

Title: Integrated spatial model estimates the fish distribution using environmental DNA and catch data

Yuki Kanamori^{1*}, Hiroshi Okamura², Shota Nishijima², Yuki Hongo², Yasuyuki Uto³, Hisatoku Mita⁴, Mitsuhiyo Ishii⁴, Kiyoharu Akimoto⁵, and Akane Kusano⁶

¹ Fisheries Resources Institute, Japan Fisheries Research and Education Agency, 25-259 Shimomekurakubo, Samemachi, Hachinohe, Aomori 031-0841, Japan

² Fisheries Resources Institute, Japan Fisheries Research and Education Agency, 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

³

⁴

⁵

⁶

* Corresponding author

Email: kana.yuki@fra.affrc.go.jp

Abstract

Keywords

1 Introduction

Understanding of spatial distribution of species and underlying its mechanism is a major goal in ecology. Field surveys using environmental DNA (eDNA) are widely used for detecting invasive or rare species and hotspot of biodiversity because the surveys of eDNA are easy to detect occurrence of target species, non-invasiveness, and high cost effectiveness rather than previous direct sampling method (Rees et al. 2014; Thomsen & Willerslev 2015). However, uncertainty of the occurrence of eDNA is an issue in eDNA studies because various factors, such as biological and environmental features, intricately affect the shedding (i.e., production of eDNA from organisms) and degradation (i.e., decay of eDNA from a system) of eDNA (Fig. 1). Therefore, it is necessary to consider the spatial uncertainty included in eDNA to estimate the spatial distribution of species.

One step towards overcoming these uncertainties is a understanding of the "ecology of eDNA": (Barnes & Turner 2016). Previous studies

Integrated species distribution models (IDMs) are now common spatial model to predict spatial pattern of species (Issac et al. 2020). The model use the different type of data with strengths and weaknesses, such as scientific survey data which is restricted spatially and quantitatively and opportunistic citizen data which is widely collected and abundant, and combine in a single model (Isaac et al. 2020; Miller et al. 2019).

The models combine the different type of data with strengths and weaknesses in a single model (). For example, scientific survey data are high quality but less abundant due to restriction of spatially costly while opportunistic data such as citizen data are widely collected and abundant but may be low quality due to not using consistent field methods.

23 Combining both types of data can capitalize on the strengths of each data and perform better
24 prediction than models when we use single data (Pacifi et al. 2017; Miller et al. 2019).

25 Tokyo Bay is a large enclosed coastal sea in Japan. In Tokyo Bay, there are many
26 commercially important species for fisheries that are called "Edomae" because these species
27 have been used for Sushi since Edo Era (about 400 years ago). Catch of some Edomae have
28 been decreased because of habitat modification due to urbanization (e.g., landfill of tidal
29 flats and water pollution). Catch statistics (total catch in each species, efforts, and
30 geographic location of fishing) have been collected for stock assessment since 1990 by
31 prefectures around Tokyo Bay. The strengths of this data are the direct evidence that a focal
32 species occupies a location of fishing and abundant because of widely collected in Tokyo
33 Bay. On the other hand, weakness of this data is like a opportunistic data because the data is
34 likely to be biased towards areas to high density of focal species due to commercially fishes,
35 consequently less zero data. In addition to this catch statistics, scientific survey of eDNA
36 has been conducted monthly since 2018 for biodiversity monitoring because biodiversity
37 also may decreased due to human-induced environmental changes in Tokyo Bay (Hongo et
38 al., submitted). The strengths are that the data is systematically collected by scientific survey
39 data and includes zero data, while the weaknesses are that the data is less abundant due to
40 spatial restriction of the survey and includes uncertainties in presence/absence as description
41 in above.

42 In this paper, to predict spatial distribution of species from eDNA, we first make a
43 model which considers uncertainties of eDNA caused by environmental factors without
44 additional laboratory experiments and numerical hydrodynamic models, by using an
45 integrated spatial distribution model (eDNA-IDM). We then apply the model to both eDNA

46 data and catch statistics for four Edomae fish in Tokyo Bay, Japan. The predicted spatial
47 distribution of four fish form our model reduced

48

49 **2 Materials and Methods**

50 **2.1 A general model to estimate species distribution from eDNA**

51 Integrated spatial distribution model that account for explicitly spatial autocorrelation in
52 occurrence were built by Pacifici et al. (2017), which shows three approaches to predict the
53 spatial distribution of species: the joint likelihood (shared), correlation, and covariate
54 methods. The joint likelihood method uses multiple data types to simultaneously estimate a
55 shared set of parameters with constraining that the likelihoods of shared set of parameters to
56 be equal across. The correlation method connects multiple data types indirectly through a
57 shared covariance matrix that captures similar patterns present in each data sources. The
58 covariate method incorporates information from a added dataset via a fixed effect.

