

RESEARCH ARTICLE

Uncertainty analysis of species distribution models

Xi Chen^{1*}, Nedialko B. Dimitrov¹, Lauren Ancel Meyers^{2,3}

1 Graduate Program in Operations Research Industrial Engineering, The University of Texas at Austin, Austin, Texas, United States of America, **2** Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America, **3** The Santa Fe Institute, Santa Fe, New Mexico, United States of America

* carol.chen@utexas.edu



Abstract

The maximum entropy model, a commonly used species distribution model (SDM) normally combines observations of the species occurrence with environmental information to predict the geographic distributions of animal or plant species. However, it only produces point estimates for the probability of species existence. To understand the uncertainty of the point estimates, we analytically derived the variance of the outputs of the maximum entropy model from the variance of the input. We applied the analytic method to obtain the standard deviation of dengue importation probability and *Aedes aegypti* suitability. Dengue occurrence data and *Aedes aegypti* mosquito abundance data, combined with demographic and environmental data, were applied to obtain point estimates and the corresponding variance. To address the issue of not having the true distributions for comparison, we compared and contrasted the performance of the analytical expression with the bootstrap method and Poisson point process model which proved of equivalence of maximum entropy model with the assumption of independent point locations. Both Dengue importation probability and *Aedes aegypti* mosquito suitability examples show that the methods generate comparatively the same results and the analytic method we introduced is dramatically faster than the bootstrap method and directly apply to maximum entropy model.

OPEN ACCESS

Citation: Chen X, Dimitrov NB, Meyers LA (2019) Uncertainty analysis of species distribution models. PLoS ONE 14(5): e0214190. <https://doi.org/10.1371/journal.pone.0214190>

Editor: Shyamal D Poddada, University of Pittsburgh Graduate School of Public Health, UNITED STATES

Received: January 8, 2017

Accepted: March 9, 2019

Published: May 23, 2019

Copyright: © 2019 Chen et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This publication was supported by Cooperative Agreement Number U01CK000512, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

Introduction

Species distribution models [1–3] are commonly used to predict the geographic distributions of animals or plants species. They are applied in species conservation [4], ecology [5], and other fields. Some SDMs, like the maximum entropy model, are used to predict the probability for the species being present. Others, like Poisson point process models, are used to model the intensity of the species per unit area.

Quantifying the uncertainty of maximum entropy models can help biologists allocate sampling efforts more efficiently. For places with the same probability estimate, different uncertainty estimates can help differentiate the need for further sampling effort. It may be possible to lower uncertainties in the estimates by choosing sampling locations carefully. However, the independence between sample units need to be guaranteed to maintain the independence

Competing interests: The authors have declared that no competing interests exist.

assumptions underlying a maximum entropy model. Quantifying the uncertainty also helps biologists have an idea of the amount of data sufficient to estimate probabilities across the geographic area. Knowledge of uncertainty can help answer questions such as: What is the benefit of collecting an additional 1000 presence only data points? What are low and high scenarios for the output estimates?

Unfortunately, most SDM methodology focuses on using point estimates. Point estimation involves using a single value for estimating target population parameters from sample data. However, the estimations are usually not equal to the target population parameters exactly, and so the accuracy of the estimations is important. A well accepted method of describing the uncertainty of the estimations is to look at their variance. With the variance of the estimates, one can compute confidence intervals, an interval that contains the true parameter with a certain confidence [6]. With current practice, SDMs only produce point estimates for the predicted probability or intensity at all species locations and background points without any corresponding uncertainty estimates at these locations.

To address this lack of uncertainty quantification in SDM, one must refer both to the SDM methodology and statistical methodology in quantifying uncertainty of point estimates. One of the most popular methods for SDM is the maximum entropy model. The conventional maximum entropy model was first formulated by Jaynes in 1957 [7] based on Shannon's measure of entropy [8] (see details in [9]). MAXENT incorporating the effect of actual occurrence data, became popular among biologists in modeling species distribution with the contribution of MaxEnt software [1, 10, 11]. The mathematical equivalence of MAXENT, model used in MaxEnt software, and Poisson point process models (Poisson PPMS has been shown in [12]. Poisson PPMS may be fitted in the 'spatstat' package in R, which provides a way of assessing model uncertainty by providing standard error estimates [13]. To quantify uncertainty in point estimates, bootstrap methods are popular. Bootstrap uses computer-intensive simulation to calculate standard deviations of the estimated parameters, and is broadly applied in the biology field [14–17]. In this paper, we adopt the maximum entropy method and compare the analytical expression of the standard deviation with the standard deviation calculated through bootstrap method and Poisson point process model (PPM) approach.

