Models of lake invasibility by *Bythotrephes longimanus*, a non-indigenous zooplankton

Alex Potapov[a,b,e], Jim Muirhead[a,c], Norman Yan[d], Subhash Lele[a,b], Mark Lewis[a,b,c]

[a]Centre for Mathematical Biology,
[b]Department of Mathematical and Statistical Sciences,
[c]Department of Biological Sciences,
University of Alberta, Edmonton, AB, T6G 2G1 Canada
[d]Department of Biology, York University, 4700 Keele Street,
Toronto, ON M3J 1P3; and Dorset Environmental Science
Centre, Dorset, ON P0A 1E0, Canada
[e]e-mail: apotapov@math.ualberta.ca

Abstract

We built a family of hierarchical risk models for the spread of invasions by the spiny waterflea (*Bythotrephes longimanus*) in lakes in Ontario, Canada. Knowledge of covariates determining lake invasibility and ability to predict risk of future invasions may help to develop management policy and slow the invasions in the future. The models are based on two component submodels. The first component was a stochastic gravity submodel for the propagule pressure between lakes via recreational boaters. The second component was a submodel for establishment risk, given that the invader has already been introduced to a lake. This component was a logistic regression model, incorporating up to 17 measured covariates that describe the physical and chemical condition of the lake. Variants of the risk model, each incorporating different subsets of the covariates, were calibrated using presence/absence data from a 300-lake survey conducted in 2005–2006 by the Canadian Aquatic Invasive Species Network (CAISN). The predictive capacity of the best model was high, giving AUC values close to 0.94. Of the model covariates considered, the most important predictors of existing invasions were propagule pressure and lake pH, and, to lesser extents, phosphorus (P) and lake elevation. Our

fitting of the propagule pressure submodel demonstrated a significant Allee effect for *Bythotrephes*. Our development of the establishment risk predictor showed that it is essential to account for temporal variability in lake physico-chemistry. We demonstrated that invasions of lake networks by the spiny waterflea follow highly predictable patterns which can be understood with a properly calibrated, hierarchical risk model.

Key words: Aquatic invasions, risk model, invasion predictions, statistical model selection, habitat suitability

Introduction

Biological invasions are increasing, mainly due to the rapid growth of international trade and transport (Lockwood et al. 2005). Invaders can cause biodiversity losses (Sala et al. 2000, Gurevitch and Padilla 2004), with ecological and economic harm measured in millions or even billions of dollars (Pimentel et al. 2005). Management efforts to reduce the spread of invaders to protect native habitats are thus clearly justified, but such management is best designed using scientifically grounded understanding, tools and/or arguments with a proven history of efficacy. Such arguments and efforts require the predictive modeling of invasion risk built on the analysis of actual invasion cases, so that reasons and ecological mechanisms for invasion success can be understood and appropriately modelled (Clark et al. 2001).

One approach to predicting invasions is to determine the types of habitat that are suitable for the invader. This is similar to determining species' niches from records of current distribution and abundance or presence/absence data (e.g. Pulliam 2000; Guisan and Thuiller 2005; Stockwell 2007). Although this approach has frequently been used to model invasible habitat based on the occurrences of the invader and its relationship with the environment in its native habitat, it has also been used successfully to model potential range expansion of an invader given current presence/absence data and habitat characteristics within the invaded range (e.g. Drake and Bossenbroek (2004)).

However, it matters not how invasible a new habitat is, if invader propagules do not reach it. Therefore, another important component of the invasion process is propagule pressure, characterizing the number of introduced individuals or the rate of their introduction (e.g. Drake et al. 2005). Propagule flow to a certain location can be estimated with the help of gravity models, which have been successful in explaining past and predicting future introductions (Bossenbroek et al. 2001; MacIsaac et al. 2004). Because both propagule pressure and invasibility are clearly implicated in the invasion process, they can be combined into one risk model, e.g. Leung and Mandrak (2007), effectively describing both introduction and establishment steps of invasion. This approach is both intellectually satisfying, and appears to provide the most promise in prediction of the invasion risk, though it does require both more data and more complicated models.

The spiny waterflea, *Bythotrephes longimanus* (Crustacea, Onychopoda, Cercopagidae) is a good candidate for building such an integrated, invasion risk model. It was first discovered in North America in the lower Laurentian Great Lakes in the early 1980s, likely transported within ballast water by ships from the Baltic Sea. Within a decade, it had occupied all the Great Lakes and started to spread to inland lakes (see review of *Bythotrephes* invasion history and corresponding references in Branstrator et al (2006)). *Bythotrephes* is a zooplankton predator, and it has dramatically decreased native zooplankton biodiversity (Yan et al. 2002; Barbiero and Tuchman 2004), but its impacts may depend on characteristics of the native zooplankton community (Strecker and Arnott 2005). Its spread has been modeled among the thousands of lakes in the Great Lakes region (MacIsaac et al. 2004; Muirhead and MacIsaac 2005; Branstrator et al. 2006; Muirhead 2007), where it appears that human-mediated traffic of recreational boats is a major mechanism of new and ongoing introductions (Weisz and Yan 2010).

Not all lakes are suitable for *Bythotrephes* establishment, however. The invader preferentially colonizes large, deep, and nutrient poor lakes (MacIsaac et al. 2000; Weisz and Yan 2010), a lake type that is common on the Canadian Shield. MacIsaac et al. (2000) developed a model of lake invasibility by *Bythotrephes* based on European lake

data. Although the model was reasonably successful in predicting invasions for lakes in which the species has been recorded, it also generated a high rate of predicted occurrences where *Bythotrephes* was not observed (i.e. false alarms). One possibility of the mismatch between model predictions and observed invasion status for Canadian Shield lakes is differences in processes governing lake water clarity in Canada, where water colour is the main determinant, vs. Europe, where phytoplankton biomass is key (Cairns et al. 2007). Alternatively, mismatches between model predictions and observed invasions may be due to the stage to which the invasion has progressed. *Bythotrephes* invasions are still ongoing in Canada (Cairns et al. 2007), given the recent colonization history. Thus it is necessary to distinguish between two situations: (a) when the invasion is in its initial stages, and (b) when the invasion is well advanced and the invader has occupied almost all suitable locations. In the latter case, when species absence is related mainly to unfavourable conditions, the relationship between species establishment and habitat suitability is relatively static and matches the standard niche determining concepts. In the former situation, this relationship is dynamic and more complicated to model: some lakes are not invaded due to their unsuitability, while the others are suitable, and likely have invasions in their future. They are currently protected from invasion only by dispersal limitation. These alternatives complicate the determination of favourable conditions for the invader, and necessitate distinguishing "true absence" from "temporary absence". Only presence data are certain if there is insufficient sampling effort or insufficient time for the observed process to develop (Lele and Keim 2006; Pearce and Boyce 2006).

To achieve a better understanding of *Bythotrephes* spread and establishment on the south-central Canadian Shield, the Canadian Aquatic Invasive Species Network (CAISN) conducted a project to collect *Bythotrephes* presence/absence data with sufficient sampling effort to confirm true current presence and absence and to collect simultaneous lake physical and chemical characteristics in order to determine the best predictors of invasion risk (Cairns et al. 2006). In 2007, a database containing the results of sampling of about 300 lakes ("300 lakes database") was compiled (Cairns et al. 2007) and provided to the modeling teams within CAISN. Here our main purpose is to develop a statistical

model for lake invasibility by *Bythotrephes* based upon the 300 lakes data. We needed a family of models since we didn't know which of the measured covariates determined lakes invasibility.

Methods

Data on Muskoka watershed invasion by Bythotrephes

300 lakes database

In 2005 and 2006, 311 lakes were sampled from June to August (as described by Cairns et al. 2006, 2007) during peak *Bythotrephes* population size (Yan and Pawson 1998). The lakes were located in the Muskoka watershed 2EB (Cox 1978) which has the longest *Bythotrephes* invasion history in North America outside of the Great Lakes (Yan et al. 1992). This region also has been identified as a hotspot for new invasions because of substantial recreational boater traffic (MacIsaac et al. 2004). Field crews collected plankton samples for *Bythotrephes* presence/absence, defined physical characteristics of the lakes, and collected water samples for subsequent chemical analysis at the Ontario Ministry of the Environment's Dorset Environmental Science Centre chemistry lab (Dorset, Ontario). All plankton samples (6/lake) were examined in their entirety for *Bythotrephes*. Later the survey data were organized into a "300 lakes database", which was provided to the modelling teams within the CAISN project. Details of survey design and lake sampling techniques can be found in Cairns et al. (2006), Cairns et al. (2007), and Weisz and Yan (2010).

Here we used n=306 lakes from the database, for which values of 17 lake covariates listed in Table 1 were available. In formulae we refer to the covariates and lakes by numbers: Cv_k for k-th covariate, and Cv_{ki} for the measurement of k-th covariate in lake i.

To improve maximum likelihood convergence and interpretation of the results of regression models, we normalized the data. For each Cv_k we calculated its mean μ_k and its "spatial" standard deviation across lakes σ_{Sk}. In actual model fitting we used the normalized values:

$$x_{ki} = \frac{Cv_{ki} - \mu_k}{\sigma_{Sk}} \quad (1)$$

The values of μ_k and σ_{Sk} are presented in Table 1.

Data for determining temporal variability of covariates

For 42 of the 306 lakes, the covariates were measured twice, in both 2005 and 2006. The measurements clearly show that chemical covariates vary with time. Information about their variability is important for invasion predictions, since such a predictor must give a reasonable answer for any measurement within a given range. To characterize variability of chemical covariates, we used records of covariates for 8 lakes in the same region as our spatial survey that were shared with us by A. Paterson of the Ontario Ministry of Environment (see Yan et al. 2008 for an overview of limnological changes in these 8 lakes). Most records were taken 1-2 times a month, mainly from the end of April to the beginning of November, and we used 1989 to 2007 data, when *Bythotrephes* were introduced and spreading in the region. For each covariate, first we estimated the variance at each lake, and then averaged the variances with the weights proportional to the number of data points at each lake. The square root of the averaged variance gives the standard deviation for temporal variability σ_k of each covariate.