59 Although each methods estimate the spatial distribution of species using multiple data
60 sets, we need to select method depending on the data features for analysis because there are
61 strengths and weaknesses (Pacifici et al. 2017; Miller et al. 2018). The joint likelihood
62 method may be problematic when the second data is of poorly quality compared to
63 correlation and covariate methods because each data can directly inform the latent
64 occurrence state (probabilities?) and the weight given to estimate the parameters is naturally
65 determined by their relative size and quality. Thus, it is not the best method when our
66 second data is low quality while it is the best method when our second data is high quality

67 (vise versa). The correlation method is added robustness to the joint likelihood because the
 68 second data indirectly inform the occurrence state. Thus, it is the best method when our
 69 second data is low quality while it is inferior to the joint likelihood method when both data
 70 are deemed reliable. The covariate method does not make full use of the information in the
 71 second data because the second data as a constructed covariate in the mean occurrence state.
 72 In addition, this method can reduce the computational cost because there are fewer
 73 parameters to estimate and the number of data locations can be reduced. Thus, it is the best
 74 method when the second data is low quality and/or there is computational limitation while it
 75 may not the best method when the information of the second data is needed.

76 When predicting the spatial distribution of species from eDNA using integrated
 77 species distribution model, the information that a species exists is needed as second data to
 78 consider spatial uncertainties of eDNA due to complex factors (Fig. 1). Hence, the second
 79 data is preferred to high quality as possible.

80 しかし、eDNA は直接的なモニタリングに比べて簡易的であるためより広い範囲で取
 81 得されている可能性が高く、eDNA のデータと同様の空間範囲で調査データのように
 82 質の高いデータを取得することは難しいかもしれない。その一方で、eDNA の空間的
 83 な不確実性を考慮するためには、種がいた証拠である 2 番目のデータの情報を eDNA
 84 のデータにしっかりと伝える必要がある。これらを考えると、integrated spatial
 85 distribution model を用いた eDNA からの空間分布の推定には、以下のような correlation
 86 method が適切である：

$$\begin{aligned}
 p_e(s_i) &= \alpha_e + \sum_k f_{e,k}(x_{e,k}(s_i)) + w\theta(s_i) + u_e(s_i) \\
 p_a(s_i) &= \alpha_a + \sum_k f_{a,k}(x_{a,k}(s_i)) + \theta(s_i) + u_a(s_i)
 \end{aligned}
 \tag{1}$$

where α and $x_k(s_i)$ are the intercept and the covariates at sites i for occurrence probabilities at sites i of the added data (p_a) and eDNA data (p_e), respectively. $u(s_i)$ is spatial error that is specific for each data following multivariate normal distributions $MVN(0, \mathbf{R})$, where the variance–covariance matrix \mathbf{R} is a Matérn correlation function. θ which is shared between two equations is the common spatial pattern between the two data, which cannot explain by each terms of the equations. That is, θ can be interpreted as "true" spatial distribution of species.

2.2 An application to a eDNA and catch data in Tokyo Bay

2.2.1 eDNA data

Field surveys

Field surveys were conducted by prefectural experimental station in Chiba, following the consistent sampling design at 14 sites in Tokyo Bay from April to December in 2018 (Fig. 1). In each sites, seawater and environmental data were simultaneously collected. For eDNA analysis, two litter of bottom seawater was collected using a Niskin water sampler, and then it was separated for two 1L samples for replicate. Each samples filtered glass fiber membrane GF/F ($0.7 \mu m$ pore size; Cytiva, Sheffield, UK) onboard and then the filters were frozen on a block of dry ice. These frozen filters were stored at -30° in the laboratory until eDNA extraction. To lower the levels of cross-contamination, equipments for eDNA sampling were changed new one or washed in each sites. During sampling the bottom seawater, seawater temperature, salinity, pH, and dissolved oxygen (DO) at the same depth of seawater sampling for eDNA were measured by CTD (メーカ一).

108 **Laboratory experiments**

109 In laboratory, eDNA extraction, eDNA amplification, and eDNA sequence were conducted.
110 Total eDNA was extracted from the frozen filters using a DNeasy Blood and Tissue Kit
111 (Qiagen, Hilden, Germany) following Yamamoto et al. 2019. Mitochondrial 12S rRNA
112 gene was amplified using MiFish universal primers referring to Miya et al. 2015 with slight
113 modification. The details was shown in Hongo et al. (受理されてないようだったら書くし
114 かない). eDNA sequence were

115 **2.2.2 Catch statistics**

116 A part of catch statistics of small-scale bottom trawl fisheries recorded by several
117 representative boats of Chiba Prefecture were provided by Chiba Prefecture. This data
118 included date, geographic location, efforts (number of tows), gear, and catch weight (kg) in
119 each fish. Almost of all gear was beam trawl although dredge net also used. The species
120 which also detected by eDNA was *Conger myriaster* (マアナゴ), *Kareius bicoloratus* (イシ
121 ガレイ), *Lateolabrax japonicus* (スズキ), and *Konosirus punctatus* (コノシロ). Thus, we
122 estimated the spatial distribution of these four species using the eDNA-IDM. マコガレイ,
123 カマス類, クロダイ, イシモチ類も解析できる??