In this article we consider quantifying the uncertainty in SDM. We focus specifically on the maximum entropy SDM methodology. A significant reason for the popularity of the maximum entropy methodology is its applicability to presence-only data with least assumptions [18]. For traditional statistical estimation methods like regression, both of the presence and absence of the species are required. However, in real cases, biologists often only know the places a species has been observed, while lacking information about absences of species.

Our main contribution is analytically deriving an expression of the standard deviation of the target species distribution probabilities and comparing the results with bootstrap methods and standard deviation calculated through Poisson PPM approach. We show that the three methods generate comparatively the same results and our analytic model uncertainty calculation procedure is dramatically faster than the bootstrap method and more proper comparing to Poisson PPM without independence assumption and provided a direct result to maximum entropy model.

Materials and methods

Maximum entropy model

Consider a region with geographic divisions given by $X = \{x_1, x_2, \dots, x_n\}$. Suppose some species lives in the region, and the fraction of the species that lives in division i is p_i . A basic goal in SDM is to reconstruct the geographic distribution $P = \{p_1, p_2, \dots, p_n\}$. To do this, we have

some species occurrence data $O = \{o_1, o_2, \dots, o_n\}$, where each o_i specifies the number of times the species has occurred in division i . The occurrence data can be viewed as a sample from the distribution P . In addition we are given k layers of environmental data for the region described by features $f_j(X)$ for $j = 1, \dots, k$. For example, one such function could be the average elevation in each geographic division.

Jaynes' maximum entropy model attempts to reconstruct P . Let $\hat{P} = \{\hat{p}_1, \hat{p}_2, \dots, \hat{p}_n\}$ be the reconstructed density. Let $\hat{E}(f_j(X)) = (\sum_{i=1}^n o_i f_j(x_i)) / (\sum_{i=1}^n o_i)$ be the empirical estimate of $E_p(f_j(X)) = \sum_{i=1}^n p_i f_j(x_i)$ given by the occurrence data O . Jaynes' maximum entropy model attempts to reconstruct P through an optimization problem. The optimization uses Shannon's measure of entropy as the objective (1), subject to the moment constraints (2). Constraints (3) and (4) ensure that the optimal solution for the optimization is a probability distribution [9]. A mathematical formulation of the maximum entropy problem is

$$\max_{p_i} - \sum_{i=1}^n p_i \log p_i \tag{1}$$

$$\text{s.t. } \sum_{i=1}^n p_i f_j(x_i) = \hat{E}(f_j(X)) \quad j = 1, \dots, k \tag{2}$$

$$\sum_{i=1}^n p_i = 1 \tag{3}$$

$$p_i \geq 0 \quad i = 1, \dots, n \tag{4}$$

Bootstrap method

Table 1 describes the bootstrap method for estimating the uncertainty of the estimate resulting from maximum entropy. The core of this bootstrap procedure is thinking of the distribution P as parameterized by the values it assigns to each geographic division. The procedure starts by estimating the parameters once, yielding a probability distribution. Then, it samples the data from that estimated distribution to construct several new estimates.

Analytic deduction of uncertainty

In this section, we demonstrate the basic idea of the analytical method for quantifying uncertainty in maximum entropy. The data $O = \{o_1, o_2, \dots, o_n\}$ follow a multinomial distribution with unknown parameters P . A maximum likelihood estimator for P follows a certain multivariate normal distribution as the number of samples grow large. The maximum entropy model can be viewed as a function mapping this estimator to \mathcal{R}^n . The input is the empirical expectations, $\hat{E}(f_j(X))$, derived from the observation data, $O = \{o_1, o_2, \dots, o_n\}$. The output is the estimate of the probability distribution over geographic regions, $P = \{p_1, p_2, \dots, p_n\}$. The analytical method of quantifying uncertainty describes how the output, P , changes as the input, O , changes. This is essentially a quantification of the way the optimization mapping warps the data input space, to the output space. We show the detailed deduction of the analytic method for uncertainty in the S1 Appendix.

For brevity, let $a_j = \hat{E}(f_j(X))$ and the vector of a_j can be expressed as $A = (a_1, a_2, \dots, a_k)^T$. Let $g(A)$ denote the maximum entropy optimization, model (1),(2),(3),(4), as a function from \mathcal{R}^k

Table 1. Bootstrap method.