Propagule pressure estimates

The direct measurement of *Bythotrephes* propagule pressure is infeasible, but models for propagule pressure may be developed based on movement patterns of a well-known vector: recreational boating. Now common tools for predicting boater's behaviour are gravity models (Thomas and Hugget 1980; Keller et al. 2009), which have been

developed to model migration and flows of economic trade (e.g. Zipf 1946; Linneman 1966) Here we used a stochastic form of gravity model for Ontario lakes developed in Potapov et al. (2010). The model predicts the mean boater traffic between lakes *i* and *j* as a Poisson process with the mean intensity

$$\lambda_{ij} = C_\lambda \left(A_{Ni} A_{Nj}\right)^\gamma d_{ij}^{-1} \left(b_i + b_j\right)^\alpha, \quad \alpha = 1.37, \ \gamma = 0.58, \tag{2}$$

where

$$A_{Ni} = \frac{A_i}{1 + A_i / A_0} = \frac{A_0 A_i}{A_i + A_0}, \qquad A_0 = 3200\,\mathrm{km}^2 \tag{3}$$

is the normalized lake area, characterizing its attractivity for boaters; d_{ij} is the distance between the lakes; and b_i characterizes the number of boaters visiting lake *i*. We assume that the number of boaters in a geographic area is proportional to the total population in this area, Pop_k, and their willingness to travel to the lake is inversely proportional to the distance from the area to the lake l_{ik}. For these inhabited areas we take regions corresponding to the first two digits of a postal code (FSA2), and

$$b_i = C_b \sum_{\substack{k \text{ over all} \\ \text{FSA2 areas}}} \mathrm{Pop}_k \times l_{ki}^{-1} . \tag{4}$$

C_b is a normalization constant.

To predict absolute mean boater flow, C_λ in (2) must be proportional to the total number of trips in the lake system or the total number of boaters. These numbers may be known only approximately or may be unknown. In spite of this, (2) can successfully predict relative boater traffic, showing ratios of the flows for different lakes. In this case we can choose C_λ arbitrarily. In this work, we have normalized λ such that the maximum estimate corresponds to λ_i=1, which is equivalent to choosing C_λ=15.23. To obtain the absolute boater flow λ_A, we need a conversion factor *C* such that $\lambda_A = C \times \lambda$.

Let us assume that the number of invader individuals transported by each boat is a binomially distributed random number with the mean n_I. Then the flow of invader individuals into a lake is a random variable:

$$Y \sim Poisson(\mu), \qquad \mu = C_\mu \lambda, \qquad C_\mu = C \times n_I .$$

The coefficient C_μ is unknown, but can be estimated from the data.

Single covariate predictors and current invasion stage

Logistic regression based on presence/absence data is a standard approach to risk modeling when there is no additional information about the processes to be modeled. We tried it with only one covariate, that is we compare 18 models for probability of lake *i* invasion of the form

$$P_{INV,i} = S(a_0 + a_1 x_i), \qquad S(u) = (1 + \exp(-u))^{-1}, \tag{5}$$

where $S(u)$ is logistic function and x_i the value of one of the 18 covariates listed in Table 1 related to lake *i*. We fit these models to the data with the help of R function `glm` (R 2009; Crowley 2007). The models were ranked according to AIC value.

This step may give information about current invasion stage. In the initial stage of invasion there are many suitable lakes, but only a few are accessed by the invader, and the most important invasion predictor must be propagule pressure for the lake. On the other hand, if sufficient time has elapsed for the potential spread of the invader, most of uninvaded ones are not suitable for the invader. Thus, the most important invasion predictors should be covariates related with habitat type, most probably related with water chemistry.

The results are presented below, and propagule pressure appeared to be the best predicting covariate. Therefore, the most accurate invasion risk model must account for two invasion stages: introduction and establishment. We implemented this approach in a hierarchical risk model.

Hierarchical risk model and model selection

Our hierarchical risk model couples a stochastic model for population introduction, which is based on propagule pressure, with an establishment risk model, based on the local environmental conditions. This hierarchical risk model can be expanded into a family of such when only subsets of local environmental conditions are included in the establishment risk submodel. In the last part of this section we develop methods for assessing which out of the family of models best reflects biological reality. From this selection process we can deduce which environmental conditions have a significant impact on *Bythotrephes* establishment.

Modeling techniques used for the establishment risk model are determined by our goals: 1) to obtain risk estimates for *Bythotrephes* invasion into lakes and 2) to determine covariates important for lake invasibility diagnostics. Since all covariates are numbered in Table 1, such a subset can be represented as a collection **K** of covariate indices. For example, if we use elevation (Elev), phosphorus (P_1) and pH, then **K**=(3,10,16). If we need to denote that index *k* takes all values from such a subset, we shall write $k \in \mathbf{K}$.

Invasion as a random process

We introduce a random variable X where $X = 0$ corresponds to uninvaded lake and $X = 1$ to invaded one. Also, let us denote:

$P(Y = j \mid \mu)$ — probability that *j* invader individuals arrive at a lake given propagule pressure μ;

$P(X = 1 \mid Y = j)$ — probability that the invader establishes given that $Y = j$ individuals arrive.

The probability of the lake invasion is then:

$$P(X = 1) = \sum_{j=0}^{\infty} P(X = 1 \mid Y = j) \times P(Y = j \mid \mu). \qquad (6)$$

The probability that an invader establishes given a level of propagule pressure $P(X=1|Y=j)$ is often approximated by a binomial distribution with the independent separate establishment of each individual in many invasion models (Jerde and Lewis 2007; Jerde et al. 2009), but this approach does not seem appropriate for *Bythotrephes*. During the summer, *Bythotrephes* has several parthenogenetic reproduction cycles, and by the end of summer the introduced individuals create a small population. Then through sexual reproduction, resting eggs are produced, which is the only life stage that has been demonstrated to survive the winter in North America. Establishment depends on how many resting eggs are produced, which in turn depends on the size of the end-of-summer population. If the sexually reproducing population is too small, the invader may go extinct due to Allee effects (Wittmann et al. this issue). The presence of Allee effects implies a threshold in the terminal population size.

If the lake is suitable for *Bythotrephes*, 3-6 cycles of parthenogenetic reproduction with 2-4 eggs at each stage may result in an "avalanche" that increases the number of individuals ~100 times and more, and the initial number of individuals needs only to be enough to start the avalanche. If the lake is not suitable or less suitable, then the avalanche does not arise and the initial number of individuals may not matter, unless hundreds are introduced, which we consider improbable for recreational boaters. Therefore, $P(X=1|Y=j)$ is actually a probability of a reproductive avalanche for a certain lake, given j individuals are introduced.

The stage of parthenogenetic reproduction can be modeled by a branching process, and to the sexually reproduction stage the model from Jerde et al. (2009) can be applied, but such a model would be quite complicated. To simplify the approach, we assume that there is a threshold number of invader individuals m such that:

- for $j<m$ $\quad P(X=1|Y=j)\approx 0$ (the invader does not survive if fewer than m arrive);
- for $j>m$ $\quad P(X=1|Y=j)\approx P(X=1|Y=m)$ (the further increase of j individuals practically does not increase the probability of establishment. If the lake is suitable, m is enough for establishment; if it is not suitable, additional arrivals do not help).

Then the probability of establishment may be approximated by:

$$P(X=1) \approx \sum_{j=m}^{\infty} P(X=1 \mid Y=m) \times P(Y=j \mid \mu) = \\ = P(X=1 \mid Y=m) \times Q(\mu, m) \tag{7}$$

where

$$Q(\mu, m) = \sum_{j=m}^{\infty} P(j \mid \mu) = \sum_{j=m}^{\infty} P(j \mid C_{\mu}\lambda)$$

It can be shown (see Appendix A), that $Q(\mu, m)$ can be approximated by a simpler expression:

$$Q \approx 1 - \exp(-\kappa\lambda^m), \qquad \kappa = C_{\mu} / m!. \tag{8}$$

Eventually we obtain the relation:

$$P(X=1) \approx \left[1 - \exp(-\kappa\lambda^m)\right] \times P(X=1 \mid Y=m), \tag{9}$$

where the last term depends only on the lake covariates. This expression has the same form as habitat suitability models, where the first factor describes invader flow and the second one describes lake suitability.

Model (9) has some similarity with the hierarchical model in Leung and Mandrak (2007) for zebra mussels, where the invasion probability equals to a product of propagule pressure term and lake suitability term. However, their derivation is different. The term (8) has been used in Leung et al. (2004) to account for Allee effect, but without derivation.

Model of lake suitability

We assume that the establishment probability $P(X=1 \mid Y=m)$ depends on the covariates that have been measured for each lake. As mentioned above, these covariates vary with time and we treat them as normally distributed random variables $\xi_k \sim \mathrm{n}(x_k, \sigma_k^2)$ (Sect. 2).

If the covariates were constant, a reasonable approximation may be the logistic regression:

$$P(X=1|Y=m)=S\left(a_0+\sum_{k\in\mathbf{K}}a_k x_k\right).$$

A simple substitution of ξ_k instead of x_k would give "instant lake suitability", while the invader establishment takes time and should depend on averaged characteristics. For this reason, we use the approximation:

$$P(X=1|Y=m)=\left\langle S\left(a_0+\sum_{k\in\mathbf{K}}a_k\xi_k\right)\right\rangle_{\xi} \qquad (10)$$

Since the sum in the internal brackets is also distributed normally, it may be simplified to:

$$v=a_0+\sum_{k\in\mathbf{K}}a_k\xi_k \sim \mathrm{n}\left(a_0+\sum_{k\in\mathbf{K}}a_k x_k,\sum_{k\in\mathbf{K}}a_k^2\sigma_k^2\right), \qquad (11)$$

or

$$\nu=a_0+\sum_{k\in\mathbf{K}}a_k x_k+\xi_0\sqrt{\sum_{k\in\mathbf{K}}a_k^2\sigma_k^2}, \qquad \xi_0\sim\mathrm{n}(0,1).$$

Therefore, averaging in (10) need to be done over only one Gaussian random variable ξ_0. Exact averaging requires evaluation of an integral over ξ_0, but we approximate it by a finite sum (see Appendix B). Let ξ_{0j} are n_0 points such that probabilities $\mathrm{Prob}(\xi_0<\xi_{0j})=(j-0.5)/n_0$, j=1,…,n_0, that make a uniform grid in [0,1]. If n_0 is big enough, then

$$P(X=1|Y=m)\approx\frac{1}{n_0}\sum_{j=1}^{n_0}S\left(a_0+\sum_{k\in\mathbf{K}}a_k x_k+\xi_{0j}\sqrt{\sum_{k\in\mathbf{K}}a_k^2\sigma_k^2}\right) \qquad (12)$$

