Cryptic mussels in a contact zone: towards to species distribution modelling

**V. M. Khaitov, A. A. Zaychikova, P. Y. Safonov, M. V. Katolikova, M. V. Ivanov, P. P. Strelkov**

## Introduction

Species distribution models (SDM) being a numerical tools describing the relationship between species occurrence and environmental parameters allow to predict distribution patterns of ecologically and economically important species both in space and time (Elith, Leathwick, 2009). The SDM framework is very popular in terrestrial ecology (+++) but the amount of papers devoted to SDMs for marine species has been growing significantly in recent years as well (Robinson et al 2017). SDMs are mostly applied for «good», morphologically distinct, species (e.g. Reiss et al., 2011; Lindegren et al., 2022) that can be involved in routine studies which in the case of marine ecology require numerous samples. However, the increasing evidence for cryptic species among marine animals (Bastrop et al 1998; Fiser et al 2010) makes it urgent to find possibility to build SDMs for sympatric or parapatric cryptic species, and even to infraspecific taxa such as subspecies or phylogenetic lineages within polytypic species. Only few studies have been conducted in this direction in the case of marine organisms. (e.g. Lowen et al 2019; Dennis, Hellberg, 2010).

In the marine realm, the longest scientifically known and best studied complex of cryptic species is the “*Mytilus edulis*”, or blue mussels, complex (+++++++), which includes a number of species that are morphologically similar and capable of hybridization (++++++). Blue mussels are powerful ecosystem engineers playing an important role in structuring of coastal communities (+++) and considered as one of the prime target of aquaculture (+++) in the temperate and subpolar seas. In the North Atlantic sector, the dominant blue mussel species are *M. edulis* (thereafter *ME*) and *M. trossulus* (*MT*) , which form a series of contact zones (zones of sympatry) in different parts of the region from Scotland and the Gulf of Man in the south to Greenland and Svalbard in the north (Varvio et al., 1988; Hilbish et al., 2000; Brooks, Farmen, 2013; Katolikova et al. 2016; Vainola, Strelkov 2011; Wenne et al. 2020). Scientists agree that *ME* and *MT* coexisting in sympatry are ecologically differ (RC, Katolikova et al.) and in particular have different values in aquaculture (+++), but surprisingly little is known about the environmental factors of their segregation i.e. ecological niche partitioning of these species.

On a biogeographic scale, the distribution of *Mytilus* species is thought to be regulated by regional temperature and its correlates (Popovic & Riginos, 2019; Wenne et al., 2020). *MT* is believed to be a more northerly species, more tolerant to cold waters inhabiting areas that once existed in glaciation zones. *ME* potentially can distribute in cold waters but tend to form settlements in more temperate latitudes (Wenne et al., 2020). However the greatest progress in comparative ecological studies of *ME* and *MT* has been made in the zones of species coexistence in the Baltic Sea and in the waters surrounding the Kola Peninsula (White and Barents Seas). In the Baltic Sea species are distributed in space according to salinity: *MT* is confined to the desalinated areas of the inner part of the sea, and *ME* to the more saline areas of the Kattegat and western Baltic. At the boundary of these areas species actively hybridize, hybrids dominate in settlements and the frequencies of two species’ genes change gradually in space along salinty gradient (++++ Kijewski et al., 2019). Another patterns were observed in the White and Barents Seas (++++). Hybridization is limited in this area, hybrids are always rare. Species are distributed in space as a two-level mosaic. In regional scale, the frequency of *MT* is elevated in port areas, possibly due to introduction of this species into the region with ship traffic (+++). Locally, the frequencies of species vary broadly in scale of meters and kilometers but no clear evidences of the association between this patchiness and salinity was proven (VS, 11, Katolikova et al. 2016). Katolikova et al. 2016 revealed that an important factor of local species segregation on the White Sea is the habitat substrate: in mixed settlements the proportion of *MT* are greater on fucoid thallus, *ME* - on abiotic bottom substrates (mud, sand, stones, gravel, etc.). The taxonomic structure of mussel’s settlements in the White Sea can be influenced by biotic factors as well: predators (sea stars) preferentially eat out *MT* decerasing proportion of this species in the mixed settlements (Khaitov et al., 20++; Khaitov et al., 2023). The investigation of the *MT*-*ME* contact zone in the Barents Sea showed that the proportion of *ME* increases with depth when moving from higher intertidal levels to sublitoral (Marchenko et al. 2023). Again, no clear association of taxonomical structure with salinity was revealed in this region (Marchenko et al. 2023).