Algorithm	Bootstrapping
1	function Bootstrapping (N)
2	$\hat{P} = M(O)$
3	For $i = 1: N$ do
4	$\hat{O}_n = B(\hat{P}, m)$
5	$P' = M(\hat{O}_n)$
6	Record P'
7	Return $SD(P', N)$
N	Repeat the procedure N times
$O = \{o_1, o_2, \dots, o_n\}$	Original occurrence data
$M(O_n)$	Fit a maximum entropy model given a set of species occurrence data $O_n = \{o_1, o_2, \dots, o_n\}$ and return probability density estimation \hat{p}
\hat{P} and P'	A reconstructed density over the geographic region
$B(\hat{p}, m)$	Sample m occurrence data following probability density \hat{p} , where $m = \sum_{i=1}^n o_i$
$\hat{O}_n = \{\hat{o}_1^n, \hat{o}_2^n, \dots, \hat{o}_n^n\}$	The n^{th} new sampled occurrence data with $\sum_{i=1}^n \hat{o}_i^n = \sum_{i=1}^n o_i$
$SD(P', N)$	Calculate standard deviation of the set of P' s

<https://doi.org/10.1371/journal.pone.0214190.t001>

to \mathcal{R}^n . In other words, the function takes as input the vector A with j^{th} entry specified by a_j , specifying right hand sides of the equality constraints $\hat{E}(f_j(X))$, and outputs a probability estimate across the geographic region P . We would like to understand the uncertainty in the output $g(A)$ as a function of the uncertainty of the input A . This can be done following steps similar to those in the delta method [19, p.75].

To understand the uncertainty in the output $g(A)$, we begin by writing a first order Taylor expansion of g around $E(A)$

$$g(A) \approx g(E(A)) + \nabla g(E(A)) \cdot [A - E(A)] \approx g(\mathbf{F} \cdot \hat{P}) + \nabla g(\mathbf{F} \cdot \hat{P}) \cdot [A - E(A)], \tag{5}$$

where \mathbf{F} is $k \times n$ matrix of k features with entry (i, j) specified by $f_i(x_j)$ and $\nabla g(\cdot)$ is an $n \times k$ matrix of partial derivatives, with entry (i, j) specified by $\frac{\partial p_i}{\partial a_j}$. If we can compute an expression for these partial derivatives, then everything on the right hand side above is constant, except $[A - E(A)]$ whose distribution we know because we know the distribution of A . $g(A)$ is an affine transformation of $[A - E(A)]$, and can be approximated as

$$g(A) \sim \text{Normal}(g(\mathbf{F} \cdot \hat{P}), \nabla g \cdot \frac{\mathbf{F} \cdot \Sigma \cdot \mathbf{F}^T}{m} \cdot (\nabla g)^T), \tag{6}$$

where Σ is proportional to the covariance matrix of \hat{P} with entry (i, j) specified by $-\hat{p}_i \hat{p}_j$ for $i \neq j$, and entry (i, i) specified by $\hat{p}_i(1 - \hat{p}_i)$.

We express the $\frac{\partial p_i}{\partial a_j}$ as (Detailed deduction shown in [S1 Appendix](#))

$$\frac{\partial p_i}{\partial a_j} = \sum_{r=1}^k p_r (a_r - f_r(x_i)) ((-\Psi)^{-1})_{rj}, \tag{7}$$

where $\Psi_{rj} = \text{cov}_P(f_r, f_j)$ is the covariance matrix of features with respect to the maximum entropy model results, and f_j denotes the j^{th} feature in constraint (2). We denote the inverse covariance matrix as Ψ^{-1} and refer to its (r, j) th entry as $(\Psi^{-1})_{rj}$.

To summarize, one can compute analytical estimates of the uncertainty as follows:

1. Gather data for $f_r(\cdot)$ and the right-hand sides of constraints (2), a_r .
2. Solve the maximum entropy model to get a vector of P of probabilities p_i .
3. Compute the matrix $-cov_P(f_r, f_j)$, using the features and the vector P .
4. Compute the derivatives $\frac{\partial p_i}{\partial a_j}$ using (7), giving the matrix ∇g .
5. The covariance of the output P can then be estimated as $\nabla g \cdot \frac{F \cdot \Sigma \cdot F^T}{m} \cdot (\nabla g)^T$, following Eq (5).

Results

We demonstrate the applications of the analytical expression of the uncertainty through two examples, Dengue virus and *Aedes Aegypti* mosquito, and compare the analytical results with the uncertainty calculated using the bootstrap method and Poisson PPM approach. The analytic method results aligned well with bootstrap method results, but Poisson PPM approach gave much larger standard deviations. We only show the results and comparison of analytic and bootstrap below but include results and comparison of Poisson PPM in [S1 Figs](#). The resolution of the Dengue virus example is at county level while the resolution of the *Aedes Aegypti* mosquito is at 1 km^2 area level through Texas.