Hierarchical invasion risk model

Combining (9) and (12), we come to the full invasion risk model

$$P(X=1)=\left[1-\exp\left(-\kappa\lambda^m\right)\right]\times\left[\frac{1}{n_0}\sum_{j=1}^{n_0}S\left(a_0+\sum_{k\in\mathbf{K}}a_k x_k+\xi_{0j}\sqrt{\sum_{k\in\mathbf{K}}a_k^2\sigma_k^2}\right)\right]. \qquad (13)$$

There are no standard functions for fitting such models, so we calculated log-likelihood

$$L(\kappa, m, \{a_k\}) = \sum_{i=1}^{n} \left(\ln\left[1 - \exp\left(-\kappa\lambda_i^m\right)\right] + \right.$$
$$\left. + \ln\left[\sum_{j=1}^{n_0} S\left(a_0 + \sum_{k\in\mathbf{K}} a_k x_{ik} + \xi_{0j} \sqrt{\sum_{k\in\mathbf{K}} a_k^2 \sigma_k^2} \right) \right] - \ln n_0 \right) \qquad (14)$$

and maximized it with respect to $\kappa, m, \{a_k\}$ using R internal function `vmmin`, a part of `optim` routine, corresponding to BFGS method (R 2009). To obtain a good initial guess for gradient descend maximization in `vmmin`, we did 5000 steps of non-annealing "hide and seek" random maximization algorithm (Romeijn and Smith 1994; Spall 2003; Potapov 2009). To increase the speed of computations, the function implementing calculations of log-likelihood, random maximization and calls to `vmmin` has been written in C++ as an R extension.

Model selection

We have tested models of various complexities containing different subsets of covariates. The simplest models contained only gravity scores λ, more complicated used λ and one lake covariate, λ and two covariates, and so on. To compare performance of different models and to select an optimal set of covariates, we apply methods of statistical model selection.

We used information-based criteria, associating the best model with the minimum value of AIC or BIC (Burnham and Anderson 2001, 2004; Ghosh and Samanta 2001). Interpretation of their results has been considered in a number of publications. Most important for us are two conclusions.

1. Sometimes AIC and BIC give different conclusions according to which model should be considered as the best. According to (Burnham and Anderson 2004), BIC gives better results, when the true model is simple and has 1-4 covariates, AIC is better when the true model is complicated. Ghosh and Samanta (2001) put it slightly different: AIC is better when we want to build a better predictor, BIC is better when we need to "discover the

truth", that is to find a model related to most important underlying mechanisms. Therefore, in our case the primary model selection tool must be BIC.

2. Burnham and Anderson (2001) consider interpretation of model comparison by means of AIC. They conclude that the difference of AIC for two models should be big enough to decide that one is definitely better than the other. If $\Delta AIC \leq 2$, the two models should be considered as similar, and only for $\Delta AIC > 8$ one of the models is almost certainly better. Since both AIC and BIC are both based upon calculation of log-likelihood, one should expect that similar approach should be in case of BIC as well.

To compare the predictive ability of the models we compare their values of AUC or area under ROC curve, (e.g. Pepe 2003). In our case, typically the models having the least BIC values have the greatest AUC value as well.

Results

Current invasion stage

The best single covariate for predicting the invasion status of lakes was propagule pressure λ with AIC=151 (Fig. 2). The next best was pH with AIC=164, and AIC difference between these models was 13, which means that λ was a much better predictor. Therefore, the invasion status of a lake was best explained by the chances for invader introduction. Water chemistry parameters were of secondary importance to establishment, significant only for lakes with big enough λ. This is an argument in favour of using habitat suitability model (13), that takes into account the current invasion dynamics, and considers invasion as a process with two independent stages: introduction and establishment (see Leung and Mandrak 2007 as well).

Chemistry data variability

The estimates of variances and standard deviations σ_k (Table 1) are important not by themselves, but only in comparison with variability of the same covariates between the lakes or spatial variability σ_{Sk}, that may allow us to predict lake invasibility. If both types of variability are close, it is impossible to say what is the reason for the difference in covariate values between the lakes: temporal variability or a significant difference in the lake types. For a covariate to be a potentially valuable invasion predictor, the condition $\sigma_k / \sigma_{Sk} << 1$ must hold. The relative variability in covariates σ_k / σ_{Sk} is shown in Fig. 3. The highest ratio in variability is found for Secchi depth (close to 0.6) with most of the ratios for other covariates below 0.2. Thus, spatial variability for them is much more important than temporal.

Results for Hierarchical model (13)

To determine the optimal value of parameter n_0 in (13), we did model fitting with all possible combinations of 1 to 4 lake covariates for n_0=1, 3, 5, 11, 25, 51, 99. Models were characterized by AIC, BIC, and AUC values. As we increase n_0, the mean deviation of the results from those for n_0=99 stabilize (Fig. 4), and for $n_0 \geq 51$ there is practically no difference.

The second effect related with n_0 was model overfitting. Models of logistic regression in presence of measurement errors may be biased (Stefanski and Carroll 1985; Carroll et al. 1995). When the data by chance appear to be separated better than they should be according to their distribution, the fitted maximum likelihood model has a threshold point, such that below the threshold it predicts risk close to 0, and above the threshold close to 1. This means that the risk is underestimated below the threshold and overestimated above it. This effect can be called overfitting since the model interprets a random feature as a significant one. In our case, overfitting is caused by temporal variability rather than measurement error, the latter being much less. Overfitting is

reflected in the values of logistic regression coefficients. Typically for a regression model with normalized covariates $|a_k|$~1, and in case of overfitting the values may be ~100 and more. In test calculations with n_0=1 and 6-7 covariates, we have observed values a_k even exceeding 1000. The hierarchical model accounting for variability of the covariates automatically corrects for occasional data separation and often eliminates overfitting. The values of max$|a_k|$ averaged over 20 AIC-best and BIC-best models decrease with n_0 (Fig. 6) and saturate for $n_0 \geq 51$, in agreement with Fig 5. For this reason, below the results are presented for for n_0=51.

Model selection for the models including n_c=0, 1, 2, 3 and 4 lake covariates gives different conclusions for AIC and BIC criteria, see Table 2. According to BIC, the best was the model with n_c=1 using propagule pressure λ and lake pH. According to AIC, the best was the model with n_c=3 using λ, pH, phosphorus and elevation. One model with n_c=4 has AIC slightly smaller than models with n_c=3, but its coefficients show signs of overfitting: the model actually uses the difference of two measures of alkalinity, that is mainly random series, to improve likelihood minimization. Therefore, this model has to be discarded, and other models with n_c=4 are not better than models with n_c=3.

Model selection also shows evidence of a certain structure in combinations of covariates within 21 best models with n_c=3 (see Table 3). The models naturally split into two groups. Five or six models (#7, 14, 17, 19, 20, and perhaps 15) demonstrate signs of overfitting: coefficients for pH are 3-5 times more than that for the other models. The rest of the models demonstrate a clear pattern. All of them have the structure:

λ + pH + {P or DOC} + {one of: Elev, Ca, Alk, Alki, K, CD, Na, Mg}. (15)

All covariates in each group enclosed in braces are strongly correlated with each other, (see Table 4) and hence must be related with the same effect.

AIC and AUC for all models tested are well correlated, so the model with the lowest AIC has the highest AUC (Fig. 6). BIC splits the results into four bands, according to the number of covariates used, and the model with the best BIC does not have the highest AUC.

Estimates of *m* for all tried models are grouped into three clusters (Fig. 7). The best models according to both AIC and BIC criteria form a cluster of estimates around 2.2.

The best invasion predictors

We can conclude that there are two candidates for the "best" predictor:

a) using λ and pH, AIC=121.8, BIC=133.6, AUC=0.914,

$$P(X=1)=\left[1-\exp\left(-13.0\cdot\lambda^{2.23}\right)\right]\times S\left(5.50\,\frac{\mathrm{pH}-6.18}{0.57}-2.12\right)= \\ =\left[1-\exp\left(-13.0\cdot\lambda^{2.23}\right)\right]\times S\left(9.65\cdot(\mathrm{pH}-6.40)\right) \qquad (16)$$

$$\kappa=13.0\pm0.8,\quad m=2.23\pm0.12,\quad a_0=-2.12\pm0.23,\quad a_{\mathrm{pH}}=5.50\pm0.90$$

(see Sect. 2 and Table 1 for covariates normalizing).

b) using λ, pH, P_1/P_2, and Elev, AIC=115.8/117.1, BIC=134.4/135.7, AUC=0.935/0.933. Averaging coefficients for two series of phosphorus, we obtain

$$P(X=1)=\left[1-\exp\left(-13.9\cdot\lambda^{2.15}\right)\right]\times \\ \times S\left(5.92\,\frac{\mathrm{pH}-6.18}{0.57}-1.76\,\frac{\mathrm{Elev}-327.4}{83.0}-5.01\frac{\mathrm{P}-10.23}{8.44}-5.21\right)= \\ =\left[1-\exp\left(-13.9\cdot\lambda^{2.14}\right)\right]\times \\ \times S\left(10.39\cdot\left[\mathrm{pH}-6.68-0.002(\mathrm{Elev}-327.4)-0.057(\mathrm{P}-10.23)\right]\right) \qquad (17)$$

$$\kappa=13.9\pm0.9,\quad m=2.15\pm0.11,\quad a_0=-5.21\pm0.31,$$
$$a_{\mathrm{pH}}=5.92\pm0.43,\quad a_{\mathrm{Elev}}=-1.76\pm0.34,\quad a_{\mathrm{P}}=-5.01\pm0.30,$$

Model (16) is supported by BIC criterion and parsimony. Model (17) is favoured by AIC, and has a higher AUC value. Note that many models of (15) family also have AUC higher than model (16).

Discussion

Composition of the family (15) of the best predictors shows that it accounts for 4 different effects, with the first two effects, related with propagule pressure and pH being much more important than two latter ones. The effect of propagule pressure is obvious and has been discussed above. pH may be a classifier for type of lake based on water inflow, the characteristics of the surrounding watershed or underlying geology.

