Работа посвящена одной из самых молодых тем морской биологии - анализу связи биологических черт видов с градиентами параметров среды. Потенциально, работы в этом направлении дадут ответ на вопрос, как факторы среды отбирают определенные композиции видов, сосуществующих в сообществе. Предполагается, что отбор сосуществующих видов происходит на основе функциональных черт видов. Работа в этом направлении позволит лучше понять взаимосвязи между функциональными чертами видов и факторами среды.

Методической основой для изучения этой взаимосвязи являются Joint species distribution models позволяющие оценить взаимосвязь несколько матриц данных (Ovaskainen & Abreg, 2020): матрицу обилий видов, матрицу биологических черт видов, матрицу средовых данных (эти три матрицы фигурировали в оцениваемой работе), а также матрицу филогенетических связей между видами, матрицу описывающую дизайн сбора материала, и матрицу географических координат. Последние две матрицы обуславливают то, что модель по своей природе оказывается смешанной, в которых случайную часть составляют пространственные эффекты и взаимозвисимости данных, определяющиеся дизайном сбора материала.

Именно в этом и заключается главная проблема рассматриваемой работы: авторы попытались проанализировать очень непростые данные с помощью очень простого инструмента (RDA), который по своей природе позволяет оценить только фиксированные эффекты и не позволяет строить смешанные модели. Самый корректный путь заключался бы в том, чтобы полностью изменить схему анализа и применить Joint Species Distribution Modelling (see Ovaskainen, Abrego, 2020) с использованием пакета “Hmsc”. Этот подход позволит решить поставленные задачи значительно глубже.

В принципе, можно было бы развить и тот подход, который изначально был предложен авторами. Однако для успешного движения в этом направлении необходимо исправить несколько важных недочетов, которые допустили авторы.

Самый главный недочет заключается в том, что авторы, используя RDA, забывают, что это частный случай регрессионного анализа. Стало быть модель, построенная с его помощью, требует оценки статистической значимости. Авторы ни разу этого не сделали.

Вместе с тем, мощности пакета vegan позволяют это сделать с помощью пермутационных методов. Это можно сделать с помощью функции anova.cca() из пакета vegan. При этом можно использовать следующие параметры anova(RDA\_Model, by = NULL) даст информацию о статистической значимости всей модели; anova(RDA\_Model, by = “axes”) даст информацию о статистической значимости каждой из constrained axes; anova(RDA\_Model, by = “margin”) даст информацию о статистической значимости отдельных предикторов. I do not recommend to use anova(RDA\_Model, by = “terms”) to avoid dependency on sequention of predictors in the model. В любом случае без оценки статистической значимости RDA неясно насколько можно доверять полученным результатам.

Еще одной статистической проблемой имеющегося анализа является множественность сравнений без адекватной корректировки выводов. Так, например, адекватное сравнение двух шельфов возможно только если предиктор «Шельф» включен в общую модель регрессионного анализа (в том числе в RDA), наравне с другими предикторами. Сделать это можно было бы, но это серьезно затруднило бы трактовку ограниченных осей. Поэтому, раздельный анализ материала из двух разных шельфов можно себе позволить. Однако с огромным количесвтом оговорок. Если же для описания двух разных шельфов строится две разные модели, то мы неизбежно попадаем в область множественных равнений. Если анализ происходит для двух шельфов независимо, то это, как минимум, требует корректировки уровня отвержения нулевой гипотезы (например, за счет поправки Бонферронни). Если предиктор «Шельф» не включен в общую модель, то отвержение нулевой гипотезы должно проходить при уровне alpha = 0.05/2 = 0.025.

Авторы, вроде как не используют оценку статистической значимости, однако сейчас станет ясно применяемая везде процедура подбора оптимальной модели, основана у авторов на использовании стандартного порога отвержения нулевой гипотезы равного 0.05.

Авторы использовали процедуру подбора оптимальной модели около 20 раз(!). Для этого была использована функция ordiR2step() из пакета ”vegan”. Cудя по коду, приведенному в файле “Sutton\_EFcode\_FiMS.R”, эта функция была использована со следующим набором параметров:

ordiR2step(ChukchiRDA0, scope=formula(ChukchiRDA1),direction="forward", Pin=.05, permutations = 999, R2scope = FALSE)

«R2scope = FALSE» (это не есть значение по умолчанию). Такое значение параметра означает, что критерием остановки подбора модели является не adjusted R2 (что было бы логично), а оптимальной моделью считалась модель с пермутационным p-value меньшим, чем p=0.05 («Pin=0.05»). Таким образом никакой корректировки на множественность тестирования гипотез не производилось. Авторам следует в явной форме объяснить почему они взяли именно такой набор параметров, который не является набором по умолчанию в данной функции.