Dengue importation probabilities

Dengue virus is often imported into Texas from endemic counties. We aim to estimate the probability that the next importation case will happen in each county of Texas. Historical case import data, $O = \{o_1, o_2, \dots, o_n\}$ with n equal to 254 counties in Texas, present empirical samples from this distribution. Each o_i counts the number of imports in county i . We are also given features $f_j(X) \in R^{1 \times 254}$ for $j = 1, \dots, 10$ that represent socio-economic, demographic, and environmental features selected for all 254 counties across the Texas counties. This completely defines the inputs necessary for a maximum entropy model.

Specifically, we use ten years, 2002 to 2012, of Dengue importation data into Texas received from the Texas Department of State Health Services. The features $f_j(X)$ represent features listed in [Table 2](#). The ten final features were selected through a series feature selection procedures, including representative variable selections and most predictive variable selections, which

Table 2. Ten features included in maximum entropy model. The data for these features is derived at a county level from the 2009-2013 American Community Survey 5-year estimates [21] and WorldClim Database [22].

Features
Population of Educational Attainment with Bachelor's degree
Minimum Temperature of Coldest Month
Percentage of Using Public Transportation to Work
Population of Educational Attainment in some college(no degree)
Population of Walked to Work
Population of Commuting to Work with Other Means
Population of Educational Attainment less than 9th grade
Percentage with Graduate or professional degree
Percentage of Walked to Work
Average Artificial Surface (Percentage)

