Итого. В качестве данных для тестового датасета взяты следующие популяции (см. таблицу). Для BL критерием отбора было то, что мидии собраны за пределами Кольского залива в эстуариях рек, для BH - за пределами Кольского залива но без заметного опреснения, для W - отобрано 4 популяции, в которых Ptros была близка к 0.2, 0.4, 0.6 и 0.8. Все эти данные НЕ будут участвовать в процессе подбора парамтеров регерссионных моделей или построения калькуляторов.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subset | pop | N\_Tm\_T | N\_Em\_T | N\_Tm\_E | N\_Em\_E | Ptros |
| W | chupa | 29 | 30 | 3 | 22 | 0.7023810 |
| W | kovda | 6 | 2 | 0 | 41 | 0.1632653 |
| W | rya | 23 | 5 | 18 | 28 | 0.3783784 |
| W | umba\_pil | 34 | 0 | 5 | 5 | 0.7727273 |
| BL | ivan2 | 4 | 2 | 0 | 45 | 0.1176471 |
| BL | ivan3 | 33 | 7 | 1 | 11 | 0.7692308 |
| BL | yokanga | 9 | 5 | 1 | 44 | 0.2372881 |
| BH | bukhtovka | 7 | 2 | 18 | 13 | 0.2250000 |
| BH | dz\_banka | 7 | 1 | 23 | 57 | 0.0909091 |
| BH | dz\_lit | 2 | 1 | 45 | 41 | 0.0337079 |
| BH | ivan\_buy | 15 | 0 | 7 | 0 | 0.6818182 |
| BH | ivan\_us | 8 | 0 | 8 | 33 | 0.1632653 |
| BH | yarn02 | 6 | 0 | 7 | 10 | 0.2608696 |

На всякий случай привожу таблицу с описанием тех популяций, которые ВОШЛИ в моделинг датасет.

Эти две таблички могут быть предложены в качестве первичных данных для приложения к статье.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subset | pop | N\_Tm\_T | N\_Em\_T | N\_Tm\_E | N\_Em\_E | Ptros |
| W | berzakol | 3 | 1 | 0 | 43 | 0.0851064 |
| W | kal | 35 | 6 | 4 | 1 | 0.8913043 |
| W | luv\_korg | 2 | 1 | 0 | 40 | 0.0697674 |
| W | luv\_mat | 6 | 0 | 4 | 38 | 0.1250000 |
| W | niva\_sl | 16 | 2 | 7 | 16 | 0.4390244 |
| W | oenij | 30 | 5 | 0 | 5 | 0.8750000 |
| W | padan | 0 | 1 | 0 | 29 | 0.0333333 |
| W | porya | 4 | 2 | 0 | 36 | 0.1428571 |
| W | salnij | 32 | 4 | 0 | 3 | 0.9230769 |
| W | umba | 37 | 15 | 9 | 44 | 0.4952381 |
| W | umba\_06 | 0 | 0 | 0 | 29 | 0.0000000 |
| W | umba\_bridge | 0 | 0 | 0 | 20 | 0.0000000 |
| W | umba\_kamni | 0 | 0 | 0 | 43 | 0.0000000 |
| W | umba\_pikut | 7 | 6 | 0 | 17 | 0.4333333 |
| W | umba\_pioner | 0 | 1 | 0 | 39 | 0.0250000 |
| W | umba\_sovhoz | 1 | 4 | 0 | 33 | 0.1315789 |
| W | vor1 | 1 | 2 | 8 | 32 | 0.0697674 |
| W | vor2 | 0 | 2 | 0 | 37 | 0.0512821 |
| W | vor5 | 2 | 0 | 1 | 40 | 0.0465116 |
| W | voronya | 5 | 2 | 0 | 39 | 0.1521739 |
| BL | abram | 11 | 7 | 2 | 25 | 0.4000000 |
| BL | belok | 14 | 7 | 6 | 39 | 0.3181818 |
| BL | belok2 | 45 | 4 | 1 | 0 | 0.9800000 |
| BL | fr | 9 | 7 | 4 | 35 | 0.2909091 |
| BL | mi | 41 | 25 | 1 | 18 | 0.7764706 |
| BL | nm | 0 | 0 | 0 | 37 | 0.0000000 |
| BL | nm\_last | 16 | 2 | 4 | 28 | 0.3600000 |
| BL | sevsk | 21 | 8 | 0 | 25 | 0.5370370 |
| BL | zmis | 45 | 1 | 3 | 8 | 0.8070175 |
| BH | banka | 3 | 5 | 40 | 36 | 0.0952381 |
| BH | kuvsh | 20 | 2 | 16 | 2 | 0.5500000 |
| BH | seredi | 55 | 1 | 38 | 4 | 0.5714286 |
| BH | seredi\_sub | 39 | 17 | 40 | 20 | 0.4827586 |
| BH | tu\_old | 37 | 1 | 4 | 3 | 0.8444444 |
| BH | ustie | 51 | 7 | 31 | 17 | 0.5471698 |
| BH | ustie\_sub | 27 | 10 | 43 | 39 | 0.3109244 |
| BH | vol | 32 | 10 | 10 | 13 | 0.6461538 |