The other *MT*-*ME* conact zones studied are more similar to the White and Barents Seas than to the Baltic. Hybridization is limited (++++), distribution is mosaic (+++++). Data on the relationship to salinity are contradictory as well. In some cases the higher *MT* proportion was found in dissalinated areas and *ME* proportion increased when moving to areas with normal salinity (e.g. in Bergen area, Ridgway & Naevdal, 2004) but in other (e.g. in Gulf of St. Lawrence) no statistically significant association with salinity was was found but the highest *MT* proportion was revealed in the river estuaries (Mareau et al. 2005). Wave action was also considered as an important factor regulated species distribution in contact zones situated in the northwest Atlantic: *MT* has been shown to gravitate toward more turbulent habitats and *ME* toward quieter areas (Tam, Scrosati, 2014; Comesaña et al., 1999).

A common feature of the attempts described above is the search for some key factor determining the segregation of MT and ME. The gradient of this factor is considered as an environmental axis along which the ecological niches of the two species diverge. However, such a key factor can not always be reliably isolated from other environmental parameters, which can also play an important role in the niches’ divergence. For example, harbors (which considered as a hot spots of *MT* concentrations) are often located in storm-protected areas, frequently with low salinity. This can prevent the isolation of anthropogenic influences from influence of surf and salinity. Potentially, all of the factors considered above (++) could simultaneously influence the distribution of MT and ME or their role can only be assessed when other factors are taken into account.

The best way to resolve this problem is to use SDM as formal and rigorous tool for analysis of ecological niche structure. SDMs allow to consider several plausible factors of species segregation simultaneously and distinguish their contribution in species distribution pattern. However to our knowledge, in the history of *ME* and *MT* studies the SDM-like approach have been applied only twice, by Kijewski et al.(+++) and by Wenne et al. (+++). Both times the machine learning techniques, popular in SDM framework, were used to model the macro-geographic distribution of species (technically, of allele frequencies at taxonomically informative genes) in the space of multiple climatic and oceanographic characteristics available from public databases. The conclusions of these studies are summarized by recognizing temperature and salinity as important factors influencing the macrogeographical distribution of species with *MT* tolerate lower salinities and temperatures than *ME* (Kijewski et al. Wenne et al.). The populations involved in the works cited were scattered over a very wide geographic range, with distances between them reaching hundreds and thousands of kilometers. On the other hand, the gradients of factors regulating species distribution can vary on much smaller spatial scales. For example, salinity can vary significantly on scales of a few kilometers in river estuaries (Van Damme, 2005) or the strength of wave action can vary dramatically on different sides of the same intertidal boulder (Denny ???++++). Does it mean that the patterns of species distribution revealed in large geographic scale would be found at smaller scales as well?

The effective study of cryptic species distribution coexisting in small contact zones, require to handle large numbers of specimens collected from numerous sites. Expensive methods for genetic identification of cryptic species inhibit the development of such research. The finding of a conchological trait associated with the pattern of formation of the nacreous layer in the ligament region (Khaitov et al., 2021; Zolotarev +++) made it possible to identify probabilistically *ME* and *MT* and to estimate the taxonomic structure of settlements without genotyping. It is shown (Khaitov at al., 2023) that in *MT* a poorly developed nacreous layer is usually exposed under the ligament’s nympha, leaving a strip of dark prismatic layer uncovered (mussels of T-morphotype), whereas in the majority of *ME* this pattern is rare and the nacre completely covers the prismatic layer under the nympha (mussels of E-morphotype). Both species are universally differ in the frequencies of this trait, but the magnitude of the differences varies between regions investigated. However in some localities, e.g. in the White Sea, the probability of correct identification of species using this trait is rather high, 86% of correct species identification in settlements where species are represented in equal proportions. The proportion of *MT* (Ptros) in mixed populations is highly correlated with proportion of T-morphotype in the White Sea (Khaitov at al., 2023).

