

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

Yuki Kanamori<sup>1\*</sup>, Hiroshi Okamura<sup>2</sup>, Shota Nishijima<sup>2</sup>, Yuki Hongo<sup>2</sup>, Yasuyuki Uto<sup>3</sup>, Hisatoku Mita<sup>4</sup>, Mitsuhiyo Ishii<sup>4</sup>, Kiyoharu Akimoto<sup>5</sup>, and Akane Kusano<sup>6</sup>

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

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

<sup>3</sup>

<sup>4</sup>

<sup>5</sup>

<sup>6</sup>

\* Corresponding author

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

**Abstract**

**Keywords**

# 1 Introduction

Understanding of spatial distribution of species and underlying its mechanism is an essential issue in ecology. Field surveys using environmental DNA (eDNA) are widely used for detecting invasive or rare species and hotspot of biodiversity (面倒なのでレビュー論文を引用) because the surveys of eDNA are easy to detect presence/absence of target species, non-invasiveness, and high cost effectiveness rather than previous direct sampling method (Rees et al. 2014; Thomsen & Willerslev 2015). However, the presence/absence of eDNA includes many types of uncertainties due to relating to environmental factors such as temperature and advection (). For example, in aquatic habitats, it is not sure whether target species are in a location or not when eDNA of target species is detected because eDNA are transported passively. Therefore, the consideration to the influence of environmental factors on eDNA is necessary for estimation of species distribution when we use eDNA methods.

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

23 collected and abundant but may be low quality due to not using consistent field methods.  
24 Combining both types of data can capitalize on the strengths of each data and perform better  
25 prediction than models when we use single data (Pacifi et al. 2017; Miller et al. 2019).

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

43 In this paper, to predict spatial distribution of species from eDNA, we first make a  
44 model which considers uncertainties of eDNA caused by environmental factors without  
45 additional laboratory experiments and numerical hydrodynamic models, by using an

integrated spatial distribution model (eDNA-IDM). We then apply the model to both eDNA data and catch statistics for four Edomae fish in Tokyo Bay, Japan. The predicted spatial distribution of four fish from our model reduced

## 2 Materials and Methods

### 2.1 A general model to estimate species distribution from eDNA

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

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

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

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

81 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  
82 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  
83 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  
84 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  
85 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 spetial  
86 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  
87 method が適切である:

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

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

## 95 **2.2 An application to a eDNA and catch data in Tokyo Bay**

### 96 **2.2.1 eDNA data**

#### 97 **Field surveys**

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

107 seawater, seawater temperature, salinity, pH, and dissolved oxygen (DO) at the same depth  
108 of seawater sampling for eDNA were measured by CTD (メーカー).

## 109 **Laboratory experiments**

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

### 116 **2.2.2 Catch statistics**

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

### 125 **2.2.3 Estimation of spatial distribution**

126 To estimate the spatial distribution of four focal species using eDNA and catch data by  
127 considering with spatial uncertainties of eDNA, we fitted the model (equation 1) to the



128 presence/absence of eDNA and of catch collected in Tokyo Bay as follows:

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

129

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

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

140

## 141 Acknowledgments

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

## 143 Authorship

144 YK conceived of the research idea. YH, YU, HM, MI, KA, and AK conducted field  
145 sampling. YH performed the laboratory experiments. YK, HO, and SN designed statistical  
146 analyses. YK wrote programs and performed the analyses. YK wrote the manuscript with

<sup>147</sup> input from all co-authors' comments.

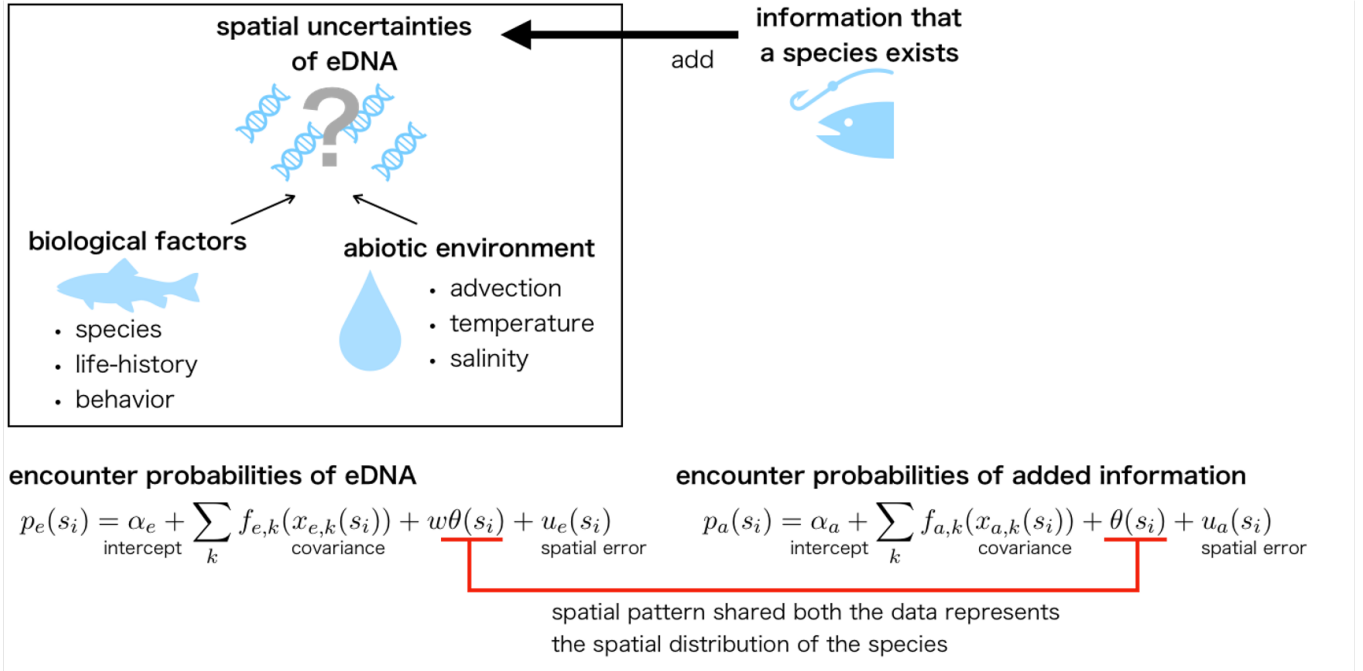


Fig. 1: Conceptual diagram of this study.