Я рекомендовал бы авторам отказаться от применения подбора оптимальной модели (это совсем немеханическая и небезобидная процедура). Вместо этого я предлагаю продумать более компактную схему анализа, основанного на подборе всего двух моделей (одна для одного моря, другая для другого). Каждая из моделей была бы подобрана как partial RDA модель. В модели следовало бы включить ВЕСЬ набор средовых предикторов плюс матрицу географических координат станций, as a conditioning matrix, the effect of which is removed (“partialled out”) before next step. После построения таких моделей можно будет увидеть «чистые» градиенты, связанные с факторами среды, а не с географическими градиентами, которые вносят шумы если не учитывать пространственные эффекты. Для реализации предложенной идеи можно использовать вот такой код (for Beaufort Sea the code is analogous).

station\_data <- read.csv("Both Seas/Arctic\_Sea\_Lat\_Long.csv")

Chukchi\_coord <- station\_data %>% filter(Sea == "Chukchi") %>% select(Latitude, Longitude)

RDA\_model\_partial <- rda(ChukchiLQ ~ . + Condition(as.matrix(Chukchi\_coord)), data = ChukchiRstand)

anova(RDA\_model\_partial)

anova(RDA\_model\_partial , by = "axis")

anova(RDA\_model\_partial , by = "margin")

Вы увидите, что при учете географических координат, как контролирующей матрицы, количество статистически значимых предикторов («чистых» природных градиентов) окажется очень невелико. Это конечно, не желаемая смешанная модель, которую можно бы бостроить с помощью пакета «Hmsc” но, по крайней мере географический компонент в модели был бы учтен.

В принципе, в уже существующей схеме анализа можно было бы учесть и пространственные корреляции. Можно было бы использовать partial RDA. Для этого можно применить очень простой код: Model <- rda(X = LQ\_matrix, Y = R\_matrix, Z = Geographical\_coordinates\_matrix). Geographical\_coordinates\_matrix will consist of two columns.

Честно говоря, я не понял, зачем нужно было проводить отдельную процедуру подбора моделей для каждого типа функциональных черт (Figure 4). Никто ведь не мешает рассматривать ординацию в ограниченных осях не только объектов (станций), но и признаков (функциональных черт). Эта ординация может быть взята непосредственно из той модели которая использовалась для построения рисунка Figure 3. Надо просто написать вот такой код

plot(ChukchiRDAstep, display = c("species", "cn"))

Здесь будет видна ординация функциональных черт в ограниченных осях, заданных средовыми предикторами.

Why the name of this paper is “Environmental filtering influences functional community assembly of **epibenthic** communities”? The analysis of species list revealed numerous infaunal elements inhabiting soft sediments (I.e. Phyllodoce.groenlandica, Atylus.carinatus, Cerianthus.lloydi, Cylichna.alba, Diastylis, Nepthys.ciliata, Pontoporeia femorata, Serripes.groenlandicus). Я взял только те виды, которые точно встречаются в Северо-Западной Атлантике и точно являются инфауной. Создается впечатление, что критерием выделения эпибентоса является орудие лова. Трал собирает все, и эпибентос и инфауну. Следовательно выражение «epibenthic community” некорректно.

На этом раунде рецензирования имеет смысл остановиться только на методически вопросах, так как работа изобилует проблемами, связанными с некорректным использованием методов и отсутствием статистических оценок значимости результатов. Поэтому я пока не буду касаться трактовок, приведенных в Discussion, так как пока еще не ясно насколько полученные результаты статистически значимы.

1. На строчке 198 сказано, что максимальное значение модальности равно 3. НО! В матрице из файла ChukchiTraitsbyTaxon.csv присутствуют коды до 6. Либо надо исправить, либо пояснить почему такая разница.
2. На строке 200-202 сказано «The Q matrix was then multiplied with a square root transformed, biomass-weighted taxon by station matrix (L matrix)». Матрица Q (e.g. in the Chukchi Sea) имеет размерность 327х38, матрица L имеет размерность 67х327. Следовательно для соблюдения правил умножения матриц должна матрица L быть умножена на матрицу Q. Или объясните, что имеется ввиду.
3. Количество видов должно быть равно количеству строк в матрице Q (соответственно, столько же столбцов в матрице L). НО!