Phosphorus and dissolved organic carbon may characterise lake productivity. *Bythotrephes* is more likely to be found in oligotrophic lakes (MacIsaac et al 2000; Branstrator et al 2006). This assumption agrees with the negative sign of the coefficients a_k for both P and DOC (Tables 2-3). Secchi depth is strongly anticorrelated with both of them (correlation coefficients are –0.59/–0.63 and –0.67, see Table 4), and could also be a useful predictor. *Bythotrephes* is a visual predator requiring light to hunt, and thus low light (small Secchi depth) may increase its death rates by reducing its ability to find food. Hence, lower Secchi depth should be associated with lower establishment risk, as propagules will die if they can't see to find food (Weisz and Yan 2010). On the Canadian Shield, Secchi depth is controlled by DOC (Pérez-Fuentetaja et al 1999). Unfortunately, Secchi depth has the highest variability among all covariates, and cannot make a reliable predictor (Fig. 3, Table 1).

The last effect is related with big group of covariates in the last braces in (15). Conductivity, alkalinity, and concentration of base cations are well correlated with each other. All of them are correlated with lake elevation as well. Since concentrations of the ions are not very high, it is hard to explain their effect on *Bythotrephes* establishment. We may assume that all these chemistry variables just characterize lake elevation. The latter, in turn, determines order of the lake in a chain of water reservoirs connected by rivers or streams. If an upstream lake becomes invaded, this increases chances of all lakes below it to become invaded as well. Also many of the higher elevation lakes in our case are protected as they are in Algonquin Park, a large provincial park in Ontario, so they have much less motorized boat traffic. In other words, as elevation increases,

chances of a lake to become invaded are less. This agrees with the negative sign of a_k related to elevation (Tables 2-3). If this assumption is true, then it has two consequences.

First, the effect of elevation or alkalinity is an indirect measure of propagule pressure as well, though less explicit. It might be specific of Muskoka watershed (e.g. the presence of Algonquin Park), and be not so important in other regions. Second, it is better to use elevation itself in the predictor, rather than metal ion data correlated with it. This leaves only two predictors with 4 covariates: (λ, pH, P, Elev) and (λ, pH, DOC, Elev). If we decide to drop lake elevation Elev as a not very reliable propagule pressure predictor, then, according to Table 2, accounting only for P or DOC does not provide a significant decrease in AIC or increase in AUC compared to the simplest predictor (16), containing only λ, pH. The latter then appears to be optimal.

If we consider only the establishment part of the model (9), namely equation (12), it considers only properties of the lake without explicit accounting for propagule pressure. Therefore, it describes lake suitability as a potential habitat for *Bythotrephes*. In other words, it provides risk for the lake to be invaded sometimes in the future. We have calculated it for all 306 lakes used for model fitting. It appears that more than 100 of them have an invasibility risk >0.5, and for the year 2006 only 28 of them were invaded. This means that numerous new invasions may be expected in the future. As more lakes with both chemical and P/A data become available, models may improve, and other variables, such as predators may well help improve future models. Clearly this is a work in process.

We expect that our predictors will work in other Ontario lakes, because so much of Ontario is covered by the similar terrain of the Canadian Shield. We also expect our predictors should apply to other lakes with similar temperature regimes and ecosystem structure, and recreational boater activity. *Bythotrephes* survival and growth strongly depend on water temperature (Kim and Yan 2010). Hence, lake temperature may be another important covariate for predictors targeted for wider areas.

If we consider only the propagule pressure-related part of the model (9), namely Eq. (8), it mainly describes invader introduction. However, it also has a part related to establishment, parameter *m*. Leung et al (2004) have pointed out that its value greater than 1 means the presence of Allee effect: a successful establishment requires introduction of a certain minimum number of individuals even for most suitable lake. Our estimate of $m \approx 2.2$ we interpret as an experimental evidence of Allee effect for *Bythotrephes* predicted earlier by Wittmann et al. (2010). The presence of an Allee effect may be an important factor for the rate of *Bythotrephes* spread, and potentially for its controllability (Taylor and Hastings 2005). The reduction of propagule pressure below critical threshold below which Allee effects prevent population expansion (e.g. due to better management of equipment by boaters) may prevent new invasions. The spread of other species, such as zebra mussels, are also influenced by Allee effects (Leung et al 2004).

There is a possibility of further development of our hierarchical model. For the lake suitability component we used logistic regression, while there are other possibilities such as neural networks (Leung and Mandrak 2007) or GARP (Herborg et al. 2007). In our case logistic regression has the following advantages:

1. There are well established techniques for estimating parameters and errors in logistic regression; and, it is computationally efficient, which is very important for comparison of several thousand candidate models.
2. Its results are very easy to report and to reuse, see (16), (17), and Tables 2-3.
3. Compared to more complex techniques (e.g. neural networks), it is much easier to detect the effects of overfitting and, in case of linear function of covariates, to implement corrections for variability in covariates. There are theoretical results for the case of measurement errors.

Use of other approximation techniques for the lake suitability part is possible. However, accounting for covariates' variability would be a much harder computational problem.

Further applications of our risk models may be related with lake management. Knowledge of propagule pressure should allow managers and ecologists to identify lakes

where invasion may be expected, and thus might require additional signage or public outreach to reduce propagule introduction. Knowledge of the pH of these lakes might increase the accuracy of the determination of the risk of invasion.

Our modeling approach may be of interest not only for *Bythotrephes*-related research community, but for developing models for other invaders and to invasion biologists in general. Describing invasion risk at different stages of invasion may require combinations of submodels for propagule pressure, Allee effect and habitat properties. Models of habitat suitability type with explicit accounting for the probabilities of propagule arrival and establishment are useful at early stages of invasion, when only the presence data are certain and will not change in future, and absence may be related with problems of the invader detection or delays with the invader arrival.

Temporal variability in habitat characteristics may produce many modelling problems. Accounting for it appears to be important in three aspects. 1) Covariates with too much variability in time appear to be predictors of limited reliability, like Secchi depth in our case. 2) A single covariate measurement may lead to overestimated or underestimated invasion risk. 3) Predictors which account for variability (hierarchical model) appear to be computationally efficient and stable against the effects of overfitting.

Acknowledgements

This research has been supported by Canadian Aquatic Invasive Species Network, by NSERC and by a Canada Research Chair (ML). We want to thank A. Paterson, Ontario Ministry of Environment, for data on variability of chemical covariates, and A. Cairns of York University for running the sampling program for the 300 lakes, assembling the 300 lake database.

Appendix A

For the function Q in (7) we have the following expression:

$$Q(\mu, m) = \sum_{j=m}^{\infty} P(j \mid \mu) = \sum_{j=m}^{\infty} \frac{\mu^j}{j!} \exp(-\mu) = \frac{\mu^m}{m!} \exp(-\mu) \sum_{j=0}^{\infty} \frac{m! \mu^j}{(j+m)!}$$

For small μ $Q(\mu, m) \approx \mu^m / m!$, for μ big $Q(\mu, m)$ approaches 1. It is convenient to replace the sum by an empirical approximation with qualitatively similar behavior:

$$Q(\mu, m) \approx Q_1(\mu, m) = 1 - \exp\left(-\frac{\mu^m}{m!} \right).$$

Or, in terms of the relative boater flow λ,

$$Q(\lambda, m) \approx 1 - \exp\left(-\frac{(C_\mu \lambda)^m}{m!} \right) = 1 - \exp\left(-\kappa \lambda^m \right), \qquad \kappa = \frac{C_\mu^m}{m!},$$

where κ should be fitted from data.

We have tried another possible approximation to Q,

$$Q_2(\mu, m) = \left[1 - \exp\left(-\mu / (m!)^{1/m} \right) \right]^m.$$

It have shown slightly worse model performance, though formally it provides closer approximation to $Q(\mu, m)$. Both approximations are compared in Fig. 9.

Appendix B

With the help of (11), formula (10) can be written as

$$P(X=1|Y=m)=\int_{-\infty}^{\infty} S(v(\xi_0))\exp\left(-\frac{\xi_0{}^2}{2\sigma_0{}^2}\right)d\xi_0\,, \qquad \text{(B1)}$$

where

$$v(\xi_0)=a_0+\sum_{k\in\mathbf{K}} a_k x_k+\xi_0\sigma_0, \qquad \sigma_0=\sqrt{\sum_{k\in\mathbf{K}} a_k^2\sigma_k^2}\,. \qquad \text{(B2)}$$

Introduce a change of variables,

$$\eta=\int_{-\infty}^{\xi_0}\exp\left(-\frac{u^2}{2\sigma_0{}^2}\right)du, \qquad d\eta=\exp\left(-\frac{\xi_0{}^2}{2\sigma_0{}^2}\right)d\xi_0\,, \qquad \text{(B3)}$$

then we have one-to-one relation between ξ_0 and η, there exists the inverse change $\xi_0(\eta)$, and hence we can write

$$P(X=1|Y=m)=\int_0^1 S(v(\xi_0(\eta)))d\eta\,. \qquad \text{(B4)}$$

We approximate this integral by a finite sum over n_0 points using midpoint rule. The interval [0,1] we split into n_0 segments of the length $1/n_0$, with the middle in the points $\eta_j=(j-0.5)/n_0$, $j=1,...,n_0$. The corresponding values of ξ_{0j} can be obtained from (B3), and they coincide with the probabilities that $\text{Prob}(\xi_0<\xi_{0j})=\eta_j=(j-0.5)/n_0$. Substituting these values of ξ_{0j} into (B4), we obtain (12).

References

Barbiero RP, Tuchman ML (2004) Changes in the crustacean communities of Lakes Michigan, Huron and Erie following the invasion of the predatory cladoceran *Bythotrephes longimanus*. Can J Fish Aquat Sci 61:2111-2125

Bossenbroek JM, Kraft CE, Nekola JC (2001) Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes, Ecol. Appl. 11: 1778-1788.

Burnham KP, Anderson DR (2001) Model selection and multimodel inference. A practical information-theoretic approach. 2nd ed., Springer, NY, USA, 488 pp.

Branstrator DK, Brown ME, Shannon LJ, Thabes M, Heimgartner K (2006). Range expansion of *Bythotrephes longimanus* in North America: evaluating habitat characteristics in the spread of an exotic invader. Biol. Invas. 8:1367-1379.

Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Method. Res. 33:261-304.