# 

# Regression models parameters

In all models outcome variables were supposed as binomialy disributed (). Logit was used as a linck function in all models. In all cases initially full models including all terms and their inteactions were constructed. After the full models were constructed they were simplified accordingly to backward selection protocol (Zuur et al., 2009).

The full models list

Model1:

Model2:

Model3:

Model4:

Model5:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Terms | Estimate | SE | z-statistic | p-value |
| **Model 1** |  |  |  |  |
| (Intercept) | -2.9 | 0.17 | -17.08 | < 0.001 |
| Ptros | 5.2 | 0.35 | 14.73 | < 0.001 |
| Subset(BL) | 0.5 | 0.30 | 1.67 | 0.096 |
| Subset(BH) | 2.6 | 0.28 | 9.26 | < 0.001 |
| Ptros:Subset(BL) | -1.1 | 0.55 | -2.01 | 0.044 |
| Ptros:Subset(BH) | -2.3 | 0.58 | -3.88 | < 0.001 |
| **Model 2** |  |  |  |  |
| (Intercept) | -4.3 | 0.42 | -10.17 | < 0.001 |
| Ptros | 4.6 | 0.96 | 4.81 | < 0.001 |
| Subset(BL) | 0.8 | 0.88 | 0.96 | 0.335 |
| Subset(BH) | 4.4 | 0.89 | 4.92 | < 0.001 |
| Species(*M.trossulus*) | 4.3 | 0.50 | 8.47 | < 0.001 |
| Ptros:Subset(BL) | -2.2 | 1.78 | -1.25 | 0.211 |
| Ptros:Subset(BH) | -3.2 | 1.78 | -1.82 | 0.069 |
| Ptros:Species(*M.trossulus*) | -2.3 | 0.97 | -2.41 | 0.016 |
| Subset(BL):Species(*M.trossulus*) | -1.3 | 0.92 | -1.40 | 0.162 |
| Subset(BH):Species(*M.trossulus*) | -4.9 | 0.79 | -6.21 | < 0.001 |
| Ptros:Subset(BL):Species(*M.trossulus*) | 3.0 | 1.78 | 1.70 | 0.089 |
| Ptros:Subset(BH):Species(*M.trossulus*) | 5.5 | 1.57 | 3.50 | < 0.001 |
| sd\_(Intercept).pop | 0.8 |  |  |  |
| **Model 3** |  |  |  |  |
| (Intercept) | 2.7 | 0.18 | 15.42 | < 0.001 |
| Ptros | -1.7 | 0.33 | -5.04 | < 0.001 |
| Subset(BL) | -0.9 | 0.32 | -2.67 | 0.008 |
| Subset(BH) | -3.2 | 0.27 | -11.82 | < 0.001 |
| Ptros:Subset(BL) | 1.2 | 0.55 | 2.17 | 0.03 |
| Ptros:Subset(BH) | 3.5 | 0.53 | 6.62 | < 0.001 |
| **Model 4** |  |  |  |  |
| (Intercept) | 3.8 | 0.25 | 15.09 | < 0.001 |
| Morph(T) | -3.7 | 0.36 | -10.33 | < 0.001 |
| Ptros | -5.1 | 0.52 | -9.89 | < 0.001 |
| Subset(BL) | -0.5 | 0.42 | -1.14 | 0.253 |
| Subset(BH) | -1.7 | 0.40 | -4.12 | < 0.001 |
| Morph(T):Ptros | 8.4 | 0.63 | 13.34 | < 0.001 |
| Morph(T):Subset(BL) | 0.4 | 0.39 | 0.91 | 0.361 |
| Morph(T):Subset(BH) | -1.0 | 0.33 | -2.95 | 0.003 |
| Ptros:Subset(BL) | 0.8 | 0.76 | 1.06 | 0.288 |
| Ptros:Subset(BH) | 2.3 | 0.75 | 3.00 | 0.003 |
| sd\_(Intercept).pop | 0.1 |  |  |  |
| **Model 5** |  |  |  |  |
| (Intercept) | -2.9 | 0.17 | -16.94 | < 0.001 |
| PT | 6.1 | 0.40 | 15.25 | < 0.001 |
| Subset(BL) | 0.6 | 0.31 | 1.80 | 0.072 |
| Subset(BH) | -0.3 | 0.45 | -0.64 | 0.522 |
| PT:Subset(BL) | -0.3 | 0.74 | -0.43 | 0.669 |
| PT:Subset(BH) | -1.9 | 0.69 | -2.70 | 0.007 |