В матрице ChukchiQ имеется 327 строк, в матрице BeaufortQ - 246 строк. В тексте (строка 193-194) сказано, что «with almost identical taxon richness in the Chukchi and Beaufort seas (n=227, n=226, respectively)». Почему расходятся эти числа?

1. Для подбора оптимальной модели в коде используется, функция ordiR2step(), Function ordiR2step performs forward model choice solely on **adjusted R2 and P-value.** НО в тексте (строка 214-215) авторы пишут, что «a forward model selection using the lowest Akaike Information Criterion (AIC) was performed for each shelf». Actually, constrained ordination methods (such as RDA) do not have AIC. Only AIC-like statistic could be used. But in any way function ordiR2step() do not use AIC at all. Авторы должны объяснить, почему они в функции ordiR2step() поставили параметр R2scope = FALSE. Это означает, что оптимальной моделью считалась модель с пермутационным p-value меньшим, чем p=0.05 (параметр Pin=0.05). Так можно, но надо учесть, что две акватории рассматривались независимо. То есть предиктор «Sea” не был включен в общую модель. Это переводит данный случай в область multiple comparison. Следовательно p-value should be adjusted using some correction technique (e.g. Bonferronni correction). Thus the threshould p-value should be equal to 0.05/2 = 0.025.
2. Вообще не понятно, зачем авторы использовали подбор оптимальной модели (forward selection). Количество предикторов у них очень невелико, всего 7 (Gravel, Sand, Mud, Depth, Temp, Salinity, Chl.a). Для столь небольшого количеcтва проще рассматривать весь набор предикторов в каждом из морей и смотреть на то? где какие предикторы статистически значимы, а какие нет (естественно, с поправками на множественность сравнений).
3. Проверялось ли наличие мультколлинеарности между предикторами? Если предикторы коррелируют, то влияние одного может маскироваться влиянием другого. Проверить наличие мультиколлинеарности можно с помощью функции vif.cca().
4. Важная проблема примененного анализа заключается в том, что фактор «Sea” не включен в общую модель. Формально это не дает оснований для статистических сравнений двух акваторий. Как минимум надо объяснить в тексте, что все сравнения двух акваторий не имеют статистического подкрепления.
5. Авторы применяли подбор оптимальной модели для каждого типа функциональных черт. Это должно было приводить к разному набору предикторов в финальной модели для каждого типа функциональных черт. Соответственно доли дисперсии определяются разными наборами предикторов для разных функциональных черт. Это делает несопоставимыми результаты RDA для разных черт. К тому же критерий отбора финальной модели в данном случае потребует еще более суровой корректировки порогового уровня для p-level (так как во всех случаях в селеция производилась с параметром R2scope = FALSE) так как количество сравнений еще более увеличивается. Я рекомендую отказаться от подбора оптимальной модели во всех случаях применения RDA.
6. Авторы используют RDA как основу для для выявления градиентов. RDA это частный случай регрессионного анализа, реализованного для многомерных данных. Стало быть все ограничения, характерные для обычных линейных моделей здесь будут иметь значение. Одно из важных ограничений - это независимость величин. Один из источников нарушения независимости - это пространственные корреляции. Авторы должны предоставить информацию, что остатки в модели RDA не имеют пространственных паттернов. Для этого надо хотя бы визуализировать на картах значения unconstrained axes PC1 и PC2, полученных из моделей RDA. PC для каждой станции характеризуют остатки (Legendre&Legendre, 2012) от модели. Эти карты можно привести в приложении, наподобие того, как сделаны карты в FigureS2. Если на картах будет выявлен какой-то неслучайный паттерн распределения значений PCA, то это будет свидетельствовать о нарушении условий применимости линейных моделей. Для более строгой оценки наличия паттерна в остатках следует строить вариограммы. Авторы знакомы с этим подходом, судя по тому, что они применяют его для анализа отсуствия пространственной автокорреляции в анализе, где зависимой переменной является FDis. Отсутствие автокорреляции для FDis не является доказательством отсутствия пространственных корреляций в RDA.
7. Главная проблема в том, что авторы строят очень простую модель RDA для очень непростых данных. Учтенные в модели RDA предикторы формируют фиксированную часть модели, а важная для пространственно выраженных данных случайная часть модели (географические координаты) не учитываются. Иными словами модль должна быть смешанной. RDA этого не умеет. Эту сложность можно обойти двумя путями. Самый корректный - это полностью изменить схему анализа и применить Joint Species Distribution Modelling (see Ovaskainen, Abrego, 2020) с использованием пакета “Hmsc”. Этот подход позволит решить поставленные задачи значительно глубже. Второй подход, который не потребует коренного пересмотра всего анализа - это использование partial RDA. Для этого можно применить очень простой код. Во всех случаях, где применяется функция rda() можно написать вот так Model <- rda(X = LQ\_matrix, Y = R\_matrix, Z = Geographical\_coordinates\_matrix). Geographical\_coordinates\_matrix will consist of two columns. В этом случае не получится провести подбор оптимальной модели, но как я уже говорил выше крайне рекомендуется отказаться от этого подхода.