Cairns A, Elliott M, Yan ND, and Weisz E (2006) Operationalizing CAISN project 1.V, Technical Report No. 1: lake selection. Technical report prepared for the Canadian Aquatic Invasive Species Network. Dorset Environmental Science Centre, Dorset, Ontario.

Cairns A, Yan ND, Weisz E, Petruniak J, and Hoare J (2007) Operationalizing CAISN project 1.V, Technical Report No. 2: the large, inland lake, Bythotrephes survey — limnology, database design, and presence of Bythotrephes in 311 south-central Ontario lakes. Technical report prepared for the Canadian Aquatic Invasive Species Network. Dorset Environmental Science Centre, Dorset, Ontario.

Carroll RJ, Ruppert D, Stefanski LA (1995). Measurement error in nonlinear models. London : Chapman & Hall.

Clark JS, Carpenter SR, Barber M et al (2001) Ecological forecasts: an emerging imperative. Science 293:657-660

Cox ET (1978). Counts and measurements of Ontario lakes: watershed unit summaries based on maps of various scales by watershed unit. Ontario Ministry of Natural Resources, Toronto, Ontario.

Crawley MJ (2007) The R book. Wiley, USA, 942 pp.

Drake JM, Bossenbroek JM (2004) The potential distribution of zebra mussels (*Dreissena polymorpha*) in the U. S. A. BioScience 54:931-941

Drake JM, Baggenstos P, Lodge DM (2005) Propagule pressure and persistence in experimental populations. Biol. Lett. 1:480-483.

Ghosh JK, Samanta T (2001) Model selection - an overview. Curr. Sci. 80:1135-1144.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993-1009

Gurevitch J, Padilla DK (2004) Are invasive species a major cause of extinctions? Trends Ecol. Evol. 19: 470-474

Herborg L-M, Mandrak NE, Cudmore BC, MacIsaac HJ (2007) Comparative distribution and invasion risk of snakehead (Channidae) and Asian carp (Cyprinidae) species in North America. Canadian Journal of Fisheries and Aquatic Sciences 64:1723-1735.

Jerde C, Lewis MA (2007). Waiting for invasions: A framework for the arrival of non-indigenous species. American Naturalist: 170: 1-9.

Jerde CL, Bampfylde CJ, Lewis MA (2009). Chance establishment for sexual, semelparous species: Overcoming the Allee effect. American Naturalist: 173(6):734-46.

Keller RP, Lodge DM, Lewis MA, Shogren JF, eds. (2009) Bioeconomics of invasive species. Oxford University Press, NY,USA, 298 pp.

Kim, N. and Yan, N.D. 2010. Methods for rearing the invasive zooplankter *Bythotrephes* in the laboratory. Limnology and Oceanography: Methods. 8: (in press).

Lele S, Keim JL (2006) Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021-3028

Leung B, Drake JM, Lodge DM (2004) Predicting invasions: propagule pressure and the gravity of Allee effects. *Ecology* 85:1651-1660.

Leung B, Mandrak NE (2007) The risk of establishment of aquatic invasive species: joining invasibility and propagule pressure. Proceedings of the Royal Society of London Series B 274: 2603-2609

Linneman HV (1966) An Econometric Study of International Trade Flows. North-Holland Publishing Company, Amsterdam

Lockwood JL, Cassey P, Blackburn T (2005) The role of propagule pressure in explaining species invasions. Trends in Ecology & Evolution 20:223-228

MacIsaac HJ, Borbely JVM, Muirhead JR, and Graniero PA (2004) Backcasting and forecasting biological invasions of inland lakes. Ecol. Appl. 14(3): 773–783. doi:10.1890/02-5377.

MacIsaac HJ, Ketelaars HAM, Grigorovich IA, Ramcharan CW, Yan ND (2000). Modeling *Bythotrephes longimanus* invasions in the Great Lakes basin based on its European distribution. Archive für Hydrobiologie. 149:1-21.

Muirhead JR (2007) Forecasting dispersal of nonindigenous species. PhD thesis, University of Windsor, Windsor, Canada.

Muirhead JR, MacIsaac HJ (2005) Development of inland lakes as hubs in an invasion network. J. Appl. Ecol. 42: 80-90.

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl. Ecol. 43:405-412

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford : Oxford University Press.

Pérez-Fuentetaja A., P.J. Dillon, N.D. Yan, and D.J. McQueen. 1999. Significance of dissolved organic carbon in prediction of thermocline depth in small Canadian Shield lakes. Aquatic Ecol. **33**: 127-133.

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics 2005: 273-288.

Potapov A (2009) Stochastic model of lake system invasion and its optimal control: neurodynamic programming as a solution method. Nat. Res. Mod. 22, 257-288.

Potapov A, Muirhead JR, Lele SR, Lewis MA (2010) Stochastic gravity models for modeling lake invasions. Ecological Modelling (to appear). Preprint is available at http://www.math.ualberta.ca/~apotapov/Papers/stochastic_gravity_models.pdf

Pulliam HR (2000) On the relationship between niche and distribution. Ecol Lett 3:349-361

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Romeijn HE, Smith RL (1994) Simulated annealing for constrained global optimization. J. Global Optimization, 5:101-126.

Sala OE, Chapin, III FS, Armesto JJ, Berlow E, Bloomfield J, Dirzo R, Huber-Sanwald E, Huenneke LF, Jackson RB, Kinzig A, Leemans R, Lodge DM, Mooney HA, Oesterheld M, Poff NL, Sykes MT, Walker BH, Walker M, Wall DH (2000) Global Biodiversity Scenarios for the Year 2100. Science. 287:1770-1774.

Spall JC (2003) Introduction to stochastic search and optimization. Wiley, Hoboken, NJ.

Stockwell D (2007) Niche modeling: predictions from statistical distributions. Chapman and Hall, Boca Raton

Stefanski LA and Carroll RJ (1985). Covariate measurement error in logistic regression. Ann. Stat. 13:1335-1351.

Strecker AL, Arnott SE (2005) Impact of Bythotrephes invasion on zooplankton communities in acid-damaged and recovered lakes on the Boreal Shield. Can J Fish Aquat Sci 62:2450-2462

Taylor CM; Hastings A (2005) Allee effects in biological invasions. Ecology Letters 8(8): 895-908.

Thomas RW, Huggett RJ (1980) Modeling in geography: a mathematical approach. Rowman and Littlefield, Lanham, MD, USA, 338 pp.

Weisz EJ and Yan ND (2010). Relative value of limnological, geographic, and human use variables as predictors of the presence of *Bythotrephes longimanus* in Canadian Shield lakes. Can. J. Fish. Aquat. Sci. 67(3): 462–472.

Wittmann M, Lewis MA, Yan N (2010). A mechanistic model for the establishment success of *Bythotrephes* in North American lakes. To be submitted.

Yan ND, Dunlop WI, Pawson TW, and MacKay LE (1992) Bythotrephes cederstroemi (Schoedler) in Muskoka lakes: first records of the European invader in inland lakes in Canada. Can. J. Fish. Aquat. Sci. 49(2): 422–426. doi:10.1139/f92-048.

Yan ND, Paterson AM, Somers KM and Scheider WA (2008) An introduction to the Dorset Special Issue: Transforming understanding of the factors that regulate aquatic ecosystems on the southern Canadian Shield. Can. J. Fish. Aquat. Sci. 65: 781-785.

Yan ND, Girard R, Boudreau S (2002) An introduced invertebrate predator (*Bythotrephes*) reduces zooplankton species richness. Ecology Letters 5:481-485

Yan ND and Pawson TW (1998) Variation in size and abundance of the exotic invader *Bythotrephes cederstroemi* in Harp Lake, Canada. Hydrobiologia 361:157-168

Zipf GK (1946) The P1P2/D hypothesis: on the intercity movement of persons. Am Soc Rev 11:677-686

Tables

Table 1. Lake covariates in 300-lakes database, mean and standard deviation for 306 lakes, and estimates of temporal variability (s.e.) σ_k.

k	Covariate	Symbol	units	μ_k	σ_{Sk}	σ_k
1	Lake area	A	Ha	67.95	120.4	—
2	Lake perimeter	Per	m	5996.2	8861.0	—
3	Lake elevation	Elev	m	327.4	83.02	—
4	The bottom of strata sampled or maximum depth of composite sample	D	m	4.69	2.30	—
5	The Secchi depth of the lake at sample date and time	SD	m	3.74	1.76	1.00
6	Sodium unfiltered total.	Na	mg/L	3.49	8.53	0.21
7	Potassium unfiltered total	K	mg/L	0.42	0.40	0.050
8	Magnesium unfiltered total	Mg	mg/L	0.70	0.40	0.064
9	Calcium unfiltered total	Ca	mg/L	2.78	2.17	0.24
10	Total Phosphorus; unfiltered total, field replicate 1	P_1	µg/L	10.16	8.34	1.48
11	Total Phosphorus; unfiltered total, field replicate 2	P_2	µg/L	10.29	8.54	1.48
12	SiO_3 unfiltered reactive	Si	mg/L as Si	0.64	0.52	0.26
13	Dissolved Organic Carbon	DOC	mg/L	6.01	3.11	0.48
14	Total inflection point alkalinity	Alki	mg/L as CaCO3	4.15	4.43	0.41
15	Total fixed end point alkalinity to pH 4.5	Alk	mg/L as CaCO3	6.19	4.38	0.41
16	pH	pH	—	6.18	0.57	0.19
17	Conductivity at 25°C	CD	µS/cm	42.09	61.08	2.76
18	Propagule pressure (added to data set)	λ	$Year^{-1}$	—	—	—

Table 2. The best models for different number of covariates. Within each category AIC and BIC criteria give the same results. The values of coefficients κ, m, a_k are given for normalized covariates x_k.

