Potapov, A., Muirhead, J. R., Lele, S. R., & Lewis, M. A. (2011). Stochastic gravity models for modeling lake invasions. Ecological Modelling, 222(4), 964-972.

# Stochastic gravity models for modeling lake invasions

Alex Potapov[a,b,*], Jim R. Muirhead[a,c], Subhash R. Lele[a,b], Mark A. 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

## Abstract

Freshwater aquatic systems in North America are being invaded by many different species, ranging from fish, mollusks, cladocerans to various bacteria and viruses. These invasions have serious ecological and economic impacts. Human activities such as recreational boating are an important pathway for dispersal. Gravity models are used to quantify the dispersal effect of human activity. Gravity models currently used in ecology are deterministic. This paper proposes the use of stochastic gravity models in ecology, which provides new capabilities both in model building and in potential model applications. These models allow us to use standard statistical inference tools such as maximum likelihood estimation and model selection based on information criteria. To facilitate prediction, we use only those covariates that are easily available from common data sources and can be forecasted in future. This is important for forecasting the spread of invasive species in geographical and temporal domain. The proposed model is portable, that is it can be used for estimating relative boater traffic and hence relative propagule pressure for the lakes not covered by current boater surveys. This makes our results broadly applicable to various invasion prediction and management models.

Key words: gravity models; biological invasions; dispersal models; model selection

# Introduction

Biological invasions are widespread amongst freshwater aquatic ecosystems.  In North America, macroscopic invasive species cover a broad range of phyla, from plants (e.g. Eurasian watermilfoil *Myriophyllum spicatum* L., *Cabomba caroliniana*), exotic fish (e.g. Asian carp) and invertebrates (e.g. zebra mussels *Dreissena polymorpha*, spiny waterflea *Bythotrephes longimanus*). Added to these are the microscopic invaders, including bacteria and viruses.  The better-known biological invaders have major ecological and economic impacts on the ecosystems they encounter.  These impacts include interference with the feeding, growth, movement and reproduction of native species, bioaccumulation of pollutants, and the fouling of recreational and industrial facilities (Parker et al., 1999; Burbidge and Manly, 2002; Lovell et al., 2006; Crowl et al., 2008).

In many systems, it is human activity that is the dominant factor governing transfer of invasive propagules. For example, with zebra mussels and spiny waterfleas, the most important factor in overland transportation from invaded to uninvaded lakes is tied to recreational boats (Johnson and Padilla, 1996; Schneider et al., 1998; MacIsaac et al., 2004). When a boat is used at an invaded lake, invaders can attach to the boat, trailer or fishing equipment. If the boat or equipment is used at a different lake within a short enough time, the invader may be released into a new lake.  This may, in turn, lead to the establishment of new invader populations in previously uninvaded lakes.

Controlling the spread of invasion due to human transport requires knowledge of the present invasion status and quantification of spread rate. Intensive sampling effort has gone into measuring the current distributions of some of the more notorious invaders, such as zebra mussels and spiny water fleas (Kraft et al., 2002; MacIsaac et al., 2004; Bobeldyk et al., 2005). Much effort has also gone into assessing the *propagule pressure* (characterized by rate of transfer of invasive propagules) from invaded to uninvaded water bodies. To estimate it, one needs to know the number of boaters traveling between the lakes or the boater flow.  For $N$ lakes, assuming the flows symmetric, this requires estimating $N(N\text{-}1)/2$ parameters, and hence the proportional amount of data about boater movement, that is detailed surveys.

Similar problem aroused long ago in economics and geography, where it was necessary to estimate transport flows between the cities. At present the common approach to it is the use of so-called gravity models (Thomas and Huggett, 1980). In deterministic gravity models, the propagule pressure is characterized by its mean value $\lambda_{ij}$ from location $i$ ("origin" or "source") to location $j$ ("destination"). In turn, this mean propagule pressure is approximated by product of three factors: $\lambda_{ij} \sim T_i \times W_j \times \phi(d_{ij})$, where $T_i$ is the number of potential travelers at location $i$ or "repulsiveness" of the location; $W_j$ is "attractivity" of location $j$; $\phi(d_{ij})$ is a "distance deterrence" factor, which describes the fact that short trips are more frequent than long ones; $d_{ij}$ is the distance between the source and the destination (Thomas and Huggett 1980; Sen and Smith 1995). $T_i$ and $W_j$ depend on covariates characterizing source and destination locations respectively. Sometimes, when sources and destinations possess similar characteristics, $T_j$ and $W_j$ may coincide: in economics both may be proportional to population in the locations. Then

$$\lambda_{ij} \sim W_i \times W_j \times \phi(d_{ij}). \qquad (1)$$

The gain from using such model is quite obvious: the number of parameters to be estimated from data become fewer. The name "gravity model" appeared because the form of the first models resembled Newton's Law of Gravity, where city population played the role of attracting mass, and $\phi(d) = d^{-2}$.

Deterministic gravity models have been successfully applied in invasion ecology for estimating propagule pressure of invaders transported by the boaters into a lake (Baxter and Ewing, 1981; Bossenbroek et al., 2001; Leung et al., 2004; Muirhead 2007). Building a gravity models for recreational boater movement includes the following steps:

1. Define the set of covariates determining attractivity and repulsiveness terms in (1) and functional forms of the dependencies for these covariates. We assume that the covariates and the distances are available, so only a few parameters in the model remain unknown.
2. Obtain survey data on actual boater movement between sources and destinations. The surveys provide the actual number of travelers between the locations $n_{ij}$, or some data, from which $n_{ij}$ may be derived.
3. Obtain the model parameters by fitting the model to the data. Different types of gravity models vary according to ways of fitting the model parameter in (1) to the data $n_{ij}$.

For the step 1 a number of functional forms have been tested. For example, the deterrence term is taken in the form of $\phi(d) = d^{-\beta}$ (Bossenbroek et al., 2001) or $\phi(d) = \exp(-\beta d)$ (Muirhead, 2007) with β unknown. Attractivity of a lake may be postulated as proportional to lake area $A_i$ as $W_i$=$C_iA_i$ (Bossenbroek et al., 2001) or to its logarithm (Muirhead, 2007).  Step 2 is the hardest to do.  Step 3 depends on the model type and will be discussed below in more details.

In applications there are two important requirements for a gravity model: accuracy and portability. Accuracy is achieved mainly in step 1, by the choice of the model terms and parameters.  Portability means that one can add to the model sources and/or destinations, not covered by the boater survey, without rebuilding the model. In particular, portability allows one to fit a model to one set of lakes and later apply it to another one. Portability of a model is important for prediction and management of biological invasions, when the number of invaded lakes changes in time, because surveys are expensive and time-consuming.

It appears that not all existing types of gravity models are portable. For example, to achieve better accuracy, source-specific or destination-specific coefficients are introduced like in the above example, $W_i$=$C_iA_i$. Then these coefficients are not available for new locations, and the model loses portability towards sources, destinations, or both.  Therefore, portability requires a reduction of the number of region-specific parameters to a minimum.