Applying such a cheap morphotype-test allows us to reliably assess the relationship between taxonomic composition and multiple predictors variation of which is well expressed in the White Sea. In this paper, we will consider the case where *ME* and *MT* come into contact forming mixed populations in the Kandalaksha Bay (Vainola, Strelkov, 2011; Katolikova et al. 2016; Khaitov et al, 2021). This contact zone found in the relatively small area is extremely convenient for investigation of environmental factors’ role in the regulation of the *MT:ME* ratio and therefore SDM construction.The salinity gradients are well expressed in this area (Filatov et al., 2005) which is determined by the inflow of several large and small rivers into the sea (Katolikova et al., 2016). Further, the coast of the gulf possesses numerous sheltered areas (inlets and skerries) and areas open to the surf which provide the opportunity to assess the wave impact gradient. There are several active and several abandoned ports (Katolikova et al., 2016) that give an opportunity to trace the role of vessel trafic in species distribution. And finally important is that the small size of the Kandalaksha Bay (Fig. 1) ensures that there are no clear temperature gradients in both warm and cold seasons. Therefore, the influence of temperature, which probably regulates the distribution of *Mytilus* species on a broader geographic scale (Wenne et al. 2020) can be excluded from the discussion.

Using the material collected in this contact zone we will construct model which would describe the association of *MT:ME* ratio with some potentially important factors mentioned above. Basing on the model constructed we would assess the niche divergence between *MT* and *ME* coexisting sympatrically. And finally we will check the ability of the model constructed to predict species distribution pattern in other contact zones.