n_c	AIC	BIC	AUC	Covariates	κ	m	a_k, k=0,…,n_c
0	146.6	154.0	0.853	λ	2.50	1.86	7.49
1	121.8	132.9	0.914	λ, PH	13.03	2.23	-2.12 5.50
	133.7	144.9	0.890	λ, K	4.04	1.75	3.47 20.22
	137.3	148.5	0.884	λ, Alki	14.59	2.28	-0.08 2.70
2	118.7	133.6	0.926	λ, P1, PH	10.82	2.11	-5.28 -5.55 8.98
	120.0	134.9	0.923	λ, P2, PH	12.96	2.21	-3.93 -3.72 6.88
	120.5	135.4	0.924	λ, SD, PH	10.11	2.11	-7.03 5.96 15.60
	120.9	135.8	0.919	λ, Elev, PH	12.32	2.17	-2.74 -1.16 5.71
	121.2	136.1	0.919	λ, K, PH	14.90	2.28	-1.96 3.06 4.76
3	115.8	134.4	0.935	λ, Elev, P1, PH	12.11	2.09	-5.96 -1.96 -5.82 6.87
	117.1	135.7	0.933	λ, Elev, P2, PH	14.90	2.19	-4.58 -1.59 -4.16 5.30
	118.0	136.6	0.929	λ, Ca, P1, PH	15.76	2.28	-3.79 1.90 -3.41 5.82
	118.1	136.7	0.928	λ, P1, Alki, PH	16.16	2.29	-3.33 -3.04 1.84 4.80
	118.2	136.8	0.930	λ, K, P1, PH	13.54	2.20	-3.82 3.49 -3.49 6.15

4	115.2	137.5	0.936	λ, P2, Alki, Alk, PH	38.44	2.65	-15.10	-17.52	64.13	-57.86	12.80
	115.8	138.1	0.937	λ, Elev, P2, Si, PH	55.95	2.52	-3.87	-1.44	-3.24	0.87	2.09
	115.9	138.2	0.939	λ, Elev, P1, Si, PH	14.06	2.03	-4.81	-1.64	-4.46	1.00	3.85
	116.8	139.2	0.936	λ, Elev, Ca, P1, PH	15.13	2.20	-5.18	-1.46	0.92	-4.87	5.85
	117.1	139.4	0.934	λ, P1, Alki, Alk, PH	18.49	2.35	-7.77	-7.40	30.03	-26.12	10.41

Table 3. 21 best models with propagule pressure and 3 lake covariates (n_c=3). According to their AIC values, all of them cannot be totally rejected as potential true model. 5 models that include covariates including perimeter, area, sampling depth, Secchi depth, and, probably, SiO_3 show signs of overfitting: too big $|a_k|$ for pH and phosphorous. Other predictors show a pattern: all of them have the structure pH+{P or DOC}+{one of: elevation, Ca, K, Na, Mg, alkalinity, conductivity}. The values of coefficients κ, m, a_k are given for normalized covariates x_k.. Figures in bold denote the predictors with possible overfitting (see text), and covariates in bold show potential source of overfitting.

#	AIC	BIC	AUC	Covariates	κ	m	a_k, k=0,…,n_c			
1	115.8	134.4	0.935	λ, Elev, P1, PH	12.11	2.09	-5.96	-1.96	-5.82	6.87
2	117.1	135.7	0.933	λ, Elev, P2, PH	14.90	2.19	-4.58	-1.59	-4.16	5.30
3	118.0	136.6	0.929	λ, Ca, P1, PH	15.76	2.28	-3.79	1.90	-3.41	5.82
4	118.1	136.7	0.928	λ, P1, Alki, PH	16.16	2.29	-3.33	-3.04	1.84	4.80
5	118.2	136.8	0.930	λ, K, P1, PH	13.54	2.20	-3.82	3.49	-3.49	6.15
6	119.0	137.6	0.927	λ, P1, PH, CD	14.52	2.25	-3.58	-3.34	6.22	2.63
7	119.0	137.6	0.929	λ, **Per**, P1, PH	9.06	1.98	**-16.98**	**1.71**	**-13.18**	**27.84**
8	119.1	137.7	0.927	λ, P1, Alk, PH	14.21	2.23	-3.67	-3.46	1.36	5.50
9	119.4	138.0	0.926	λ, Ca, P2, PH	17.89	2.35	-3.29	1.72	-2.84	5.04
10	119.5	138.1	0.926	λ, P2, Alki, PH	18.06	2.34	-3.00	-2.60	1.63	4.39
11	119.6	138.2	0.926	λ, K, P2, PH	15.52	2.27	-3.32	2.96	-2.87	5.38
12	119.7	138.4	0.925	λ, Na, P1, PH	13.47	2.22	-3.66	2.59	-3.58	6.55

13	119.7	138.4	0.924	λ, Ca, DOC, PH	17.98	2.35	-2.66	1.67	-2.14	4.16
14	119.9	138.5	0.928	λ, **SD**, P1, PH	9.80	2.06	**-10.18**	**3.52**	**-8.78**	**16.87**
15	120.0	138.6	0.927	λ, P1, Si, PH	8.64	1.98	-5.99	-6.97	1.41	9.18
16	120.1	138.7	0.925	λ, P2, PH, CD	16.56	2.31	-3.14	-2.88	5.33	2.45
17	120.2	138.8	0.927	λ, **D**, P1, PH	11.63	2.18	**-7.75**	**-1.86**	**-11.00**	**14.29**
18	120.2	138.8	0.927	λ, Mg, P1, PH	12.46	2.18	-4.46	1.71	-4.05	7.46
19	120.4	139.0	0.926	λ, **Per**, P2, PH	10.43	2.06	**-12.59**	**1.33**	**-8.90**	**21.35**
20	120.4	139.0	0.926	λ, **A**, P1, PH	10.11	2.06	**-6.93**	**0.33**	**-6.37**	**11.45**
21	120.4	139.0	0.924	λ, Elev, DOC, PH	14.36	2.20	-2.91	-1.10	-1.75	4.21

Table 4. Correlation coefficients for lake covariates in Table 1. The correlation matrix was calculated by R function cor.

#		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Covariate	A	Per	Elev	D	SD	Na	K	Mg	Ca	P1	P2	Si	DOC	Alki	Alk	PH	CD	λ
1	A	1.00	0.92	-0.01	0.47	0.19	-0.06	-0.02	0.03	-0.02	-0.18	-0.17	0.04	-0.16	-0.01	0.01	0.10	-0.05	0.86
2	Per	0.92	1.00	-0.04	0.47	0.21	-0.06	-0.03	-0.00	-0.04	-0.20	-0.19	0.02	-0.17	-0.04	-0.03	0.06	-0.06	0.82
3	Elev	-0.01	-0.04	1.00	0.21	0.11	-0.32	-0.22	-0.30	-0.39	-0.19	-0.17	0.05	-0.17	-0.38	-0.38	-0.26	-0.33	-0.22
4	D	0.47	0.47	0.21	1.00	0.63	-0.13	-0.11	-0.08	-0.11	-0.50	-0.47	-0.08	-0.48	-0.11	-0.10	0.06	-0.15	0.47
5	SD	0.19	0.21	0.11	0.63	1.00	-0.04	0.01	0.01	0.01	-0.63	-0.59	-0.29	-0.67	-0.01	-0.01	0.28	-0.05	0.24
6	Na	-0.06	-0.06	-0.32	-0.13	-0.04	1.00	0.53	0.69	0.76	0.07	0.08	0.00	0.12	0.63	0.62	0.23	0.95	-0.02
7	K	-0.02	-0.03	-0.22	-0.11	0.01	0.53	1.00	0.54	0.57	0.06	0.07	0.11	0.10	0.48	0.47	0.28	0.73	0.02
8	Mg	0.03	-0.00	-0.30	-0.08	0.01	0.69	0.54	1.00	0.81	0.08	0.07	0.26	0.10	0.82	0.80	0.46	0.71	0.08
9	Ca	-0.02	-0.04	-0.39	-0.11	0.01	0.76	0.57	0.81	1.00	0.07	0.06	0.10	0.10	0.91	0.89	0.39	0.81	0.05
10	P1	-0.18	-0.20	-0.19	-0.50	-0.63	0.07	0.06	0.08	0.07	1.00	0.91	0.25	0.69	0.11	0.10	-0.19	0.07	-0.19
11	P2	-0.17	-0.19	-0.17	-0.47	-0.59	0.08	0.07	0.07	0.06	0.91	1.00	0.29	0.67	0.09	0.09	-0.19	0.08	-0.18
12	Si	0.04	0.02	0.05	-0.08	-0.29	0.00	0.11	0.26	0.10	0.25	0.29	1.00	0.36	0.17	0.17	-0.01	0.03	0.05
13	DOC	-0.16	-0.17	-0.17	-0.48	-0.67	0.12	0.10	0.10	0.10	0.69	0.67	0.36	1.00	0.06	0.06	-0.34	0.13	-0.19
14	Alki	-0.01	-0.04	-0.38	-0.11	-0.01	0.63	0.48	0.82	0.91	0.11	0.09	0.17	0.06	1.00	0.99	0.51	0.68	0.06
15	Alk	0.01	-0.03	-0.38	-0.10	-0.01	0.62	0.47	0.80	0.89	0.10	0.09	0.17	0.06	0.99	1.00	0.50	0.67	0.07
16	PH	0.10	0.06	-0.26	0.06	0.28	0.23	0.28	0.46	0.39	-0.19	-0.19	-0.01	-0.34	0.51	0.50	1.00	0.27	0.24
17	CD	-0.05	-0.06	-0.33	-0.15	-0.05	0.95	0.73	0.71	0.81	0.07	0.08	0.03	0.13	0.68	0.67	0.27	1.00	-0.02
18	λ	0.86	0.82	-0.22	0.47	0.24	-0.02	0.02	0.08	0.05	-0.19	-0.18	0.05	-0.19	0.06	0.07	0.24	-0.02	1.00