One more typical problem in deterministic gravity model fitting to survey data is that for some lake pairs there may be only a few boaters using it. If only a small part of boaters respond to a survey, this may be true for most lake pairs. In such situation it may be more reasonable to consider $n_{ij}$ as a realization of an integer-valued random process rather than an approximation to the mean boater flow, and maximum likelihood fitting might be more natural than least squares one.

To address all the mentioned problems, we have developed a stochastic gravity model.  Although such approaches have been applied to economic gravity models (e.g. Flowerdew and Atkin, 1982) and there is some similarity with the way of gravity model parameters fitting in (Ferrari et al. 2006), they are new to ecological gravity models.  We consider the boater movement as a Poisson process with the intensity $\lambda_{ij}$, which has the form (1).  The framework of statistical model selection allows us to use maximum likelihood approach to model fitting, to use information-based model selection criteria (Burnham and Anderson, 2002) for choosing the best functional forms for lake attractivity and distance deterrence and for choosing the best set of covariates determining them. There is no need to confine their set only to lake area. Attractivity of a lake potentially may depend on many lake covariates, e.g. perimeter, depth, availability of certain facilities, fish species living in a lake and so on, and using model selection framework one can test these hypotheses and determine the relevant ones, provided they are known. Flexibility due to model selection capabilities enabled us to build a portable gravity model without location-specific coefficients, and its use in another modeling project confirmed its efficiency in a new region. Finally, if more detailed surveys will be available in the future, this approach may allow the researches to study variability of propagule pressure, which may lead to more advanced invasion risk models.

The gravity model developed in this paper was a part of bigger project on analysis of risk of invasion of Ontario lakes by a zooplanktonic invader spiny waterflea *Bytotrephes longimanus* (Yan et al. 2002), which poses a serious threat for lake ecosystems (see e.g. Boudreau and Yan, 2003). This invader quickly spreads between Ontario lakes (MacIsaac et al., 2004, ), and therefore portability of a gravity model for its propagule pressure is an important issue.

## 2. The general structure of gravity models: accuracy and portability

Most of deterministic gravity models used for predicting recreational boaters movement (Bossenbroek et al., 2001; Leung et al. 2004; Keller et al., 2009) are the so-called constrained gravity models (Thomas and Huggett, 1980). Here we briefly describe the main idea of constraining and explain, why do we need a different approach.

The idea of a doubly constrained gravity model is to ensure good agreement between the model estimates $\lambda_{ij}$ and actual data $n_{ij}$ by introducing fitting parameters and imposing constraints: at each of $N_S$ source locations the total outflow in the model should be equal to that for the data,

$\sum_j \lambda_{ij} = \sum_j n_{ij},\ i = 1,\dots,N_S$ ; similarly, at each of $N_D$ destinations the total inflows have to be equal, $\sum_i \lambda_{ij} = \sum_i n_{ij},\ j = 1,\dots,N_D$. To be able to satisfy the constraints, the model must have at least $N_S$+$N_D$ free parameters. The parameters are introduced as $N_S$ factors for the source terms ($C_{Ti}$), and $N_D$ factors for the attractivity terms ($C_{Wj}$). The expression for the mean flow becomes $\lambda_{ij} = C_{Ti} \times T_i \times C_{Wj} \times W_j \times \phi(d_{ij})$, formulas for obtaining $C_{Ti}$ and $C_{Wj}$ can be found e.g. in (Thomas and Huggett, 1980). At the same time, the dependencies of *T, W,* and $\phi(d)$ on the location covariates are fixed or contain only a few free parameters. For example, (Bossenbroek et al., 2001) use *T*={the number of boaters in a county}, *W*={lake area}, $\phi(d) = d^{-\alpha}$, where α to be fitted to data. These remaining parameters are obtained by minimizing the total squared error $\sum_{ij} (\lambda_{ij} - n_{ij})^2$ or Pearson statistic $\sum_{ij} (\lambda_{ij} - n_{ij})^2 / \lambda_{ij}$.

Partially constrained models use fewer unknown factors and constrain only outflows (production-constrained model) or only inflows (attraction-constrained model). This significantly simplifies expressions for $C_{Ti}$ or $C_{Wj}$ (Thomas and Huggett, 1980), though at the cost of quality of approximation and model efficiency (Muirhead, 2007).

In some cases the constrained gravity models are enough for practical purposes. Muirhead (2007) has successfully used constrained gravity models to calculate *Bythotrephes* flow between the lakes named in the survey. This model was a part of the bigger model for predictions of *Bythotrephes* invasion risk. In this study, propagule pressure, as estimated by a production constrained gravity model, accounted for most of the inland lake invasions when combined with data on habitat suitability and fish community composition (Muirhead and MacIsaac, submitted). This class of gravity models has also successfully captured invasions for other species. Similar production-constrained gravity models for zebra mussel invasions have captured long-distance dispersal of zebra mussels across the U.S. (Bossenbroek et al., 2007), and estimates of boater traffic from these types of models correspond well with creel surveys (Leung et al., 2006).

However, constraints impose a serious restriction on portability of the model to a new region or just to ability to add new locations. For production-constrained models one has to know $C_{Ti}$, and hence it is impossible to introduce a new source without doing a new survey and rebuilding the model. New destinations, on the other hand, can be added. Similarly, for attraction-constrained model one can add new sources, but not destinations. Finally, doubly constrained models do not allow any extensions. Modeling of invasion progress typically require adding new sources, and considering different region require adding new destinations, that is full model portability. In our studies we needed a portable gravity model that can be fitted to data for the part of Ontario lakes covered by the boater survey, and subsequently used for the lakes from a different part of Ontario, where no survey data were available.

To achieve full portability, it is necessary to minimize the number of model parameters that are uniquely related to a certain lake or the whole region. Attractivity of a location *W* must have the same functional form for all locations, and dependence on the specific location must appear only through covariates characterizing it. For example, the relation $W_i = A_i^{\alpha} \exp(\beta A_i)$ where $\alpha$ and $\beta$ are the same for all locations is portable, while $W_i = A_i^{\gamma_i}$ with different $\gamma_i$ for each location is not. We may assume that parameters $\alpha$ and $\beta$ reflect some patterns of boater decision-making, and hence may be used at a different place. Parameters $\gamma_i$ also may reflect the same patterns, but they are not defined outside the survey region.

At the same time, at least one parameter cannot be transferred to other region: all estimated $\lambda_{ij}$ must be proportional to the total number of travelers in a system. Some models satisfy explicit constraint $\sum_{ij} \lambda_{ij} = \sum_{ij} n_{ij}$ or an equivalent one, or it may implicitly emerge from data fitting. In another region the total number of travelers is different, and hence the common factor should change. However, a portable model in a new region should be able to provide relative intensities of transport flows, and hence relative invasion risks. This may be valuable information for prediction and management of invasions.