<https://doi.org/10.1371/journal.pone.0214190.t002>

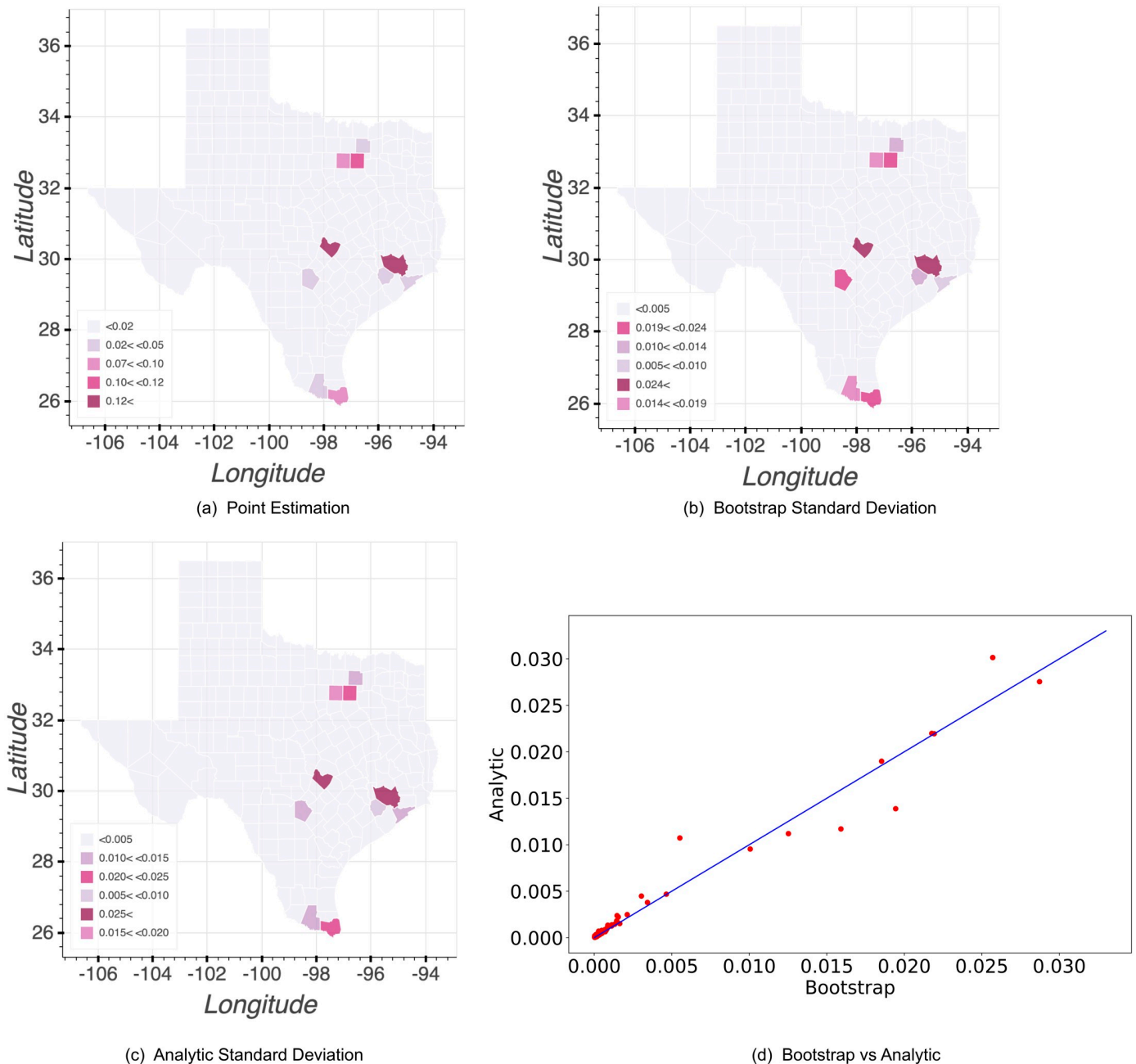


Fig 1. Standard deviation comparison for Dengue importation probability. (a) Figure shows the point estimates for the import probability \hat{p}_i . (b) Figure visually plots the bootstrap standard deviation estimates for p_i across Texas counties. (c) Figure visually plots the analytic standard deviation estimates for p_i across Texas counties. (d) Figure plots the standard deviations of bootstrap vs. analytic and shows a strong equivalence between the two. Each red dot represent the estimations for one county.

<https://doi.org/10.1371/journal.pone.0214190.g001>

demonstrated in [20]. We estimate the standard deviation using the bootstrap method, Poisson PPM approach and the analytic method. The results are presented in Fig 1.

Fig 1a shows the point estimates for the import probability estimated from maximum entropy model and Fig 1b represents standard deviation estimates from the bootstrap method and analytic method of maximum entropy model, respectively. Many Texas counties have

never had imported Dengue cases over the past ten years, and their estimates are close to zero. We map the standard deviation of the estimates p_i of each county in Fig 1b and 1c with a darker color indicating a higher standard deviation level. For bootstrap method, we did 2000 bootstrap runs and took 22403.34 seconds in total. The running time of the analytic method, using optimized matrix operations as described in the S1 Appendix, is dramatically faster than the bootstrap method and takes 0.0016 seconds in total.

Fig 1d shows the standard deviation resulting from the bootstrap against the standard deviation resulting from the analytic method. Each red dot represent a county. It also depicts a regression line between the two results— $s_a = 0.98s_b$ with $R^2 = 0.972$, where s_a and s_b stand for the standard deviation estimates from the analytic and the bootstrap methods, respectively. Regression results show a linear relationship between the standard deviation calculated from analytic expression and bootstrap method with parameter approximate 1. Both bootstrap method and analytic method generally indicate larger standard deviation for counties with larger point estimates.

Aedes aegypti habitat

The *Aedes aegypti* mosquito is the primary transmission vector of dengue, chikungunya, and zika viruses. We aim to estimate the relative probability distribution of *Aedes aegypti* in Texas. Historical presence data $O = \{o_1, o_2, \dots, o_n\}$, with n equal to the number of 1km grid squares in Texas, present empirical samples from this distribution. Each o_i is either 0, if there is no presence data for this square, or 1 if there is presence data. The features $f_j(X)$ represent environmental data for each 1 km^2 area across the Texas.

Specifically, we use 121 locations, within Texas, of *Aedes aegypti* presence data found from previous studies [23–30], DSHS. The environmental features $f_j(X)$, found from WorldClim Database [22], are listed in Table 3.

We aim to analyze the standard deviation of the estimates \hat{p}_i for each 1km square. We estimate this standard deviation using both the bootstrap method and the analytic method. The results are presented in Fig 2.

We present the point estimates of the distribution of the *Aedes aegypti* mosquito in Fig 2a. *Aedes aegypti* primarily feeds on humans and is found in urban areas, which results in higher probability estimates in those areas. The areas of concentration of *Aedes aegypti* in Texas tend to be population centers like Houston, Dallas, San Antonio, Austin, El Paso, and McAllen.

Fig 2b and 2c plots the standard deviation of the estimates p_i of each grid using the bootstrap method and the analytic method. This can give a practitioner a good sense of the standard deviation in the estimates. In applying the analytic method, one could use as input the empirical distribution or a Laplace smoothed estimator [31] to smooth the empirical

Table 3. Seven features, found from WorldClim Database [22], included in maximum entropy model.

Features
artificial surfaces
population count
temperature seasonality
elevation
precipitation seasonality
minimum temperature of coldest month
mean diurnal range