Figures

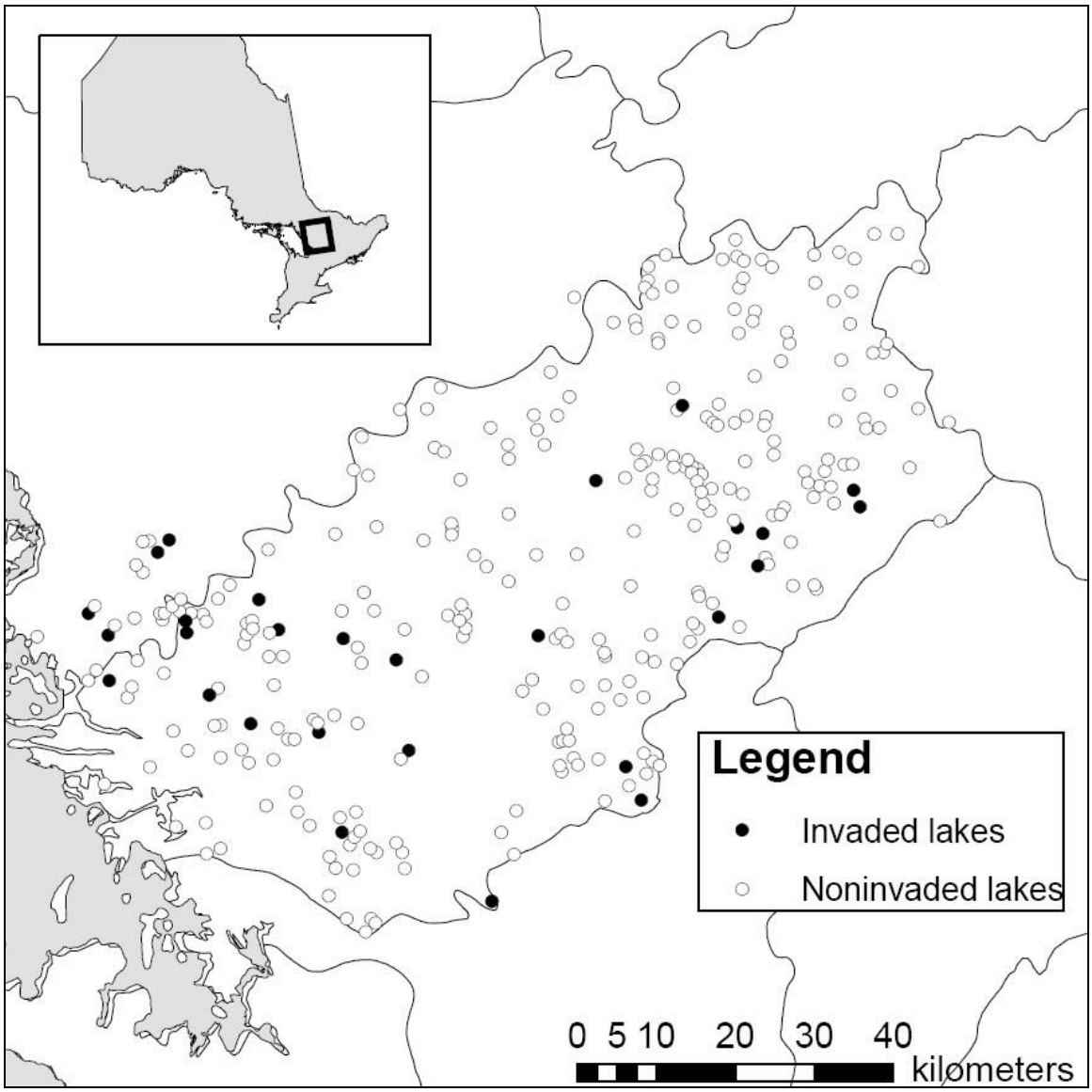

Fig. 1 Lakes in Muskoka watershed used in the study

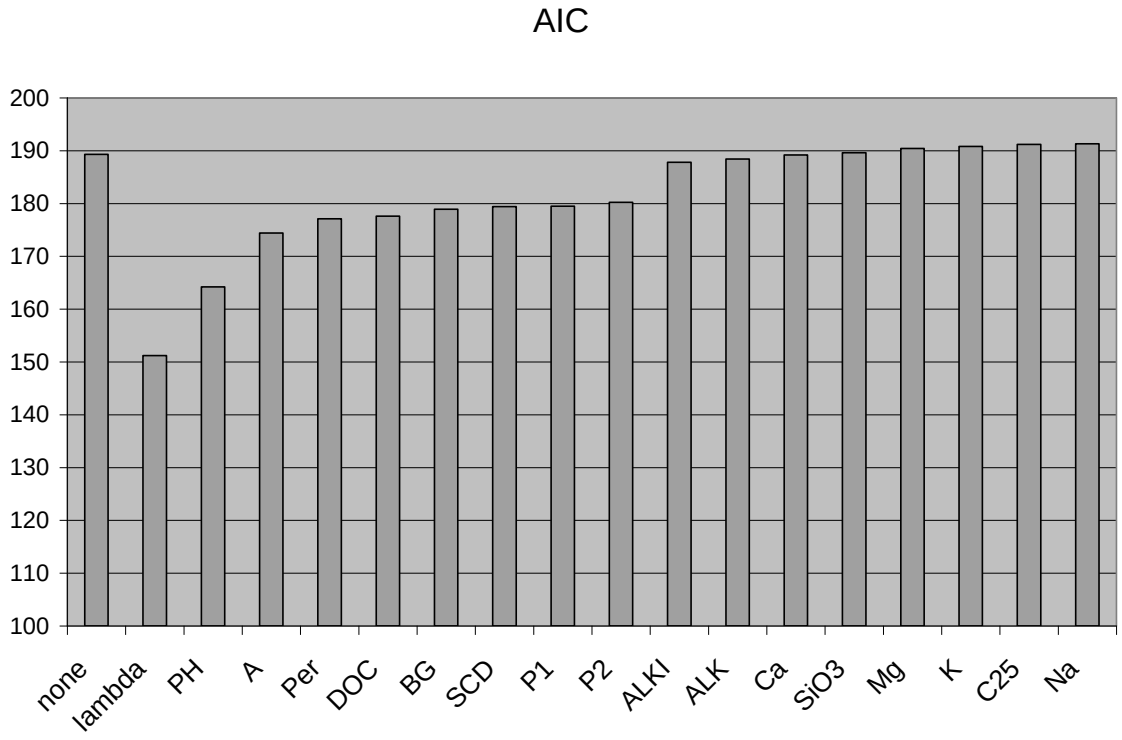

Fig. 2. Results for fitting presence-absence data with single variable logistic predictor. The best predictor is gravity score (propagule pressure). Therefore, invasion is in progress, and many suitable lakes may be not invaded because the invader has not reached them. This allows to hope that habitat suitability-type models are more appropriate for predictions.

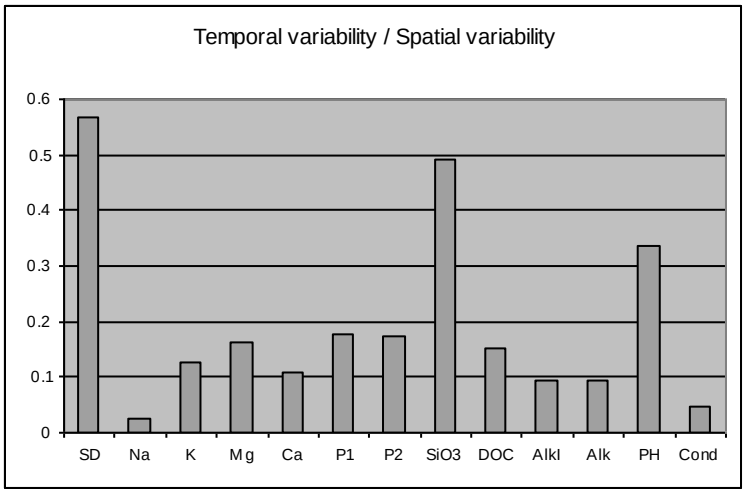

Fig. 3. Relative variability of chemical covariates σ_k / σ_{Sk}. For all covariates spatial variations exceed temporal ones at least twice.

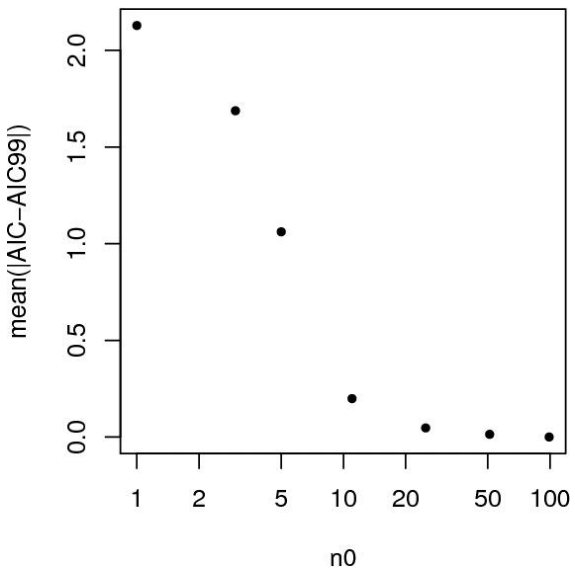

Fig. 4. Choice of n_0. Model (13) has been fitted to data for n_0=1, 3, 5, 11, 25, 51, 99 for all possible combinations of λ and up to 4 other covariate, 3214 models totally. The figure shows the difference |AIC(n_0)–AIC(99)| averaged over all models. Changes are insignificant for $n_0 \geq 25$, n_0=51 appears to be close the optimal choice.