Therefore, a natural strategy to build a portable model is to increase the number of common fitting parameters in attractivity and deterrence terms, and to reduce the region-dependent parameters to a common factor only, that is

$$\lambda_{ij} = C \times W_i(\text{covariates}_i, \text{parameters}_W) \times W_j(\text{covariates}_j, \text{parameters}_W) \times \phi(d_{ij}, \text{parameters}_\phi) \qquad (2)$$

Parameters may be fitted by standard least squares methods, as with other deterministic models. The number of parameters has to be sufficiently large to ensure good fitting properties. However, this raises a problem of model selection: what is the reasonable number of parameters that describes the transportation mechanism, but does not entail overfitting. Within deterministic framework this question cannot be easily answered. In principle, one can consider least square fitting as likelihood maximization with a special form of likelihood function $\sim \prod_{ij} \exp\left(-\text{error}_{ij}{}^2\right)$. This formally allows one to apply statistical model selection criteria, such as AIC. However, such an introduction of stochasticity is artificial.

More natural approach is to develop a stochastic gravity model and to apply information-based model selection criteria. In economical applications there are examples of stochastic gravity models (e.g. Flowerdew and Atkin, 1982; Sen and Smith, 1995). We decided to develop a portable stochastic gravity model suitable for ecological applications.

# 3. Stochastic gravity model and inference

## *3.1. The boater flow as a Poisson process*

We assume that individual boaters randomly decide to visit a certain pair of lakes, but probability of this decision depends on the lake characteristics. The total number of boaters using lakes *i* and *j*, $n_{ij}$ we describe by a Poisson distributed random variable,

$$n_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

We will apply gravity model of the form (2) for $\lambda_{ij}$, the mean number of boaters visiting one of the lakes after visiting another. As we have mentioned before, the results of the survey contain only the list of lakes visited by each boater, and therefore they do not distinguish $n_{ij}$ and $n_{ji}$. We shall assume that $n_{ij}$=$n_{ji}$. This means that our formulas have to be symmetric in $i$ and $j$.

As we have said above, in ecological applications it has been assumed that lake attractivity is determined by area of the lake, though the form of the dependency may be different. The same is true for the deterrence function. Most of the forms previously used can be combined in a single expression as follows,

$$W_i = \exp\left(a_1 \ln A_i + a_2 \ln\ln A_i + a_3 A_i / A_{\max}\right) = A_i^{a_1}\left(\ln A_i\right)^{a_2} \exp\left(a_3 A_i / A_{\max}\right), \quad (3)$$

$$\phi(d) = \exp\left(a_4 \ln d + a_5 d / d_{\max}\right) = d^{a_4} \exp\left(a_5 d / d_{\max}\right). \quad (4)$$

Similarly, one can add dependence on perimeter and any other available lake and region characteristics. Eventually this brings us to the framework of Poisson generalized linear models (GLM) with

$$\lambda_{ij} = \exp\left(a_0 + \sum a_k x_{ijk}\right), \quad (5)$$

where $a_k$ to be fitted to data $n_{ij}$, and $x_{ijk}$ represent either ln$d_{ij}$, or $d_{ij}/d_{\max}$, or a symmetric combinations of lake parameters like $\ln A_i + \ln A_j$ or $\left(P_i + P_j\right)/P_{\max}$. (See the Appendix for an example of the derivation.) We normalize distances, areas and perimeters to their maximum values to avoid situations when meaningful values of coefficients $a_k$ may be too small. The maximum likelihood fitting of such GLM models is a well-elaborated procedure that can implemented in R (Crawley, 2007).

### 3.2. Available covariates for the stochastic model

The covariates for a gravity model should reflect lake properties important for a traveler: convenience for boat launching and using, fishing, living, and so on. It is accepted that lake area is the most important covariate, though there may be others. Here we present the list of all characteristics we have tried. Others may be important as well and potentially may be used for gravity model improvement when available.

## Primary covariates

Lake parameters:

- Lake area $A_i$.
- Lake perimeter $P_i$.
- Lake coordinates, UTM easting and northing or longitude and latitude.
- Distance between two lakes $d_{ij}$. We have used distances along the roads and Euclidean distances between lake centroids. The results were practically indistinguishable. Since Euclidean distances are easier to obtain, we shall consider them as distance data.
- Elevation of lake surface above sea level.

Population data:

- 2006 population census data from Stats Canada for Ontario forward sortation areas (FSA2), corresponding to first two letters of the postal code, $\mathrm{Pop}_k$. We assume that the number of boaters in a certain region is proportional to the population of this region.
- Coordinates for the centroids of FSA2 and distances from *k*-th area to *i*-th lake, $l_{ki}$.

## Derived covariates

Our first experiments in fitting gravity models provided a number of covariate combinations that are statistically important but seem senseless from mechanistic point of view. (See Section 5 for motivation and more details.) Instead, it is more reasonable to use meaningful combinations of primary covariates as listed below.

- Normalized area $A_N$. According to our results, lake attractivity saturates for large lakes, and better attractivity measure is

$$A_{Ni} = \frac{A_i}{1 + A_i / A_0} = \frac{A_0 A_i}{A_i + A_0}, \quad A_0 \approx 3200\mathrm{km}^2. \qquad (6)$$

- Effective estimate of the number of boaters that may be attracted to a lake or boater “pressure”. We used a secondary gravity model to estimate boater pressure for lake *i* as

$$b_i = C_b \sum_{\substack{k \text{ over all} \\ \text{FSA2 areas}}} \mathrm{Pop}_k \times \phi(l_{ki}), \qquad (7)$$

where $\phi(l)$ is a distance deterrence function and $C_b$ is a normalization constant.

- "Fitted perimeter" $P_{Fi}$. If we try to fit a linear model for the dependency of lake perimeter on lake area as $\log_{10} P_i = c_0 + c_1 \log_{10} A_i + \varepsilon_i, \quad c_1 = 0.60 \pm 0.02$, this gives the expression for fitted perimeter:

$$P_{Fi} \approx 10^{c_0} A_i^{c_1}. \qquad (8)$$

## 3.3. Statistical inference

### Likelihood and parameter estimates

According to the form of our model, the log-likelihood function for the survey data is

$$l(\theta) = \sum_{ij} \ln\left( \frac{\lambda_{ij}^{\ n_{ij}}}{n_{ij}!} \exp(-\lambda_{ij}) \right) = \sum_{ij} \left( n_{ij} \ln(\lambda_{ij}) - \lambda_{ij} - \ln(n_{ij}!) \right), \qquad (9)$$

where $\lambda_{ij}$ is defined by equation (5) and $\theta$ denotes the set of covariates used in the model. For the fixed model structure we obtained ML estimates of the parameters and characterized model performance by AIC, BIC and cross-validation. Different sets of covariates were compared with the help of model selection criteria.

Whenever possible, we used R function `glm` for model fitting and parameter estimates (Crawley, 2007; R Development Core Team, 2009). Otherwise we used R routine `optim` for likelihood maximization.

