C. (2017). Angler apps as a source of recreational fisheries data: Opportunities, challenges and proposed standards. *Fish and Fisheries*, *18*, 578–595. https://doi.org/10.1111/faf.12189

Villa, E. R., & Escobar, L. A. (2006). Using moment generating functions to derive mixture distributions. *The American Statistician*, *60*, 75–80. https://doi.org/10.1198/000313006X90819

Von der Lippe, M., & Kowarik, I. (2007). Long-distance dispersal of plants by vehicles as a driver of plant invasions. *Conservation Biology*, *21*, 986–996. https://doi.org/10.1111/j.1523-1739.2007.00722.x

Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J. (2019). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*, *105*, 183–202.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fischer, S. M., Ramazi, P., Simmons, S., Poesch, M. S., & Lewis, M. A. (2023). Boosting propagule transport models with individual-specific data from mobile apps. *Journal of Applied Ecology*, *00*, 1–16. https://doi.org/10.1111/1365-2664.14356