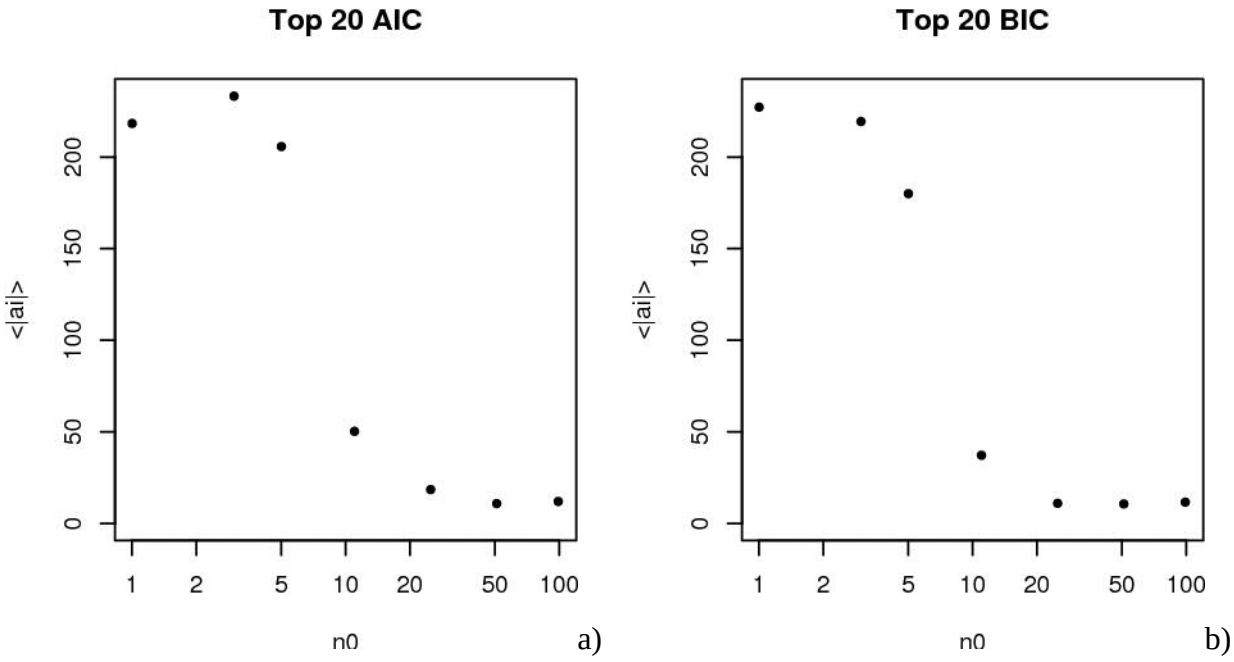

Fig. 5. Another effect related with the choice of n_0: overfitting in logistic regression. If data by chance are "too well separated", the maximum likelihood fitting makes the logistic function close to an abrupt step, which can be diagnosed by high values of $|a_i|$. The panels show $\max_i |a_i|$ averaged over for 20 best models according to AIC or BIC criteria. Without taking into account data variability (n_0=1), models with the best AIC values are overfitted. As n_0 increases, the data separation becomes less pronounced, and the absolute values of model coefficients become smaller. In agreement with Fig. 3, the choice n_0=51 is optimal.

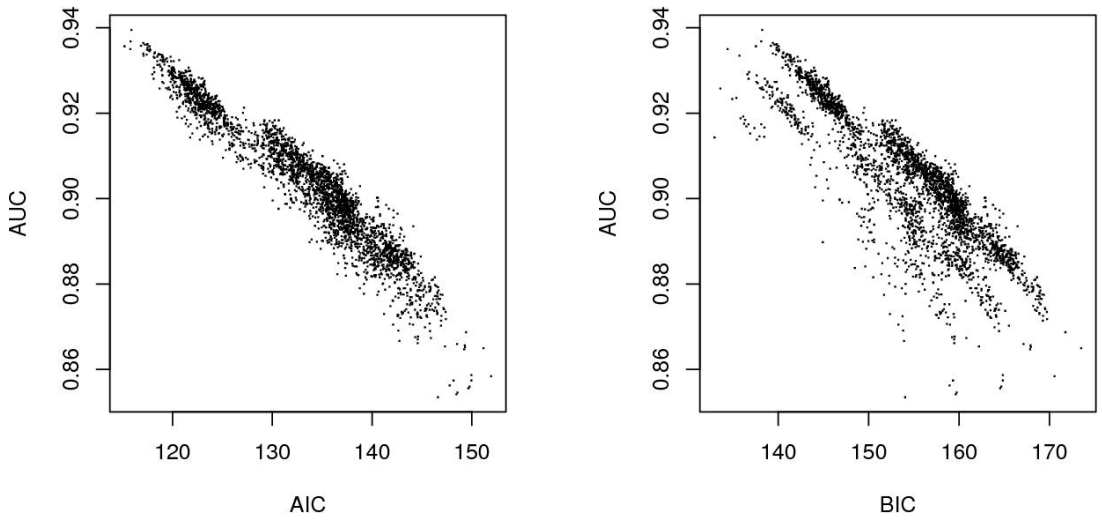

Fig. 6. Dependence of AUC value on AIC and BIC for all 3214 models, n_0=51.

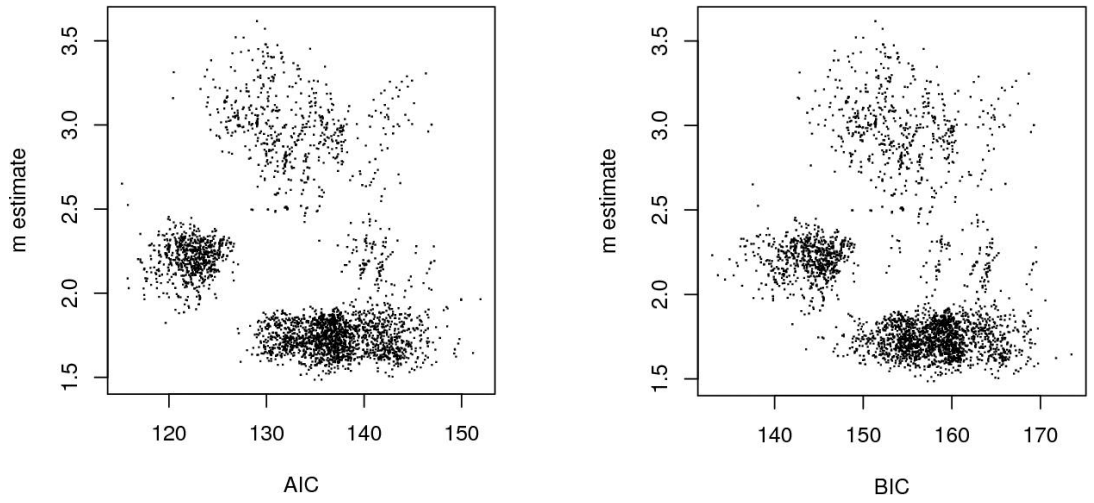

Fig. 7. Estimates of *m* for all tested models. Models with the least AIC/BIC demonstrate estimates slightly greater than 2. We interpret this as a sign of presence of Allee effect.

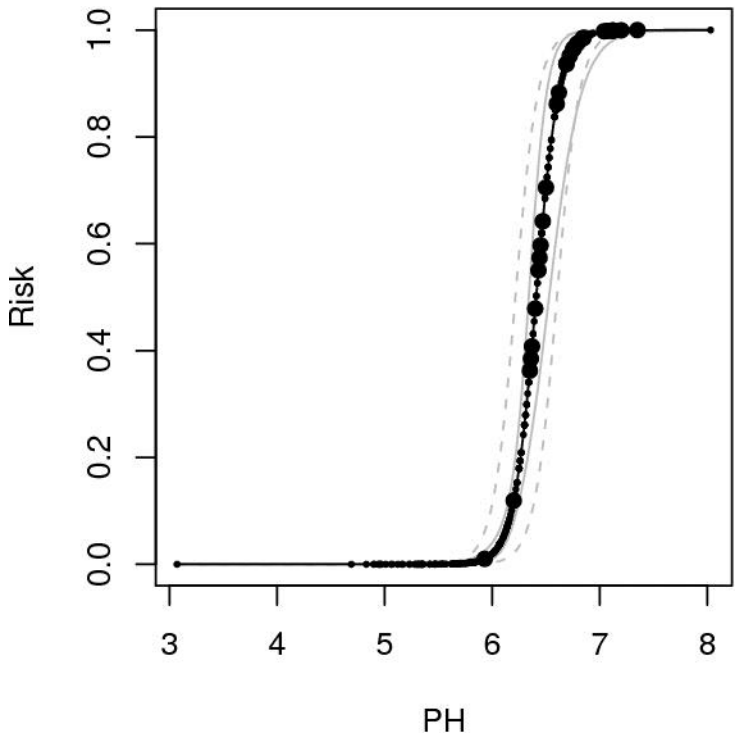

Fig. 8. BIC-best lake invasibility predictor (establishment risk estimate) using only pH (solid line). Gray solid lines shows error estimate, dashed lines show temporal variability, small bullets – uninvaded lakes, large bullets – invaded lakes. According to the predictor, 104 lakes have risk>0.5 and only 22 of them are invaded.

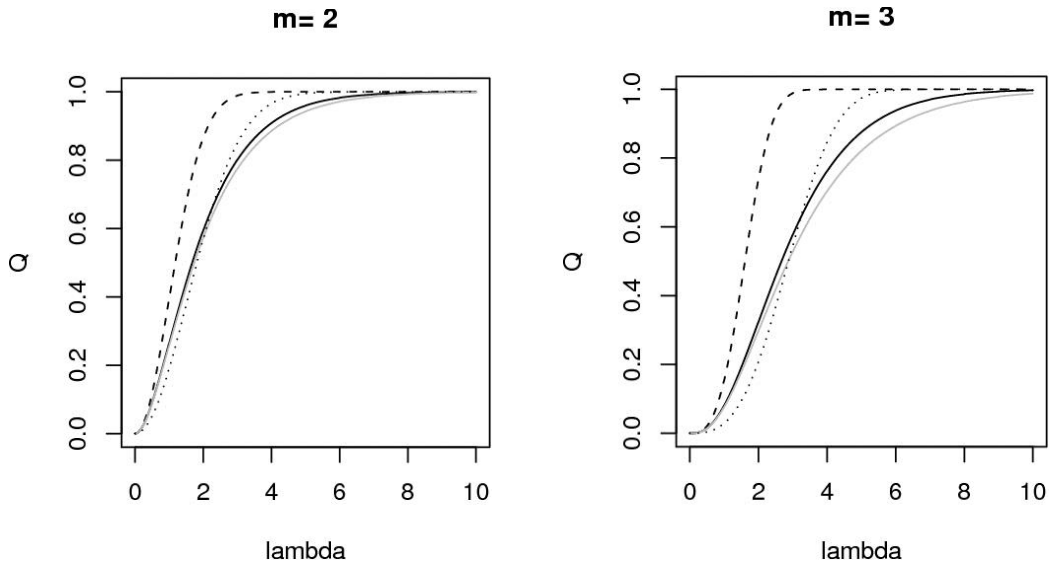

Fig. 9. Comparison of exact $Q(\lambda, m)$ and its approximations for *m*=2 and 3: black solid line – exact $Q(\lambda, m)$; dashed – $Q_1(\lambda, m) = 1 - \exp(-\lambda^m / m!)$; dotted line – $Q_1(\lambda, m) = 1 - \exp(-\alpha\lambda^m / m!)$ where α is obtained by fitting Q_1 to Q, (actual coefficient will be fitted anyway); gray line $Q_2(\lambda, m) = \left[1 - \exp\left(-\lambda / (m!)^{1/m}\right)\right]^m$.