In all models Ptros - proportion of *M.trossulus* in population; Subset - sampling area (*W*, *BL*, *BH*); PT - proportion of T-morphotype in population; Species - mussel genotype (*M.edulis* or *M.trossulus*). *M.edulis* and *W* where used as basic levels for categorical predictors in all models.

## Regression models visualisation

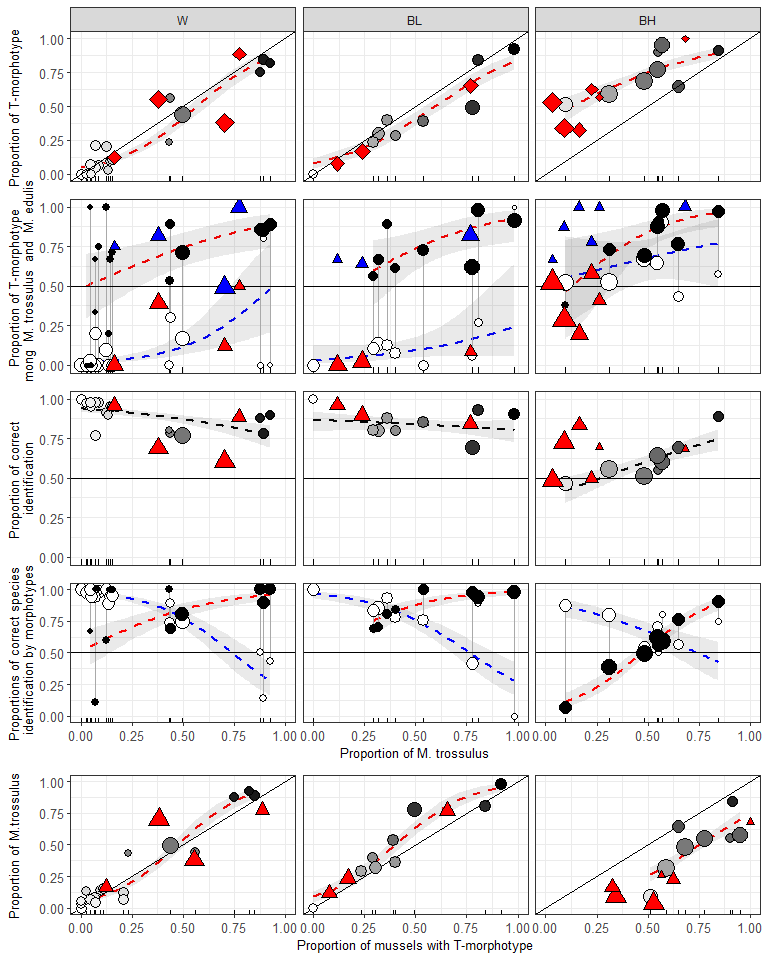


Figure ++. Visualisation of regression models. Initial data are presented as proportions of positive outcome in particular populations. Size of points is proportiol to number of mussels in the particular sample. (A) Model1: the fill intensity is proportiol to Ptros. (B) Model2: filled points - *M.trossulus* with T-morphotype; emty points - *M.edulis* with T-morphotype. (C) Model3: filled points - *M.trossulus* with T-morphotype; empty points - *M.edulis* with E-morphotype.