### On model selection criteria

At present most popular model selection criteria seem to be AIC and BIC (Burnham and Anderson 2001, 2004; Ghosh and Samanta, 2001). Their performance has been analyzed by several authors in numerical experiments. In particular, it has been noted that there are two classes of problems: building the best predictor for the observed data and "discovering the truth" (Ghosh and Samanta, 2001), that is building a model that best reflects the structure of the original system which generated the data. It has been hypothesized that AIC is better for the former task while BIC is better for the latter. In Burnham and Anderson (2004), this observation has been stated differently: BIC is better for detecting simple true models while AIC is good when the true model is complicated. As we had already said in the beginning, our major task was to make a "portable" model that can be calibrated for one set of lakes and then used for the other. A simpler model has a better chance to satisfy our requirements, and hence such problem should be closer to "discovering the truth". This makes BIC the main candidate for model selection tool.

Burnham and Anderson (2001) note that the model with smaller AIC should not automatically be accepted as the best: AIC decrease by ~2 in practice often does not mean real improvement. Doubtless improvement requires $\Delta\mathrm{AIC} \geq 8$. Unfortunately, the authors did not specify, for what data set size their recommendations have been tested, and should they be adjusted with the size of the data set or not. From the expression for $\mathrm{BIC} = -2\ln L + k\ln n$ and our data set size of *n*=13790 lake pairs with $\ln n \approx 9.5$ it means that for formally better model with $\Delta\mathrm{BIC} > 0$ implies $\Delta\mathrm{AIC} > 7.5$. Applying Burnham and Anderson's idea to our case, we can say that if $\Delta\mathrm{AIC} \sim 10$ or $\Delta\mathrm{BIC} \sim 1$, then acceptance of a more complicated model is not automatic and may be debatable. Below we apply these selection criteria and accompany it with the estimates of residual sum of squares $\mathrm{RSS} = \sum_{ij} \left(n_{ij} - \lambda_{ij}\right)^2$.

Similarly, when we compare the models with <u>the same number of variables</u>, we choose one with the smallest BIC, and hence AIC. However, there may be other variable combinations giving models with only slightly higher BIC. We shall call such models “potential competitors”. If the BIC difference between competing models is small, with $\Delta\mathrm{BIC} \sim 10$, a possible explanation may be that there are correlations among the used covariates, and the true model must contain less covariates. For definiteness, we consider as potential competitors models with BIC difference less than 30.

**Cross-validation**

For validation purposes we applied the approach resembling "leave *k* out" algorithm. To keep the amount of computations reasonable we have taken *k* equal to 10% of the data. The brief description of the algorithm is as follows. (a) We randomly split the data set into 10 groups. (b) We exclude one group, fit the model to the remaining data, and then calculate log-likelihood and RSS for the excluded part. (c) This is repeated for each of 10 groups, then all likelihoods and RSS are added up to form the validated log-likelihood, RSS, AIC and BIC. (d) Calculations are repeated for 10 different data set groupings, and we calculated the mean validated AIC, BIC and RSS.

Validated BIC and RSS were used to verify BIC and RSS obtained during model fitting.

## 4. Data for fitting the gravity model

Our data came from existing surveys (Muirhead, 2007) within one of the mentioned above projects on *Bythotrephes* invasions in Ontario. The data for patterns of recreational travel were collected by the surveys. The surveys consisted of 17 questions designed to assess the risk of transporting aquatic invasive species during recreational boating and fishing. At the basic level, trip information was collected on lakes that were visited previously, including whether recreational boaters visited lakes known to be invaded by *Bythotrephes*, and the maximum distance they trailed a boat after leaving one of these lakes.

In July 2004 10,000 surveys were sent to owners of fishing licenses registered with the Ontario Ministry of Natural Resources (OMNR). 741 answers were received, and 394 of the responded owners kept their boat at a single lake and did not move to the others. Boaters who travel with the boat have provided lists of the lakes that they visited during their trips. We assume that trips between all lakes in each list are likely. Based on survey results, boaters visited 55 invaded (source) and 273 uninvaded (destination) lakes, Fig. 1. For this set of lakes there was the table of road distances between source and destination lakes. Due to invasion history, 49 out of 55 lakes were treated as both sources and destinations, that is were present in both categories. This leaves only

13790 unique distances in the table.  The remaining 24991 pairs of destination-destination and source-source lake pairs were of no interest from invasion point of view, there were no road distance data for them, and we did not consider these pairs.  Therefore we have considered 13790 lake pairs assuming symmetrical boater movement between these lakes, which provides sufficient amount of data for studying boater movement.  Processing of the survey data gave us the values of the observed number of travelers $n_{ij}$ for 13790 lake pairs. Only 787 of the lake pairs were used by at least one boater ( $n_{ij} > 0$ ), and only 34 of them were used by more than 4 boaters.  The most popular pair of lakes was used by 17 boaters.

Geospatial data for each lake was provided by the OMNR through the Ontario Geospatial Data Exchange.  In addition to latitude and longitude of the lake centroid, data on lake area (in km$^2$) and perimeter (km) were extracted through ArcGIS software (v 9.1, ESRI).  Road distance between lake pairs ($d_{ij}$) was calculated based on the minimum distance between lake centroids.

# 5. A Hitch-Hiker's guide to fitting stochastic gravity models

## *5.1. Best model using only basic lake characteristics available from GIS data*

Building a gravity model for Ontario lakes we have started with search for the best functional form for a model using the most popular and basic lake characteristics: areas *A* and distances between the lakes *d*. Lake perimeter *P* has never been used before, but it is as basic characteristic as area, and we have added it as well. Distances between the lakes can be estimated in two ways: as a path along the roads and as Euclidean distance between lake centers. The first way is more accurate but requires information about roads as well, while the second one requires only knowledge of the lake coordinates.  For a part of lake pairs we had road distances, obtained in (Muirhead, 2007), and in likelihood estimates we used only this subset of lake pairs to compare results for both types of distances. The results were practically identical, in most cases the relative difference in AIC or BIC was less than $10^{-4}$ .  To save space and simplify the description of the results, below we present only those for the distances between lake centers.

Since we had no *a priori* information, what should be the model dependence on $A$, $P$, and $d$, we used three functional forms for area, as in (3), and two functional forms for both distance and perimeter, as in (4). This gives 7 potential GLM covariates $x_{kij}$ in (5), named in Table 1. We tried all possible combinations of these covariates, that is 128=$2^7$ different models. We split them in 8 groups, according to the number of covariates used, and within each group we have selected the model with the minimum BIC.

The results of model selection are presented in Table 1. It shows, that up to 3 covariates there is significant decrease of BIC of 200 and more. Starting with 4 variables the BIC decreases only by values ~1, and the model form becomes less and less interpretable mechanistically. For this reason we selected the 3-covariate model with minimum BIC as the best one, which was

$$\lambda_{ij} = C\left(A_i A_j\right)^{\gamma} d_{ij}^{-\beta} \exp\left(-2\frac{A_i + A_j}{A_{\max}}\right), \quad \gamma = 0.53, \quad \beta = 1.19. \tag{10}$$

This function is in good agreement with the ideas of gravity models, except for the exponential term. It corresponds to lake attractivity of the form $W_i \sim A_i^{\gamma} \exp\left(-2A_i / A_{\max}\right)$, which shows, that lake attractivity attains its maximum for $A = 0.26A_{\max}$, and then decreases. This conclusion contradicts our understanding of attractivity mechanism. We suggest the following interpretation of the exponential term: lake attractiveness increases only up to a certain lake size and does not grow further. Very big and huge lakes have almost equal attractivity. Exponential term is the best possible approximation of this dependence within our set of models.