## Material and methods

–>

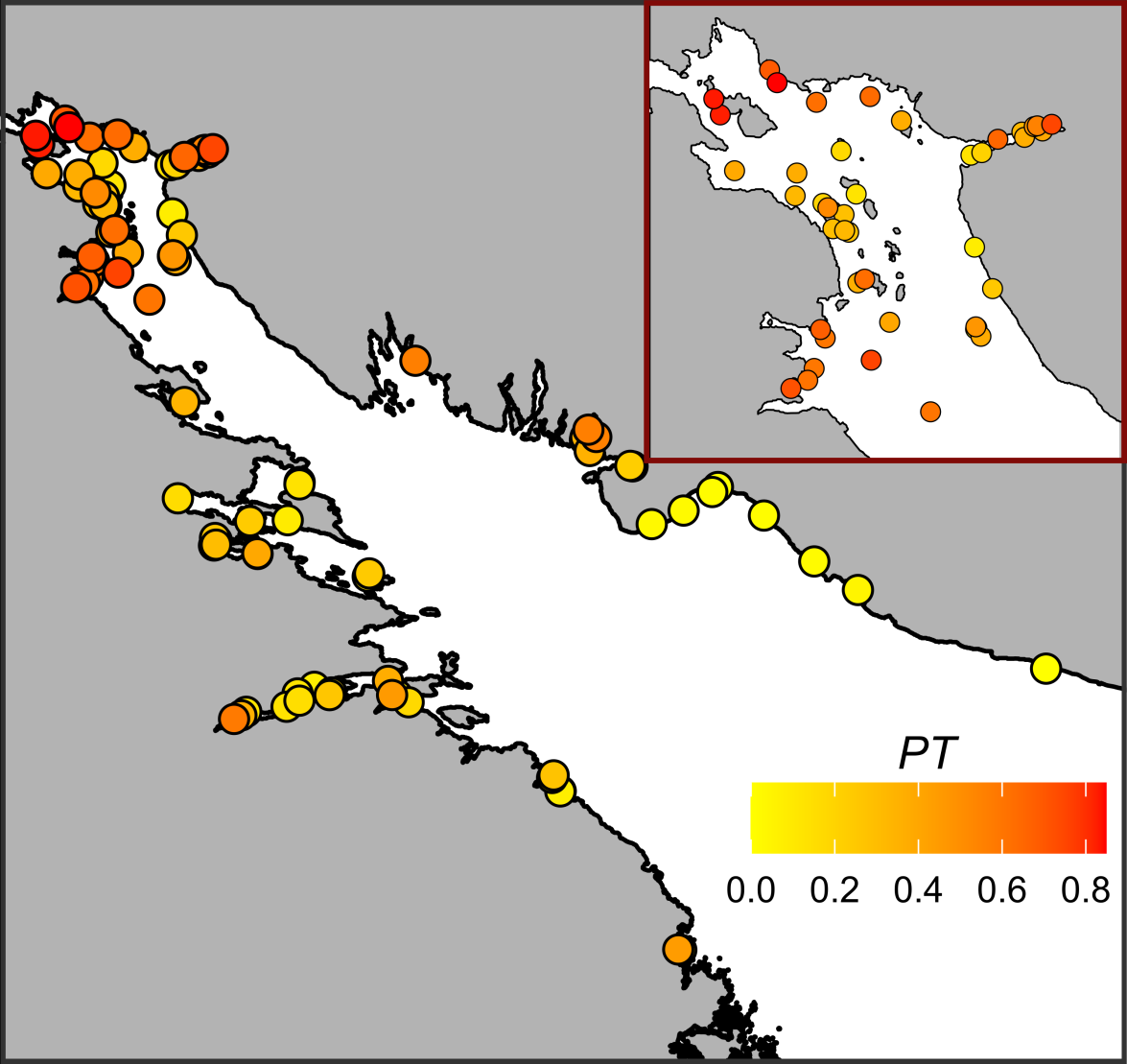
## Results

## Environmental parameters in the Kandalaksha Bay

Two of six ports presented in the area were active at the date of material collection (Fig. 1 A, STable ++ SEM). The distance from sampling sites to the nearest ports ranged from 0.11 to 82 km (Median = 18.7 km) (SFig ++, A). Sites visually categorized as “Sheltered” on average were characterized by lower values of wind fetch than those categorized as “Exposed” (SFig. ++, B). The most exposed sites (maximum values of wind fetch) were located on the western coast of the Kandalaksha Bay, and on the open shores of the islands at the top of the bay (Fig. 1 A). Salinity measured at the time of sampling ranged from 2 to 30 ppm (Median = 19 ppm). The most desalinated areas were the top of the Bay and inlets where rivers flow into (Fig. 1 A). Salinity at sites situated close to “Large” rivers was lower than at sites closer to “Small” ones (SFig. ++, C). Salinity tended to increase with increasing of distance to the mouth of the nearest river (SFig. ++ D) but no association between distance to the nearest port and salinity was found (SFig. +++ E). Sites situated close to ports tended to have lower wind fetch (SFig. +++ F). All correlations between salinity, distance to the nearest river mouth, distance to the nearest port and wind fetch were rather low (STable +++). The largest correlation revealed was between wind fetch and distance to nearest port (r = ++).

## *Ptros* spatial distribution in the Kandalaksha Bay

The maximum *Ptros* was observed in the top of the Kandalaksha Bay itself and in the small bays deeping into the mainland (Fig. 2). At the same time, sites with a lower *Ptros* were located in the immediate vicinity of sites where high *Ptros* values were represented (Fig. 2), reflecting the high variation of taxonomical structure in small spatial scale. Settlements with the lowest *Ptros* were represented in the open areas of the North-West coast.

 At! Надо доделать рисунок, заменить легенду.

## Relationship of *Ptros* to environmental parameters

The values of VIF calculated for the predictors’ set included in the analysis were rather small (the maximal VIF was detected for Fetch: 1.76). Although some pairwise correlation between environmental factors were found (see above) the low VIF values can be interpreted as a negligible collinearity between predictors (i.e. they don’t mask influence of each for other).

To assess the dependency of *PT* on all predictors considered we constructed the *Model 1* (Table 1, R2 = 0.71) which explained 77% of total deviance. The *Model 1* revealed statistically significant dependency of *Ptros* on all predictors except one.

The effective degrees of freedom (*edf*, Table 1) for most smoothers were equal to 1, reflecting the linear dependence of *PT* on such continuous predictors as distance to the nearest port (*MinDistPort*) and fetch (*AverageFetch*). The exception was *Salinity*, where a pronounced curvilinear dependence was revealed.

Table 1 . Parameters of smoothers and coefficients of parametric terms for the Model describing dependency of Ptros on predictors.

| Model term | edf | ref.edf | Chi.sq | p-value |
| --- | --- | --- | --- | --- |
| s(Salinity) | 2.4 | 9 | 396.7 | 0.0033 |
| s(DistRiver) | 0.0 | 9 | 0.0 | 0.6724 |
| s(Fetch) | 0.9 | 9 | 88.2 | 0.0417 |
| s(DistPort) | 1.0 | 9 | 276.2 | 0.0016 |
| Random effect s(Site) | 74.4 | 92 | 453.6 | 0.0000 |
| Model term | Parameter estimate | SE | z-statistic | p-value |
| (Intercept) | -1.7 | 0.1 | -11.8 | 0.0000 |
| Substrate(Algae) | 0.9 | 0.1 | 14.6 | 0.0000 |
| RiverSize(Large) | 0.4 | 0.2 | 2.6 | 0.0091 |
| PortStatus(Active) | 1.0 | 0.2 | 5.7 | 0.0000 |