## Strategy for finding of calibration samples

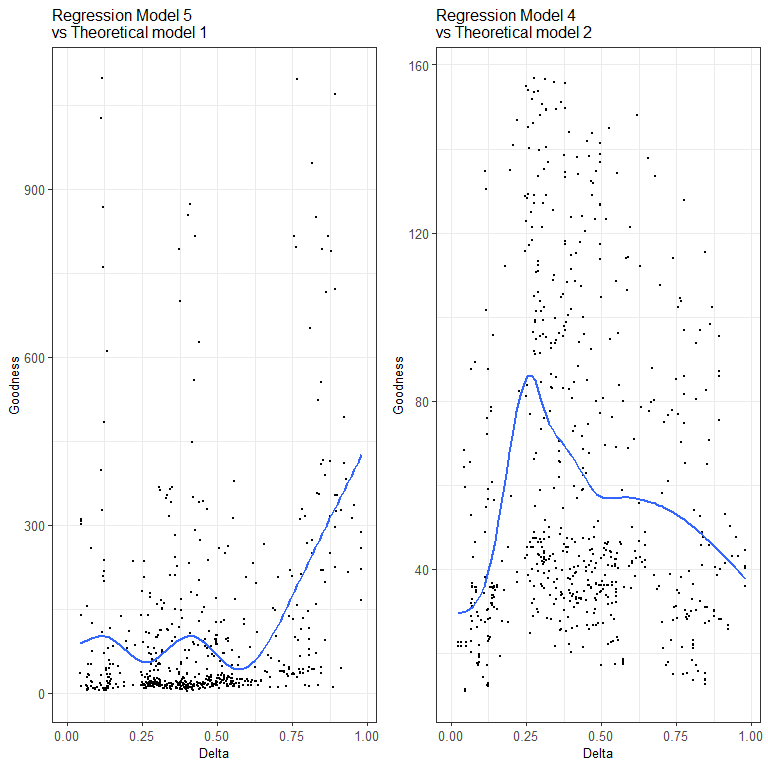


Figure +. Correspondence between regression and theoretical models. Each point corresponds to one of the possible pairs of populations from modelling data set. OX axis represents the differencу in genetic structre for each pair of populations. OY axis represents correspondence between prediction of regression model and theoretical model. Lines represent loess-smoother. (A ) Model 5 describing the dependence of proportion of M.trossulus (Ptros) on proportion of T-morphotype (P\_T) ; (B) Model 4 describing the dependence of probability of correct species identification (Pcorrect) on proportion of M.trossulus (Ptros) and morphotype (Morph).

## Generalization of data

Дальнейший анализ должен показать, как ведут себя теоретические калькуляторы, подобранные в соответствии со стратегией, описанной выше. Но есть мнение (**которого я не придерживаюсь!**), что удобнее было бы приводить данные по генерализованной модели. Под “генерализованной” моделью понимается модель, в которой фактор “Subset” имеет меньшее число градаций (или вовсе исключен из модели).