To verify this hypothesis, we introduced normalized lake area $A_N$ (6), such that for a very big lake normalized area tends to its maximum value $A_0$. To determine $A_0$, we did maximum likelihood fit to the 4-parameter model $\lambda_{ij} = C\left(A_{Ni} A_{Nj}\right)^{\gamma} d_{ij}^{-\beta}$. The resulting estimate of the limiting area is $A_0 \approx 3200\text{km}^2$. After that we have repeated the variable selection procedure replacing $A_i$ with $A_{Ni}$, and the best model is

$$\lambda_{ij} = C\left(A_{Ni} A_{Nj}\right)^{\gamma} d_{ij}^{-\beta}, \quad \gamma = 0.58, \quad \beta = 1.18. \tag{11}$$

It has almost the same BIC as (7), 5925 vs 5924, but an easy and transparent interpretation.

The cross-validation technique provided very close values of validated AIC/BIC to those given in Table 1 or cited above. For example, for model (11) the verified mean BIC was 5938. What is important, the ordering of the validated BIC values was the same as for standard BIC, and hence they led to the same conclusion. This was true for the cases below as well.

Therefore, model (11) is the best one using only basic lake parameters.

### 5.2. Model improvement with additional covariates

One can assume that lake attractivity for boaters may be related not only with lake area, but with some other characteristics as well. To check this, we have enhanced model (11) with more covariates available (see Section 3.2): geographic coordinates and elevation. Model selection procedure has shown that there was only one significant effect: use of latitude as a covariate decreases BIC from 5925 down to 5757. The only reasonable interpretation is that it is related with proximity of the lakes to the boater's residence. Therefore, model improvement requires knowledge of the boater home locations. We assumed that the number of boaters in a region is proportional to total population of the region. Information about population is available in 2006 census, and to describe potential number of visitors to a certain lake we have used the secondary gravity model (7), which uses coordinates for both lakes and FSA domains. For its distance deterrence function we used the same type of dependence as in (11), $\phi(l_{ki}) = l_{ki}^{-\nu}$. Since β is very close to 1, in test calculations we use ν=1.

The model (7) provides value proportional to potential number of visitors to lake *i*, $b_i$. Denoting lake attractivity by $W_i$, we may assume that the number of visitors to lake *i* is $\sim W_i b_i$, and the number of those who decide to visit lake *j* later that is $\sim W_i b_i W_j$. Similarly the number of visitors to lake *j*, who decide to go to lake *i* is $\sim W_j b_j W_i$. Therefore one can expect that $\lambda_{ij} \sim W_i W_j (b_i + b_j)$. We repeated the model selection procedure with elevation, coordinates, and one more covariate $b_{Aij} = b_i + b_j$. The best was a 3-variable model with BIC=5675 and cross-validated BIC=5687:

$$\lambda_{ij} = C(A_{Ni} A_{Nj})^{\gamma} d_{ij}^{-\beta} (b_i + b_j)^{\alpha}, \quad \gamma = 0.58, \quad \beta = 1.01, \quad \alpha = 1.36. \qquad (12)$$

Lake coordinates were not significant covariates any longer.

### 5.3. Final result: interpretation

Mechanistic interpretation of model terms is always an asset. Typically it is easier when the value that appears in the model has a simple dimension. It would be easy to interpret if the following would be true: a) $\gamma$=0.5, then $A^{\gamma}$ has the dimension of length and may represent lake radius or perimeter. b) $\beta=\nu=1$, then there is just inverse proportionality for distance. c) $\alpha$=1, then the number of travelers is proportional to the total population size. Otherwise, if the population increases, say, two times, the number of travelers should increase $2^{1.36}\approx 2.57$ times.

For this reason we have compared models that contain mean lake radius $R=\sqrt{A/2\pi}$ normalized by $R_0=\sqrt{A_0/2\pi}$ and some of the parameters fixed. The results are shown in Table 2. They allow us to make the following conclusions.

1) Using mean lake radius provides significantly worse results, therefore $\gamma\neq 0.5$.
2) The hypothesis $\beta=\nu=1$ is consistent with the data and so attractivity of lake pairs scales with inverse distance.
3) Setting $\alpha$=1 gives increase of BIC about 6, which does not allow us to completely reject the hypothesis that the number of travelers is proportional to population size.

For the value of $\gamma$ we can suggest the following explanation. It is very close to the value $c_1$ in (8), where lake perimeter is fitted to area. Therefore, we may assume that lake attractivity is proportional to the *estimated* or *expected lake perimeter*: the value that one may expect from glancing at the map. If this is true, then $\gamma$ is related with a psychological effect, and the best way to estimate attractivity is through the area, not actual perimeter.

Eventually we come to the best model

$$\lambda_{ij}=C\left(A_{Ni}A_{Nj}\right)^{\gamma}d_{ij}^{-1}\left(b_i+b_j\right)^{\alpha},\quad \alpha=1.37\pm 0.09,\ \gamma=0.58\pm 0.01,\ \nu=1. \qquad (13)$$

# 6. Discussion

The stochastic gravity models approach provides a number of benefits to the technical side of model building. It can be easily extended if additional covariates appear, and it may allow for more general modeling approaches with new research and management goals.

From the technical side, most important features are more realistic data fitting for rarely used lake pairs and applicability of information-based model selection criteria for choosing the optimal set of covariates and functional forms for them. These features allowed us to build a portable gravity model. Portability may be an important property of gravity models, and not all models possess it, due to data requirements and constraints under which the gravity models are parameterized. For example, the most highly constrained type of gravity model, the doubly-constrained gravity model, requires data on both outflows from invaded sources and inflows into invaded and noninvaded destinations to parameterize the model (Haynes and Fotheringham, 1984). Portability is thus limited, as the pairwise number of trips between lakes *i* and *j* must be re-calculated as new lakes are added to the data set. Unconstrained (this study) and production-constrained gravity models, on the other hand, are well suited for invasion risk management where propagule pressure estimates to each lake *j*, $\sum_i \lambda_{ij}$ , can be easily recalculated for lakes not covered by previous surveys. In production-constrained gravity models, the sum of pairwise trips for each source lake *i* must equal to data on the observed number of trips leaving each source, but pairwise trips to lakes from different data sources requires data on only measures of attraction, such as lake area (e.g. Siderelis and Moore, 1998, Leung et al., 2006) and distance between source and destination lakes.