<https://doi.org/10.1371/journal.pone.0214190.t003>

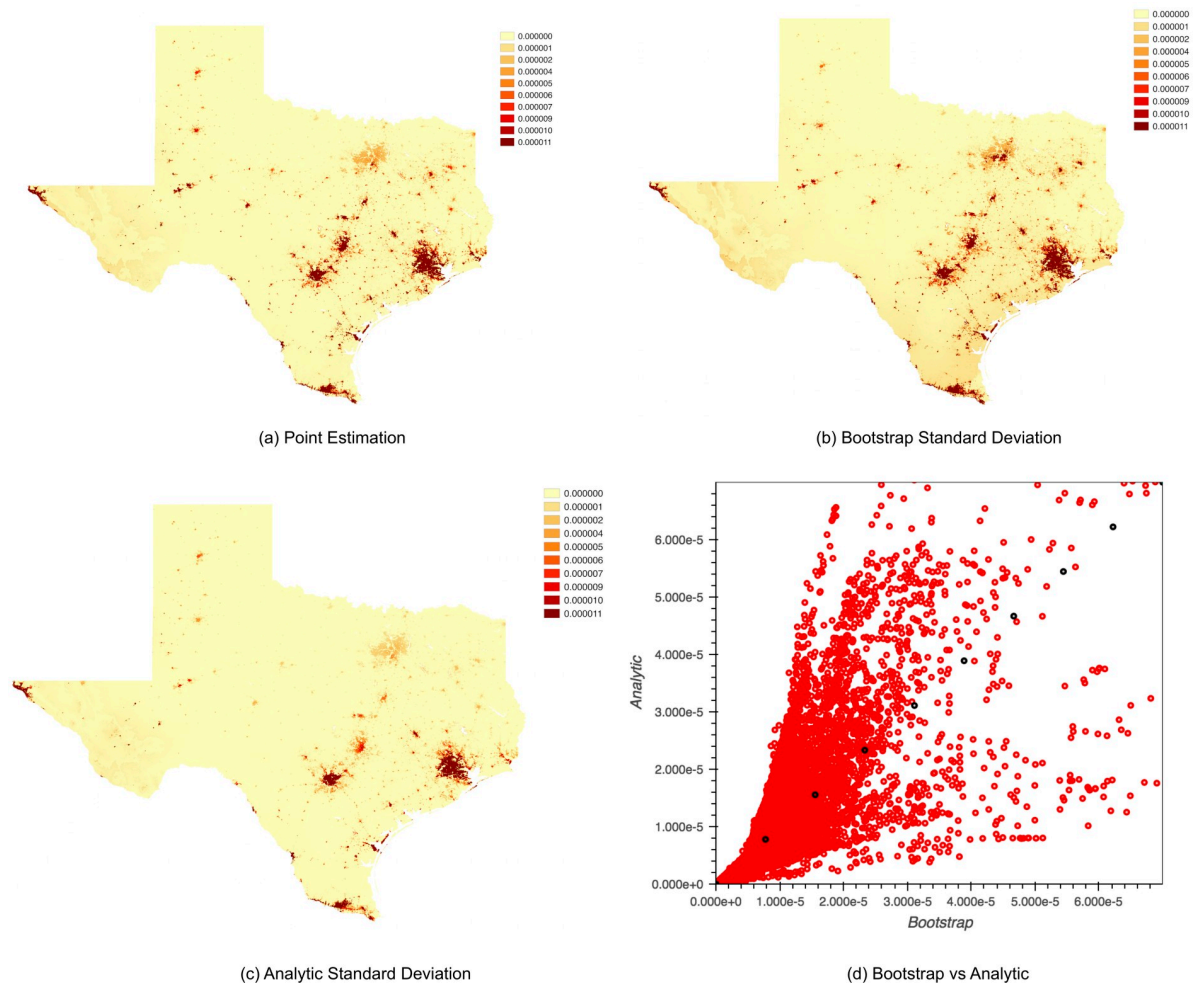


Fig 2. Standard deviation comparison for *Aedes aegypti*. (a) Figure presents the point estimates p_i . (b) Figure shows standard deviation calculated using bootstrap method. (c) Figure shows standard deviation calculated using analytic method. (d) Figure shows the standard deviation comparison between analytic method and bootstrap method.

<https://doi.org/10.1371/journal.pone.0214190.g002>

probability to be non-zero. The analytic method gives slightly higher uncertainty estimates than bootstrap as shown in Fig 2c. Each red dot represent the standard deviation estimates for each grid using the bootstrap and the analytic method respectively. The black dot shows the diagonal line when two methods aligned well. When we applied a Laplace smoothing of 0.0001, we have the relationship $s_a = 1.0744s_b$ with $R^2 = 0.802$, where s_a and s_b stand for the standard deviation estimates from the analytic and the bootstrap methods, respectively. We map the standard deviation of the estimates p_i of each 1 km^2 using the analytic method and Laplace smoothing of 0.0001 in Fig 2b. The bootstrap result and analytic result can be visually compared through Fig 2c.