Для того, чтобы проверить можно ли как-то генерализовать модель, мы должны проверить на modelling data set, приводит ли объединение данных из разных регионов к существенным изменениям в структуре модели. Единственный (на мой взгляд) путь - это сравнить по информационным критериями несколько моделей-кандидатов:

1. W + BL + BH (все регионы входят в модель)
2. (WBL) + BH (объединены W BL)
3. (WBH) + BL (объединены W BH)
4. W + (BLBH) (объединены BL BH)
5. (WBLBH)

Поблема заклчается в том, какую из моделей гонять в этом анализе. Поскольку наиболее адекватно, на мой взгляд, отражает работоспособность “morphotype-test” Model 4, то на ней и будет основан дальнейший анализ.

|  |  |  |
| --- | --- | --- |
|  | df | AIC |
| Model\_4\_fil\_cand\_1 | 11 | 1624.286 |
| Model\_4\_fil\_cand\_2 | 8 | 1620.260 |
| Model\_4\_fil\_cand\_3 | 8 | 1657.792 |
| Model\_4\_fil\_cand\_4 | 8 | 1656.967 |
| Model\_4\_fil\_cand\_5 | 5 | 1664.156 |

Приведенные данные показывают, что объединене W и BL не приводт к ухудшению модели.

## Theoretical models

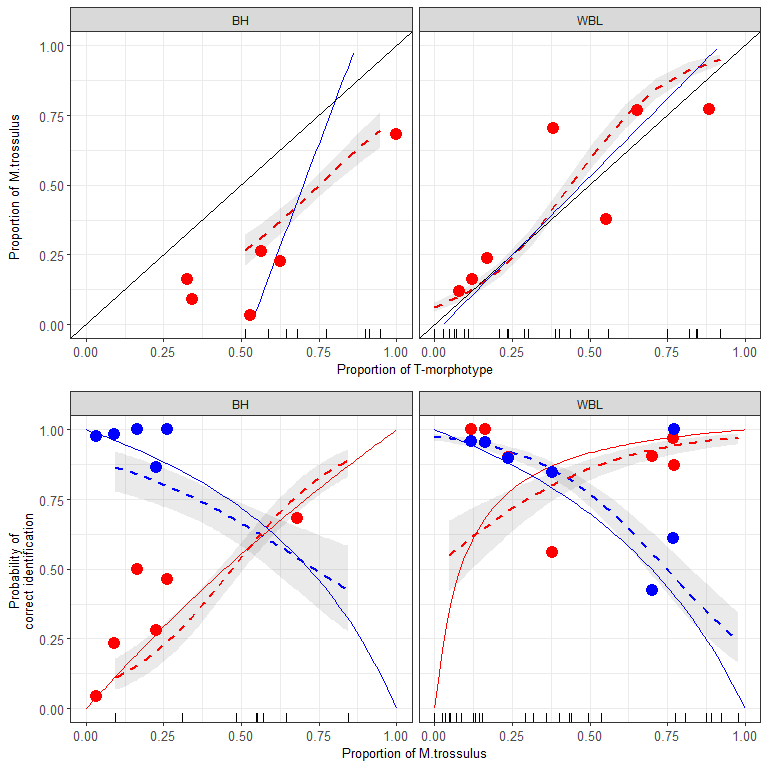


Figure +. Visualisation of regression models and theoretical models. Points represent the testing data set. На этом рисунке пунктирная линия - это эмпирическая регрессионная модель, построенная по modelling dataset. Сплошная линия - это пердсказания соответствующих теоретических моделей, которые построены на основе калибровочных выборок. Калибровочные выборки были отобраны на основе стратегии, описанной в предыдущем разделе. То есть для теоретической модели №1 - это максимально различные по генетическому составу выборки. Для теоретической моедли №2 это выборки имеющие Ptros близкое к 0.5.

Можно привести еще вот такую таблицу, в которой будут характеристики калибровочных выборок. Эти данные можно привести в виде врезки на графиках. Надо подумать.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| pop | Subset | N\_E | N\_T | P\_T\_ME | P\_T\_MT |
|  |  |  |  |  |  |
| belok2 | BL | 1 | 49 | 0.02 | 0.90 |
| nm | BL | 37 | 0 | 0.00 |  |
| banka | BH | 76 | 8 | 0.48 | 0.04 |
| tu\_old | BH | 7 | 38 | 0.09 | 0.82 |
|  |  |  |  |  |  |
| niva\_sl | W | 23 | 18 | 0.17 | 0.39 |
| porya | W | 36 | 6 | 0.00 | 0.10 |
| kuvsh | BH | 18 | 22 | 0.40 | 0.50 |
| ustie | BH | 48 | 58 | 0.29 | 0.48 |