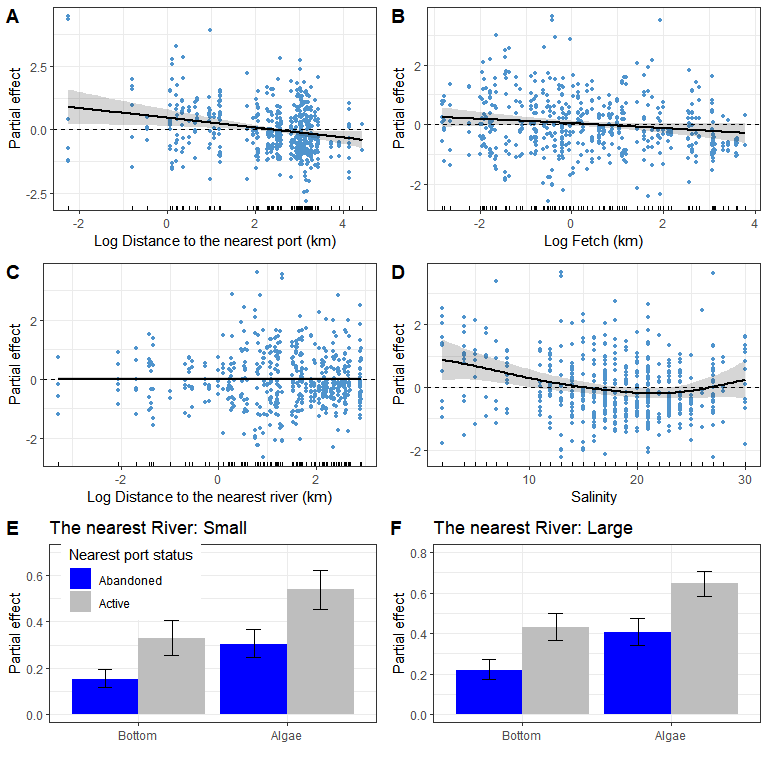


Fig. 3. Predictions of the *Model 1* describing the dependency of *Ptros* on distance to the nearest port (A), average fetch (B), distance to the nearest river (C), salinity (E) and substrate type, nearest port status and size of the largest river (E, F). The gray ribbons and wiskers demonstrate 95% confidence interval. The dotted horizontal lines in panels A-D represent the zero smoother’s value. Points on panels A-D reflect partial residuals but not raw data for each sample.

**Возможно, что для лучшей визуализации надо пересчитать шкалу OX из логарифма в обычные занчения. Это не лучший ход, но можно, если так трудно.**

According to the *Model 1* (Table 1, Fig. 3) *Ptros* has a tendency to decrease with increasing of distance to the nearest port (Fig. 1 A). The nearest port status had also significant influence on *Ptros*: the value was higher when port status was “Active” (Fig. 3 E, F). The dependency of *Ptros* on fetch (Fig. 3, B) was negative. No significant association between *Ptros* and the distance to the neares river was was found (Table 2, Fig 3 C).

The curvilinear dependence of *Ptros* on salinity (Fig. 3, D) can be described as follows. The predicted value of *Ptros* decrease when salinity moving from minimal values to apr 20-22% when it reached minimum. When salinity increases after this point the *Ptros* value increase as well. However, at low (less than 15 ppt) and high (more than 27 ppt) salinity, *Ptros* becomes higher than zero. This low and high salinity (when the Model 1 predicts *Ptros* higher than zero) was observed at 28 and 8 sites respectively (Fig. 1 B). For other 59 sites, with intermediate salinity, predicted *Pros* was less than zero (Fig. 3 D). All low salinity sites were represented in the top parts of different inlets (Fig. 1 B) but no clear pattern revealed in the distribution of sites with high salinity (Fig. 1 B).