In the final relation for the stochastic gravity model (13) we do not present the value of the factor *C*. Likelihood maximization with respect to it (or $a_0$ in (5)) results in the constraint on the total number of trips: $\sum_{ij} \lambda_{ij} = \sum_{ij} n_{ij}$ , and this equality defines *C*. Therefore, absolute values of $\lambda_{ij}$ are still tied to the survey region and survey data. However, it is natural to assume that other coefficients, $\alpha$, $\beta$, $\gamma$, $\nu$, and consequently the relative $\lambda_{ij}$ values are associated with the principles of boater’s decision making, and hence can be used at other regions where the boaters behavior is expected to be similar. In some cases the unknown constant *C* can be estimated for the new region through the total number of boaters. In other cases, e.g. in fitting a binary classifier to presence/absence data, $\lambda$ is used with a factor to be fitted, and hence the value of *C* is not important. In other words, the gravity model (13) appears to be portable from one region to another.

Stochastic gravity models, fitted by maximum likelihood estimation, have some advantages over their deterministic counterparts because it becomes possible to test hypotheses concerning additional covariates to $\lambda_{ij}$ and to simplify the functional form of the model if the covariates are not significantly different from zero. In our final model (13), the number of boat trips arriving at a lake depends on the population density of recreationalists surrounding the lakes (FSA2) and the distance of these population centers to each of the lakes. Although data was not available on the absolute numbers of recreationalists that travel from specified regions (e.g. Bossenbroek et al., 2001), here we are assuming that the proportion of the population that trailers boats is the same across the FSA zones.

When an invader spreads beyond the areas covered by surveys, portable gravity models can be applied there. We are now using the model developed in this paper to determine the most important covariates for lake invasibility by *Bythotrephes* (Potapov et al. 2010) in 306 lakes not included in Muirhead's 2004 survey (Section 4). Using the relative mean boater flows (13), we have calculated total relative inflows of propagules into each lake, which appeared to be significantly better single predictor of invasions than the next best one, lake pH, with $\Delta$AIC=13.

The stochastic gravity model approach can be easily extended if more covariates are available for the lakes covered by survey. Covariates influencing lake attractivity may include the presence of game fish species in the lake, water clarity, accessibility of the lake, as well as a suite of socio-economic factors. Interpretation of these factors in terms of perceived attractiveness is not always straightforward, however. Similar to other gravity models of recreational boater movement (Bossenbroek et al., 2001, Leung et al., 2006), we used lake area as a measure of lake attractivity. Lake area may also be confounded as a primary measure of attractiveness since area is also correlated with highly desired attributes such as the number of boat launches and availability of support facilities (Siderelis and Christos, 1998).

The approach may be also extended by adding submodels of individual-decision making process such as a Random Utility Model. The probability of choosing one destination over another is weighted by the cost to travel to those lakes, which is then nested as input into the gravity model (Siderelis and Moore, 1998).

Although the number of boater trips between lakes was used as a measure of propagule pressure in this study, data may be collected from field experiments on actual number of invasive propagules per trip. Propagule loads may be quantified for vectors associated with recreational boating such as the number of individual propagules collected from contaminated fishing lines, bait buckets, bilge water, etc. Statistical techniques such as likelihood ratio testing may be used to test the significance or relative importance of each vector in the transportation pathway of recreational traffic. Such an approach would correspond with the EPPO's (2007) recommendations to assess 1) the ease at which invasive propagules may be detected within specific vectors, 2) what is the distribution pattern of the vector with respect to destinations.

The stochastic gravity model approach may also be extended by trying other statistical distributions beyond the Poisson for the number of trips from one lake to another. As a result, statistical techniques may improve quality of gravity models in ecology. This is an area of ongoing research.

Stochastic gravity models may allow to set up new modeling problems as well. Given that the mean number of boater flows between lakes follows a statistical distribution and maximum likelihood model fitting requires casting a gravity model in a probability mass function, a series of probability-based management scenarios may be explored. Such a framework has been recommended frequently for invasive species management (Maguire, 2004). For example, a scenario may ask: given an expected mean number of boat trips between lakes from the model, what is the probability of observing at least one trip between specific invaded and non-invaded lakes. For invasive species that reproduce clonally such as zooplankton (e.g. *Bythotrephes)* or by vegetative fragmentation (e.g. Eurasian watermilfoil, *Myriophyllum spicatum*), small inocula size may be sufficient to establish a population providing environmental conditions are suitable (Drake et al., 2006). The number of individuals of sexually-reproducing species required to establish a population, however, is likely 2-3 times higher in order of magnitude, with many more required for populations that experience Allee effects due to low mate densities and to withstand environmental variability (Leung et al., 2004; Lodge et al., 2006).

Stochastic gravity models can provide both estimates of the boat traffic and information about its uncertainty arising from stochastic nature of the process errors in estimated model parameters. Availability of traffic estimates allows managers to assess the invasion risks and to make decisions, for example, about location and number of boater treatment stations, education posters etc. Uncertainty of predictions also may be important for better choice of management decisions (Regan et al., 2005; Cooney and Lang, 2007).

In this paper we have proposed to use stochastic gravity models for modeling movement of boaters, and have developed an example of such a model based upon J. Muirhead 2004 survey. This approach naturally integrates randomness and uncertainty, and uses statistical model selection instead of constraints to improve the prediction accuracy. We hope that our model may be useful in further ecological studies and management applications. At the same time, there is a lot of room for model improvement in the statistical framework.

## *Acknowledgements*

This research has been supported by Canadian Aquatic Invasive Species Network, by NSERC and by a Canada Research Chair (MAL).

# Appendix. An example of GLM form of stochastic gravity model.

Let the attractivities $W_i$ and $W_j$ have the form (3), distance deterrence function $\phi(d)$ has the form (4), and we are building a gravity model of the form (2). Then taking logarithm of $\lambda$ we obtain

$$\ln\lambda_{ij} = \ln C + \ln W_i + \ln W_j + \ln\phi(d_{ij}).$$

Denoting $a_0$=ln$C$ and substituting (3) and (4), we have

$$\begin{aligned}\ln\lambda_{ij} &= a_0 + (a_1\ln A_i + a_2\ln\ln A_i + a_3 A_i / A_{\max}) + \\ &+ (a_1\ln A_j + a_2\ln\ln A_j + a_3 A_j / A_{\max}) + (a_4\ln d_{ij} + a_5 d_{ij} / d_{\max}) = \\ &= a_0 + a_1(\ln A_i + \ln A_j) + a_2(\ln\ln A_i + \ln\ln A_j) + \\ &+ a_3(A_i / A_{\max} + A_j / A_{\max}) + a_4(\ln d_{ij}) + a_5(d_{ij} / d_{\max}).\end{aligned}$$

Now let us denote

$$x_{1ij} = \ln A_i + \ln A_j, \quad x_{2ij} = \ln\ln A_i + \ln\ln A_j,$$
$$x_{3ij} = A_i / A_{\max} + A_j / A_{\max}, \quad x_{4ij} = \ln d_{ij}, \quad x_{5ij} = d_{ij} / d_{\max}.$$

In this notation

$$\ln\lambda_{ij} = a_0 + a_1 x_{1ij} + a_2 x_{2ij} + a_3 x_{3ij} + a_4 x_{4ij} + a_5 x_{5ij},$$

or

$$\lambda_{ij} = \exp(a_0 + a_1 x_{1ij} + a_2 x_{2ij} + a_3 x_{3ij} + a_4 x_{4ij} + a_5 x_{5ij}).$$

This is the expression for Generalized Linear Model (GLM), Eq. (5). Therefore, representing the values of the lake covariates as their symmetric combinations $x_{kij}$ allows us to apply standard R routine `glm` for fitting a stochastic gravity model to survey data.