124 **2.2.3 Estimation of spatial distribution**

125 To estimate the spatial distribution of four focal species using eDNA and catch data by
126 considering with spatial uncertainties of eDNA, we fitted the model (equation 1) to the
127 presence/absence of eDNA and of catch collected in Tokyo Bay as follows:

$$\text{logit } p_e(s_i) = \alpha_e + \sum_k f_k(x_k(s_i)) + w\theta(s_i) + u_e(s_i)$$

$$\text{logit } p_c(s_i) = \alpha_c + \beta_i + \theta(s_i) + u_c(s_i)$$

128 where α is the intercept, and $x_k(s_i)$ is the covariates at sites i for occurrence probabilities of
 129 eDNA at sites i . In the study, seawater temperature, salinity, pH, and DO were used as
 130 covariates which effect on the occurrence of eDNA (i.e., $k = 4$). $u(s_i)$ is spatial error that is
 131 specific for each data following multivariate normal distributions $MVN(0, \mathbf{R})$, where the
 132 variance–covariance matrix \mathbf{R} is a Matérn correlation function. θ which is shared between
 133 eDNA and catch is the common spatial pattern between the two data, which cannot explain
 134 by each terms of the equations. That is, θ can be interpreted as the spatial distribution of
 135 species we want to know. 共変量の非線形性について書く Parameters in this model was
 136 estimated by Integrated Nested Laplace Approximation using using the R-INLA package
 137 (Lindgren, 2012) in R 3.6.1 (R Development Core Team, 2019).

138

139 Acknowledgments

140 This research was financially supported by Grant-in-Aid for Fisheries Agency of Japan.

141 Authorship

142 YK conceived of the research idea. YH, YU, HM, MI, KA, and AK conducted field
 143 sampling. YH performed the laboratory experiments. YK, HO, and SN designed statistical
 144 analyses. YK wrote programs and performed the analyses. YK wrote the manuscript with
 145 input from all co-authors' comments.

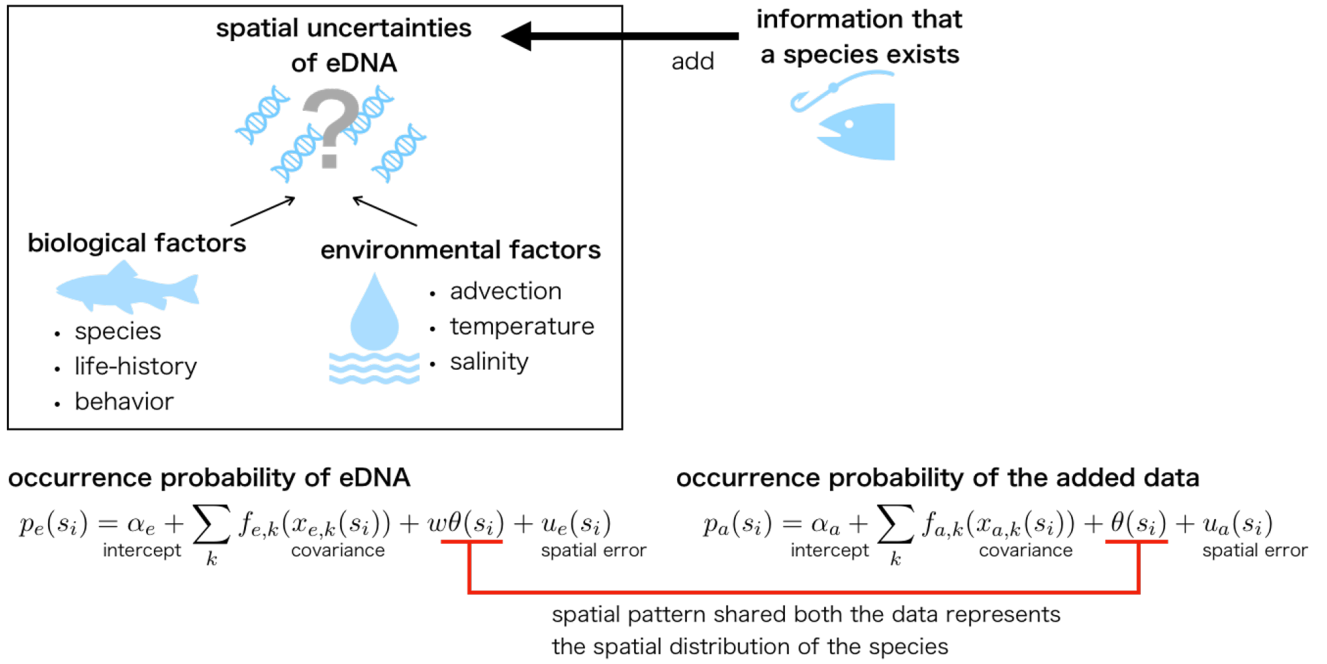


Fig. 1: Conceptual diagram of this study.