According to the *Model 1* the value of *Ptros* is dramatically higher on the Algae substrate than on the Bottom one (Table 1, Fig. 3 E, F). This value is higher in those sites which is close to active ports than to abandoned ones. Finally, the *Ptros* value is slightly higher at sites situated close to large rivers than to small ones.

## Dependency of *Ptros* on substrate type

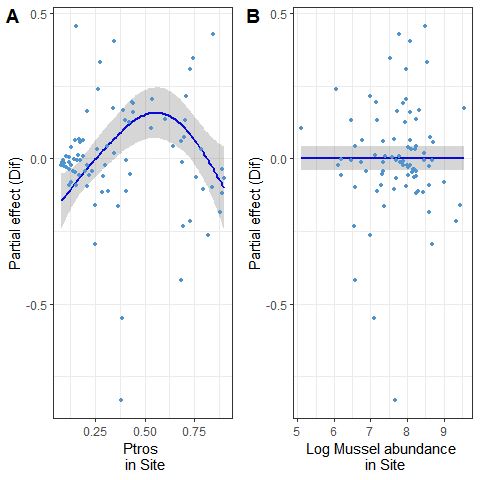


Fig. 4. The dependence of difference between *PT* on Algae and Bottom subatrates on PtrosSite (A) and total mussel abundance (B).

The difference between *Pros* on Algae and Bottom substrates was significantly dependent on *Ptros* in site (STable +, Fig. 4 A). The dependency revealed was curvelinear (edf = 2.7, STable +) with minimal *Dif* values close to zero in sites with minimal and maximal values of *PtrosSite*. The maximal *Dif* values were associated with intermediate *PtrosSite*. This pattern reflect the prevalence of *MT* at Algae substrate and prevalence of *ME* at Bottom substrate in those sites where both morphotypes are presented in equal amount. No significant dependency was revealed with total absolute mussel abundance (STable +, Fig. 4 B).

## Assessment of model’s predictive power

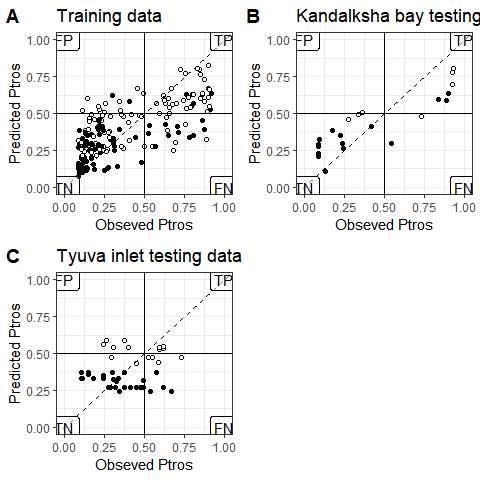


Fig. 5. Comparison of *PT*-values predicted by *Model 2* and obsreved proportion of T-morphotype for Bottom (black circles) and Algae (white circles) samples. The horizontal and vertical solid lines reflects the boundary between *MT*- and *ME*-dominated sites. Dotted lines represent the position of perfect fit when observed and predicted values are equal. Labels in the plot corners mark the quadrants with false positive (FP), true positive (TP), true negative (TN) and false negative (FN) predictions.

We plotted values predicted by the *Mode 1* against observed ones (Fig. 5).The diagram constructed for training dataset reveal relatively good correspondence between predicted and observed *Ptros*. The AUC value was rather high (AUC = 0.87) It is important to note that the most false positive preditcions for main, training, data set were made for Algae samples whereas false negative predictions were made both for Algae and Bottom samples.

In the case of testing dataset from the Kandalaksha bay (Fig. 1 B) the most sites were correctly classified by the *Model 1*. The AUC for this data was close to 1 (AUC = 0.92) indicating high predictive ability of the *Model 1*. This model being applied to testing dataset from the Tyuva inlet diplay lesser predictive power (AUC = 0.66). Only few samples were classified as true positive and the most of them were classified as true negative. All false positive samples were associated with Algae bat equal numbers of Algae and Bottom samples were in false negative region.

## Discussion

–>

–>

–>

## Literature cited

Лоция Белого моря / Гидрографическое управление ; общ. ред. А. Н. Рождественского. – 5-е изд. – Ленинград : Издание Гидрографического управления, 1932. – LXXII, 566 с., 22 вкл. л. ил., карт., схем.

# Supplementary electronic materials (SEM)

–>

–>