# References

Baxter, M., Ewing, G., 1981. Models of recreational trip distribution. Reg. Stud. 15, 327-344.

Bobeldyk, A.M., Bossenbroek, J.M., Evans-White, M.A., Lodge, D.M., Lamberti, G.A., 2005. Secondary spread of zebra mussels (*Dreissena polymorpha*) in coupled lake-stream systems. Ecoscience. 12, 339-346.

Boudreau, S. A. and Yan, N.D., 2003. The differing crustacean zooplankton communities of Canadian Shield lakes with and without the nonindigenous zooplanktivore, *Bythotrephes longimanus*. *Canadian Journal of Fisheries and Aquatic Science.* 60(11): 1307-1313.

Bossenbroek, J.M., Kraft, C.E., Nekola, J.C., 2001. Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes, Ecol. Appl. 11, 1778—1788

Bossenbroek, J.M., Johnson, L.E., Peters, B., Lodge, D.M., 2007. Forecasting the expansion of zebra mussels in the United States. Cons. Biol. 21, 800-810.

Burbidge, A.A., Manly, B.F.J., 2002. Mammal extinctions on Australian islands: causes and conservation implications. J. Biogeogr. 29, 465-473.

Burnham, K.P., Anderson, D.R., 2001. Model selection and multimodel inference. A practical information-theoretic approach. 2nd ed., Springer, NY, USA, 488 pp.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Method. Res. 33, 261-304.

Cooney, R., Lang, A.T.F. 2007. Taking uncertainty seriously: adaptive governance and international trade. Eur. J. Intern. Law, 18(3)m 523-551.

Crawley, M.J., 2007. The R book. Wiley, USA, 942 pp.

Crowl, T.A., Crist, T.O., Parmenter, R.R., Belovsky, G., Lugo, A.E., 2008. The spread of invasive species and infectious disease as drivers of ecosystem change. Front. Ecol. Environ. 6, 238-246.

Drake, J.M., Drury, K.L.S., Lodge, D.M., Blukacz, A., Yan, N.D., Dwyer, G, 2006. Demographic stochasticity, environmental variability, and windows of invasion risk for *Bythotrephes longimanus* in North America. Biol. Invasions. 8, 843-861.

EPPO, 2007. Guidelines on pest risk analysis: Decision-support scheme for quarantine pests PM513(3). European and Mediterranean Plant Protection Organization, Paris, France. http://archives.eppo.org/EPPOStandards/PM5_PRA/PRA_scheme_2007.doc

Ferrari MJ, Bjørnstad ON, Partain JL, Antonovics J., 2006. A gravity model for the spread of a pollinator-borne plant pathogen. Am Nat. 168(3), 294-303.

Flowerdew, R., Atkin, M., 1982. A method of fitting the gravity model based on the Poisson distribution. J. Regional Sci. 22, 191-202.

Ghosh, J.K., Samanta, T., 2001. Model selection - an overview. Curr. Sci. 80, No. 9-10, 1135-1144.

Haynes, K.E., Fotheringham, A.S., 1984. Gravity and spatial interaction models. Sage Publications, Beverly Hills, London, New Delhi, 88 pp.

Johnson, L.E., Padilla, D.K., 1996. Geographic spread of exotic species: Ecological lessons and opportunities from the invasion of the zebra mussel *Dreissena polymorpha*. Biol. Conserv. 78, 23-33.

Keller, R.P., Lodge, D.M., Lewis, M.A., Shogren, J.F. (eds.), 2009. Bioeconomics of invasive species. Oxford University Press, NY,USA, 298 pp.

Kraft, C.E., Sullivan, P.J., Karatayev, A.Y., Burlakova, L.E., Nekola, J.C., Johnson, L.E., Padilla, D.K., 2002. Landscape patterns of an aquatic invader: assessing dispersal extent from spatial distributions. Ecol. Appl. 12, 749-759.

Leung, B., Bossenbroek, J.M., Lodge, D.M., 2006. Boats, pathways, and aquatic biological invasions, estimating dispersal potential with gravity models. Biol. Invasions 8, 241-254.

Leung, B., Drake, J.M., Lodge. D.M., 2004. Predicting invasions: propagule pressure and the gravity of Allee effects. *Ecology* 85:1651-1660.

Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., Carlton, J.T., McMichael, A., 2006. Biological invasions: Recommendations for US policy and management. Ecol. Appl. 16, 2035-2054.

Lovell, S., Stone, S.F., Fernandez, L., 2006. The econonic impacts of aquatic invasive species: a review of the literature. Agr. Res. Econ. Rev. 35, 195-208.

MacIsaac, H.J., Borbely, J.V.M., Muirhead, J.R., and Graniero, P.A., 2004. Backcasting and forecasting biological invasions of inland lakes. Ecol. Appl. 14(3): 773–783. doi:10.1890/02-5377.

Maguire, L.A., 2004. What can decision analysis do for invasive species management? Risk Anal. 24, 859-868.

Muirhead, J.R., 2007. Forecasting dispersal of nonindigenous species. PhD thesis, University of Windsor, Windsor, Canada.

Parker, I.M., Simberloff, D., Lonsdale, W.M., Goodell, K., Wonham, M., Kareiva, P., Williamson, M.H., Von Holle, B., Moyle, P.B., Byers, J.E., Goldwasser, L., 1999. Impact: toward a framework for understanding the ecological effects of invaders. Biol. Invas. 1, 3-19.

Potapov A., Muirhead J., Yan N., Lele S., Lewis M.A. Models of lake invasibility by Bythotrephes longimanus, a non-indigenous zooplankton. Biological Invasions (submitted 2010).

Regan, H.M., Ben-Haim, Y., Langford, B., Wilson, W.G., Lundberg, P., Andelman, S.J., Burgman M.A. 2005. Robust decision-making under severe uncertainty for conservation management. Ecol. App. 15(4), 1471-1477.

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

Schneider, D.W., Ellis, C.D., Cummings, K.S., 1998. A transportation model assessment of the risk to native mussel communities from zebra mussel spread. Conserv. Biol. 12, 788-800.

Sen, A., Smith, T.E., 1995. Gravity Models of Spatial Interaction Behavior. Springer, Berlin, New York, 572 pp.

Siderelis, C., Moore, R.L., 1998. Recreation demand and the influence of site preference variables. J. Leisure Res. 30, 301-318.

