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

Yuki Kanamori[1*], Hiroshi Okamura[2], Shota Nishijima[2], Yuki Hongo[2], Yasuyuki Uto[3], Hisatoku Mita[4], Mitsuhiyo Ishii[4], Kiyoharu Akimoto[5], and Akane Kusano[6]

[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

1   Introduction

Understanding of spatial distribution of species and underling its mechanism is a 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

3

collected and abundant but may be low quality due to not using consistent field methods.

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

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

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

4

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 form 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

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

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

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 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 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 method が適切である:

$$p_e(s_i) = \alpha_e + \sum_k f_{e,k}(x_{e,k}(s_i)) + w\theta(s_i) + u_e(s_i)$$

$$p_a(s_i) = \alpha_a + \sum_k f_{a,k}(x_{a,k}(s_i)) + \theta(s_i) + u_a(s_i)$$

(1)

88   where α 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 $\mathrm{MVN}(0, \mathbf{R})$, where the

91   variance–covariance matrix $\mathbf{R}$ is a Matérn correlation function. θ 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, θ can be interpreted as "true" spatial distribution of

94   species.

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

95

2.2.1 eDNA data

96

Field surveys

97

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 μm 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 $-30°$ 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

7

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).

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. Mitochondorial 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. (受理されてないようだったら書くし

₁₁₅ かない). eDNA sequence were ....

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？？

2.2.3 Estimation of spatial distribution

₁₂₆ To estimate the spatial distribution of four focal species from eDNA and catch data by

₁₂₇ considering uncertainties caused by environmental factors, we fitted the model (equation 1)

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

$$\text{logit } p_i = \beta_p(t_i) + \omega_p(s_i) + \varepsilon_p(s_i, t_i) + \eta_p(v_i) + \lambda_p Q(i)$$

$$\log d_i = \beta_d(t_i) + \omega_d(s_i) + \varepsilon_d(s_i, t_i) + \eta_d(v_i) + \lambda_d Q(i)$$

(2)

130 where $\beta(t_i)$ is the intercept for year t, and $\omega(s_i)$ and $\varepsilon(s_i, t_i)$ are the spatial and

131 spatio–temporal random effects for year t and location s, respectively. $\eta(v_i)$ is the

132 overdispersion random effect of factor v_i, which is the interaction of year and month. λ is

133 the effect of the chatchability covariate $Q(i)$:

$$Q(i) = \log(d_{chub}(s_i) + 0.1).$$

134 That is, this term considers the effect of species misidentification between chub mackerel

135 and spotted mackerel; as mentioned earlier, we suspected overestimation of egg density of

136 spotted mackerel because the difference in egg diameter has become ambiguous according

137 to increase in egg density of chub mackerel and the distributions of egg diameters between

138 species have overlapped (Yukami et al., 2019). The constant 0.1 was added because $\log 0$

139 (i.e., no chub mackerel eggs) is undefined, and the same result was obtained when using 1 in

140 place of 0.1.

141

Acknowledgments

Authorship

145 YK conceived of the research idea. YH, YU, HM, MI, KA, and AK conducted field

sampling. YH performed the laboratory experiments. YK, HO, and SN designed statistical

analyses. YK wrote programs and performed the analyses. YK wrote the manuscript with
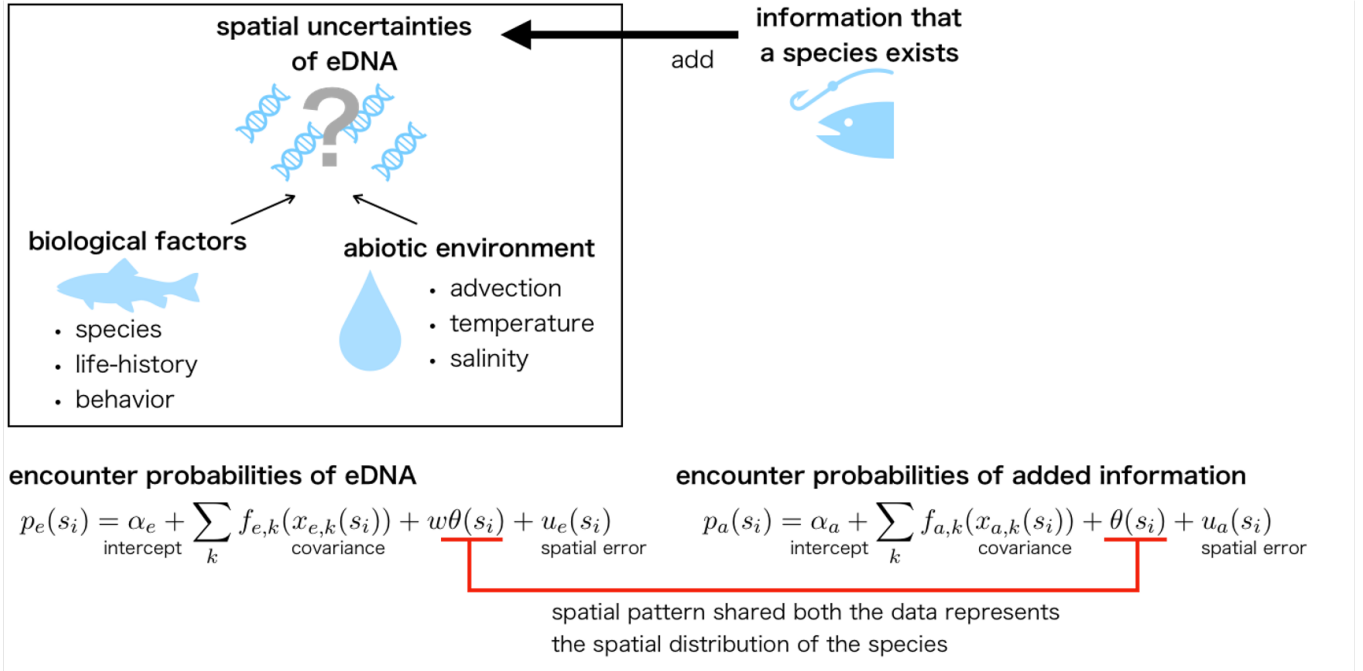
input from all co-authors' comments.

Fig. 1: Conceptual diagram of this study.