8. Авторам следует подробнее описать что такое CWM matrix. Анализ R-code показывает, что это то же самое, что LQ matrix.
9. I did not found in R-code file any analysis corresponded to PCoA described in lines 233-235.
10. I highly support the intention of authors to present their work as producible research. However the R-code provided in “Sutton\_EFcode\_FiMS.R” is not complete enough to trace all steps of analysis. I would prefer to have deal with html-file knitred from RMD-file with code included where all steps of analysis would be explined and all figures and tables included into final version of the article would be presented.
11. Table 2 includes information on adjR2 for both RDA models which is correct. However the information on total variance for two models is not comparable. This is because these models include different number of predictors. It is natural that the model with more predictor number will explain more variation. To make the comparison of explained variance proportion meaningful authors should avoid optimal model selection procedure. Otherwise this information should not be considered.
12. It is strange that proportion of variance explained by RDA1 and RDA2 presented in Figure 3 (60 +21 = 81% for Chukchi Sea and 72+28 = 100% for Beaufort Sea) is higher than total variance explained = 45% and 21%, respectively; Table 2). The basic plot.cca() function from the package “vegan” which was used for the Figure 3 construction, if believe to R-code (lines 77-80 in “Sutton\_EFcode\_FiMS.R”), does not provide the proportion of variance. However the general view of the picture in Figure 3 allow me to suppose that for its production the functions from package “ggplot2” where used. It means that the axis titles were constructed by authors. Thus there were two ways to calculate proportion of variance explained. The first one is to calculate the proportion in total Inertia (variance) including constrained and unconstrained axes. In summary(RDA-Model) this is given in “Eigenvalues, and their contribution to the variance” section. The second way is to calculate the proportion in relation to sum if inertia of constrained axes only (in summary(RDA-Model) it is given in section “Accumulated constrained eigenvalues”). I am sure proportions presented in Fig3 are calculated by the second way. It is not surprised that in the case of Beaufort Sea where after forward selection procedure only two predictors is presented in final model (“Depth” and “Gravel”). Two continuous predictors produce two constrained axes. Thus 100% of total variation explained by constrained axes will be shared between two axes only. In the case of Chukchi Sea the final model include 6 predictors and 100% of total variation explained by constrained axes is shared between 6 axes. That is why this values are senseless. If you are not going to avoid the model selection procedure (which is highly recommended), I propose to provide the proportion of variance explained by each RDA-axes in relation to total Inertia (including inertia of RDA plus inertia of PC).
13. Trichotropis sp. Ariadnaria borealis проверить по Worms.
14. Lines 273-276. Authors consider the nature of gradient revealed by redundancy analysis however no statistical support of the significance of the axes were provided. Это можно сделать с помощью функции anova.cca() из пакета vegan. При этом можно использовать следующие параметры anova(RDA\_Model, by = NULL) даст информацию о статистической значимости всей модели; anova(RDA\_Model, by = “axes”) даст информацию о статистической значимости каждой из constrained axes; anova(RDA\_Model, by = “margin”) даст информацию о статистической значимости отдельных предикторов. I do not recommend to use anova(RDA\_Model, by = “terms”) to avid dependency on sequention of predictors in the model. В любом случае без оценки статистической значимости RDA неясно насколько можно доверять полученным результатам.
15. Figure 4. In some cases the model selection procedure allow to include only singular predictor. It means that only one RDA could be calculated. What is presented in this case on the second axis? You wrote that “When only one environmental variable was selected in the model, modalities and environmental variables only centered around the first axis” (line 969-971). However in the case of Figure 4-a the points at “Reproductive strategy” panel are distributed along RDA2 axis which is impossible since the model include only one predictor. I suppose the second axis in this case is the first unconstrained axis (PC1). It should be somehow explained.
16. Line 337-338. You wrote “while the variability in percent gravel and depth were larger in the Beaufort Sea”. Visually the difference in variation of Gravel seems to be obvious. However there is no statistical significant difference revealed by Levine’s test. Either point out that there is no significant difference or do not discuss this difference.

Надо посчитать fourcorner analysis from ade4