Thomas, R.W., Huggett, R.J., 1980. Modeling in geography: a mathematical approach. Rowman and Littlefield, Lanham, MD, USA, 338 pp.

Yan, N.D., Girard, R., Boudreau, S., 2002. An introduced invertebrate predator (*Bythotrephes*) reduces zooplankton species richness. Ecol. Lett. 5, 481-485.

**Table 1**. Variable selection for Poisson generalized linear model (5) with the number of covariates $x_{kij}$ in the sum varying from 0 to 7. For each number of variables the model with the minimum BIC value has been selected. The corresponding AIC, BIC, and RSS values and the values of the coefficients $a_k$ in the sum are shown in respective columns. The potentially competing models are those with BIC not exceeding the best BIC+30. The best 3-covariate model is chosen as a result of model selection (bold).

| # of variables | Best BIC | AIC | RSS | $a_k$ | | | | | | | other models | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ln$d_{ij}$ | $d_{ij}/d_{max}$ | ln$A$ | lnln$A$ | $A/A_{max}$ | ln$P$ | $P/P_{max}$ | # of competing models | Next Best AIC | Next Best BIC |
| 0 | 9572 | 9564 | 3915 | | | | | | | | 0 | | |
| 1 | 7559 | 7544 | 3393 | | | 0.30 | | | | | 0 | 7578 | 7593 |
| 2 | 6144 | 6121 | 2466 | -1.17 | | 0.42 | | | | | 0 | 6320 | 6343 |
| **3** | **5924** | **5894** | **2097** | **-1.19** | | **0.53** | | **-2.0** | | | **0** | **6049** | **6079** |
| 4 | 5920 | 5883 | 2060 | -1.18 | | 0.56 | | -1.9 | | -0.4 | 5 | 5886 | 5924 |
| 5 | 5915 | 5870 | 2114 | -1.19 | | 0.66 | -0.37 | -2.2 | | -0.6 | 8 | 5879 | 5924 |
| 6 | 5909 | 5856 | 2101 | -1.19 | | 0.66 | -0.61 | -2.0 | 0.17 | -1.1 | 4 | 5867 | 5920 |
| 7 | 5914 | 5854 | 2087 | -1.12 | -0.7 | 0.66 | -0.60 | -2.0 | 0.17 | -1.0 | 0 | | |

**Table 2**. Testing hypotheses about model parameters. We fix part of the model coefficients according to the tested hypothesis and calculate the difference of the BIC for the tested model and the model (12) (row 4 here). The results show that hypothesis that ν=β=1 agrees with the data. The corresponding model (row 6) is reported as the best one, see (13). This model has the minimum AIC as well.

| # | Model<br>$\lambda_{ij}$ | Fixed coefficients | Fitted coefficients | AIC | BIC | ΔBIC |
|---|---|---|---|---|---|---|
| 1 | $CR_{Ni}R_{Nj}d_{ij}^{-1}(b_i+b_j)$ | ν=1, β=1, α=1 | | 5735 | 5743 | 68 |
| 2 | $CR_{Ni}R_{Nj}d_{ij}^{-\beta}(b_i+b_j)$ | ν=1, α=1 | β=0.933±0.026 | 5730 | 5745 | 70 |
| 3 | $C(A_{Ni}A_{Nj})^{\gamma}d_{ij}^{-\beta}(b_i+b_j)$ | ν=1, α=1 | β=1.05±0.03<br>γ=0.59±0.01 | 5659 | 5682 | 7 |
| 4 | $C(A_{Ni}A_{Nj})^{\gamma}d_{ij}^{-\beta}(b_i+b_j)^{\alpha}$ | ν=1 | α=1.36±0.09<br>β=1.01±0.03<br>γ=0.58±0.01 | 5645 | 5675 | 0 |
| 5 | $C(A_{Ni}A_{Nj})^{\gamma}d_{ij}^{-\beta}(b_i+b_j)^{\alpha}$ | | α=1.91±0.03<br>β=0.99±0.04<br>γ=0.58±0.01<br>ν=0.74±0.20 | 5645 | 5683 | 8 |
| 6 | $C(A_{Ni}A_{Nj})^{\gamma}d_{ij}^{-1}(b_i+b_j)^{\alpha}$ | ν=1, β=1 | α=1.37±0.09<br>γ=0.58±0.01 | 5643 | 5663 | -12 |

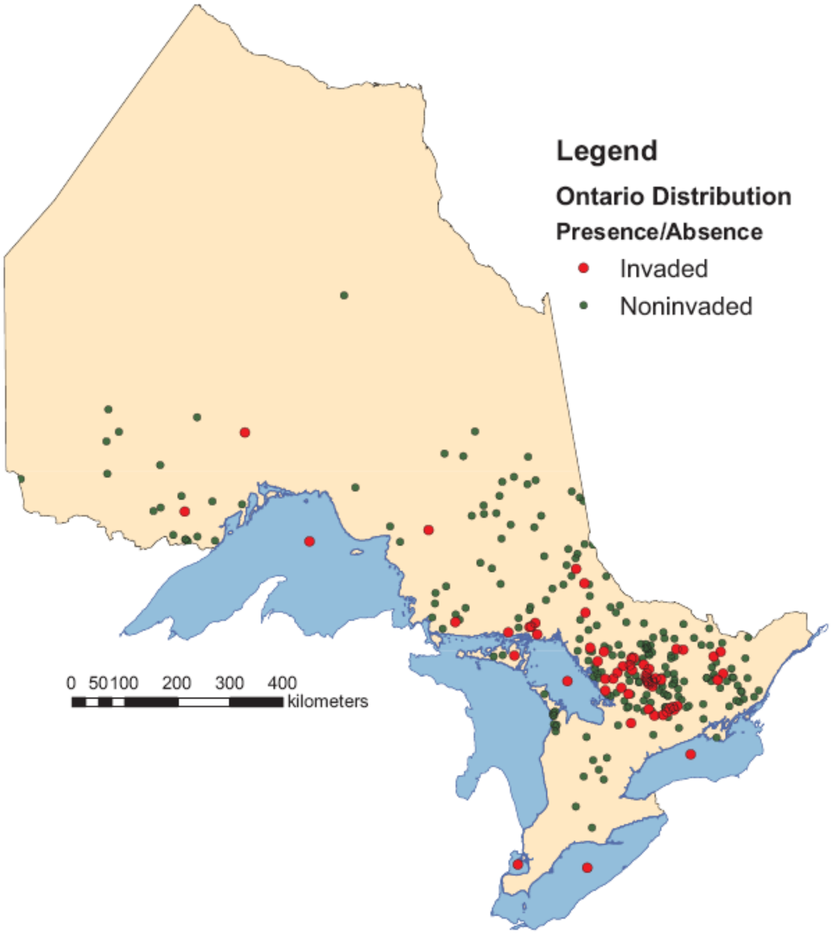

Fig. 1. Ontario lakes covered by J. Muirhead's survey and used for creating our stochastic gravity model.