We did 2000 bootstrap runs and took 30400 seconds in total. The running time for the analytic method is 6516 seconds, which is much faster than bootstrap method. As we calculated the relative probability for *Aedes aegypti* for a 1 km^2 square grid, we have 933,680 grid cells in total. Computing the covariance of the output would require matrix multiplications for matrices of size 933680×933680 , which can cause out-of-memory errors. We introduce a faster method of calculating the variance of each square grid in S1 Appendix.

Discussion

The maximum entropy model can give a point estimation of the unknown species distribution within predefined grids using presence-only data with possible influential features like environmental factors, demographic factors, social economic factors, etc. However, uncertainties come from both the model and the sample data. Some possible sources of this uncertainty are:

- The true expectation of all features $f_j(X)$ are unknown and estimated using the presence-only data.
- Species distribution data are not collected at random, but based on prior knowledge of the biologists. For example, all samples may be observed within pre-selected locations.
- Having only a few presence points relative to the size of the grid can lead to unstable models.
- The features $f_j(X)$ used within the model may be inaccurate or vary dramatically over time. So, it is unclear whether the presence only data collected is appropriate for use with the given features.

In the maximum entropy model, the output probabilities are dependent on the features $f_j(X)$. A flat $f_j(X)$ can only produce flat output probabilities. One may want to know how will the output probabilities change when the feature values change? Uncertainty quantification may help identify the features that most reduce uncertainty in a maximum entropy model.

The bootstrap method is a well accepted method of quantifying uncertainty. However, the running time of the bootstrap can be very long. In the dengue example, bootstrap method took more than 22000 seconds to generate a comparable uncertainty estimate of analytic method while the analytic method just took 0.0016 seconds in total. The analytic method uses more memory compared to the bootstrap method. In the *Aedes aegypti* example, the analytic method took only 20 percent time of running bootstrap method. However, code optimization, and element-wise matrix multiplications can significantly increase the speed of the analytic method compared to the bootstrap method. A method for increasing programming speed are shown in [S1 Appendix](#). Furthermore, the analytic method is able to approximate covariances in the output—whereas this can be quite difficult for the bootstrap method if we only use a small number of samples.

The Poisson PPM approach proved to be equivalent to MAXENT providing an alternative approach of estimating the uncertainty. However, the hidden independence assumption of species appearance locations can affect the performance of the model which gives much larger estimated uncertainty when assumptions violated.

Supporting information

S1 Appendix. Analytic expression of uncertainty, and comparison between analytic method and poisson PPM can be found in S1 Appendix.

(PDF)

S1 Figs. Analytic and poisson PPM comparison. (a) Figure plots the relationship between point estimates of Dengue importation probability vs. variance calculated through analytic method. Non-linear relationship indicates the improper use of Poisson PPM for Dengue importation cases. (b) Figure plots the standard deviations of Poisson PPM vs. analytic for Dengue importation case study and indicates that Poisson PPM provides much larger standard deviation for Dengue imports application. (c) Figure plots the relationship between point estimates of *Aedes Aegypti* existence probability vs. variance calculated through analytic method. (d) Figure shows the standard deviation comparison between analytic method and Poisson

PPM of *Aedes Aegypti* existence probability.
(TIF)

S1 Data File. Dengue importation and *Aedes aegypti* existence case study data. All the data used for Dengue importation and *Aedes Aegypti* existence case study are all included in data file.

(ZIP)

Acknowledgments

This publication was supported by Cooperative Agreement Number U01CK000512, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

Author Contributions

Conceptualization: Xi Chen.

Data curation: Xi Chen.

Formal analysis: Xi Chen.

Methodology: Xi Chen.

Project administration: Nedialko B. Dimitrov.

Software: Xi Chen.

Supervision: Nedialko B. Dimitrov.

Validation: Xi Chen.

Visualization: Xi Chen.

Writing – original draft: Xi Chen.

Writing – review & editing: Xi Chen, Nedialko B. Dimitrov, Lauren Ancel Meyers.

References

1. Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution modeling. In: Proceedings of the twenty-first international conference on Machine learning. ACM; 2004. p. 83.
2. Miller J. Species distribution modeling. *Geography Compass*. 2010; 4(6):490–509. <https://doi.org/10.1111/j.1749-8198.2010.00351.x>
3. Ward G, Hastie T, Barry S, Elith J, Leathwick JR. Presence-only data and the EM algorithm. *Biometrics*. 2009; 65(2):554–563. <https://doi.org/10.1111/j.1541-0420.2008.01116.x> PMID: 18759851
4. Guisan A, Thuiller W. Predicting species distribution: offering more than simple habitat models. *Ecology letters*. 2005; 8(9):993–1009. <https://doi.org/10.1111/j.1461-0248.2005.00792.x>
5. Austin M. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*. 2002; 157(2):101–118. [https://doi.org/10.1016/S0304-3800\(02\)00205-3](https://doi.org/10.1016/S0304-3800(02)00205-3)
6. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences*. 1937; 236(767):333–380. <https://doi.org/10.1098/rsta.1937.0005>
7. Jaynes ET. Information theory and statistical mechanics. *Physical review*. 1957; 106(4):620. <https://doi.org/10.1103/PhysRev.106.620>
8. Shannon CE. The mathematical theory of communication. 1963. MD computing: computers in medical practice. 1996; 14(4):306–317.

9. Kapur JN, Kesavan HK. Entropy optimization principles and their applications. Springer; 1992.
10. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. *Ecological modelling*. 2006; 190(3):231–259. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>
11. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*. 2011; 17(1):43–57. <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
12. Renner IW, Warton DI. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*. 2013; 69(1):274–281. <https://doi.org/10.1111/j.1541-0420.2012.01824.x> PMID: 23379623
13. Kyung M, Gill J, Ghosh M, Casella G, et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*. 2010; 5(2):369–411. <https://doi.org/10.1214/10-BA607>
14. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*. 1986; p. 54–75. <https://doi.org/10.1214/ss/1177013815>
15. Davison AC, Hinkley DV. Bootstrap methods and their application. vol. 1. Cambridge university press; 1997.
16. Manly BF. Randomization, bootstrap and Monte Carlo methods in biology. vol. 70. CRC Press; 2006.
17. Endo T, Watanabe T, Yamamoto A. Confidence interval estimation by bootstrap method for uncertainty quantification using random sampling method. *Journal of Nuclear Science and Technology*. 2015; 52(7-8):993–999. <https://doi.org/10.1080/00223131.2015.1034216>
18. Pearce JL, Boyce MS. Modelling distribution and abundance with presence-only data. *Journal of applied ecology*. 2006; 43(3):405–412. <https://doi.org/10.1111/j.1365-2664.2005.01112.x>
19. Asmussen S, Glynn PW. Stochastic simulation: algorithms and analysis. vol. 57. Springer Science & Business Media; 2007.
20. Castro LA, Fox SJ, Chen X, Liu K, Bellan SE, Dimitrov NB, et al. Assessing real-time Zika risk in the United States. *BMC infectious diseases*. 2017; 17(1):284. <https://doi.org/10.1186/s12879-017-2394-9> PMID: 28468671
21. United States Census Bureau. American Community Survey (ACS); <https://www.census.gov/programs-surveys/acs/data.html>.
22. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A, et al. Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol*. 2005; 25(15):1965–1978. <https://doi.org/10.1002/joc.1276>
23. Vitek CJ, Gutierrez JA, Dirrigl FJ Jr. Dengue vectors, human activity, and dengue virus transmission potential in the Lower Rio Grande Valley, Texas, United States. *Journal of medical entomology*. 2014; 51(5):1019–1028. <https://doi.org/10.1603/ME13005> PMID: 25276932
24. Merrill SA, Ramberg FB, Hagedorn HH. Phylogeography and population structure of *Aedes aegypti* in Arizona. *The American journal of tropical medicine and hygiene*. 2005; 72(3):304–310. <https://doi.org/10.4269/ajtmh.2005.72.304> PMID: 15772327
25. McPhatter LP, Mahmood F, Debboun M. Survey of mosquito fauna in San Antonio, Texas. *Journal of the American Mosquito Control Association*. 2012; 28(3):240–247. <https://doi.org/10.2987/12-6230R.1> PMID: 23833905
26. Gorrochotegui-Escalante N, Gomez-Machorro C, Lozano-Fuentes S, Fernandez-Salas L, Munoz MDL, Farfan-Ale JA, et al. Breeding structure of *Aedes aegypti* populations in Mexico varies by region. *The American journal of tropical medicine and hygiene*. 2002; 66(2):213–222. <https://doi.org/10.4269/ajtmh.2002.66.213> PMID: 12135296
27. Cano F, Gorski J, Eastridge J. Mosquito Surveillance in the Brazos County (Diptera: Culicidae). *Instars: A Journal of Undergraduate Research*. 2015; 1(1).
28. Kavanaugh MD. Influence of Stormwater Drainage Facilities on Mosquito Communities within the City of Denton, Texas;. http://digital.library.unt.edu/ark:/67531/metadc9765/m2/1/high_res_d/thesis.pdf.
29. Barney CE. Dengue Risk Factor Distribution in Harris County, Texas. M.Sc. Thesis, The University of Texas School of Public Health; 2008.
30. Soto D; May 4, 2015. private communication (email).
31. Russell SJ, Norvig P, Canny JF, Malik JM, Edwards DD. Artificial intelligence: a modern approach. vol. 2. Prentice hall Upper Saddle River; 2003.