*ДОПОЛНИТЕЛЬНЫЙ МАТЕРИАЛ*

УДК 595.384.2

**Структурный анализ широкопалости у крабов-стригунов статистическим языком R**

*Артеменков Д.В.1, Сологуб Д.О.1, Кивва К.К. 1*

*1* Всероссийский научно-исследовательский институт рыбного хозяйства и океанографии (ФГБНУ «ВНИРО»), г. Москва

e-mail: artemenkov@vniro.ru

Последовательность шагов для разделения особей на группы можно записать в следующем виде:

1. Считывание данных.

2. Визуализация данных для определения наличия и числа «выбросов». В случае отсутствия «выбросов» переход к пункту 5.

3. Расчёт расстояния Махаланобиса.

4. Исключение строк данных с расстоянием Махаланобиса, превышающим выбранное критическое значение.

5. Настройка модели и поиск оптимального решения.

6. Визуализация результата.

7. Запись сгруппированных значений для дальнейшего анализа.

Пример реализации данного алгоритма на языке программирования R представлен ниже.

Пояснения:

library() – вызов пакетов *mclust , dplyr , xlsx* и *lrrcovHD* используется сразу после запуска «R»;

setwd() – установление пути к файлу, содержащему исходных данных;

Crabs\_baird\_A.xlsx – произвольное имя файла, содержащего исходные данные:

CW – ширина карапакса,

LL – высота клешни,

Ln\_CW – логарифм ширины клешни,

Ln\_LL – логарифм высоты клешни.

После применения гаусовской модели получили массив данных, в котором каждой особи был присвоен числовой код 1 (УПС) или 2 (ШПС). Полученные данные обработали с помощью дискриминантного анализа в программе «STATISTICA» для исправления ошибочных определений УПС и ШПС.

Результаты разделения самцов крабов-стригунов на УПС и ШПС с применением вышеописанного алгоритма представлены на рисунке 1. При выполнении повторного разделения, для каждой единицы запаса был получен абсолютно идентичный результат.

library(mclust)

library(dplyr)

library(xlsx)

library(rrcovHD)

setwd("C:/R/Chionoecetes")

crabs <- read.xlsx2("Crabs\_2021\_baird.xlsx", sheetIndex = 1,

colClasses = c("character",

rep("numeric", 7)))

head(crabs)

summary(crabs)

crabs <- crabs %>%

filter(!is.na(Num)) %>%

select(Num, NN, CW, LL, Ln\_CW, Ln\_LL)

plot(crabs$Ln\_CW, crabs$Ln\_LL)

# Visualisation of distributions prior to further work

X <- crabs[, c(5, 6)]

dens <- densityMclust(X, G = 2)

summary(dens)

plot(dens, what = "density", data = X, grid = 200,

points.cex = 0.5, drawlabels = FALSE)

plot(dens, what = "density", type = "image",

col = "steelblue", grid = 200)

# Outliers detection with Mahalanobis distance

# Note this step is not necessary in some cases

obj <- OutlierMahdist(crabs[, c(5, 6)])

summary(as.factor(obj@flag))

crabs$flag <- obj@flag

crabs <- crabs %>%

arrange(desc(flag), Num)

head(crabs)

tail(crabs)

X <- crabs[crabs$flag == 1, c(5, 6)]

# The list of available models

model\_names <- c("EEE", "EVE", "VEE", "EEV", "VEV", "VVE", "VVV")

# In some cases EEE (1) works with the best BIC,

mod\_num <- 1

gmm.mclust <- Mclust(X, 2,

modelNames = model\_names[mod\_num])

plot(X[, 1], X[, 2], col = gmm.mclust$classification)

# Saving results of classification with zeros for outliers

crabs$classification <- 0

crabs$classification[1:dim(X)[1]] <- gmm.mclust$classification

write.csv(crabs, file = "02\_Crabs\_2021\_baird\_KG\_classified.csv", row.names = F)

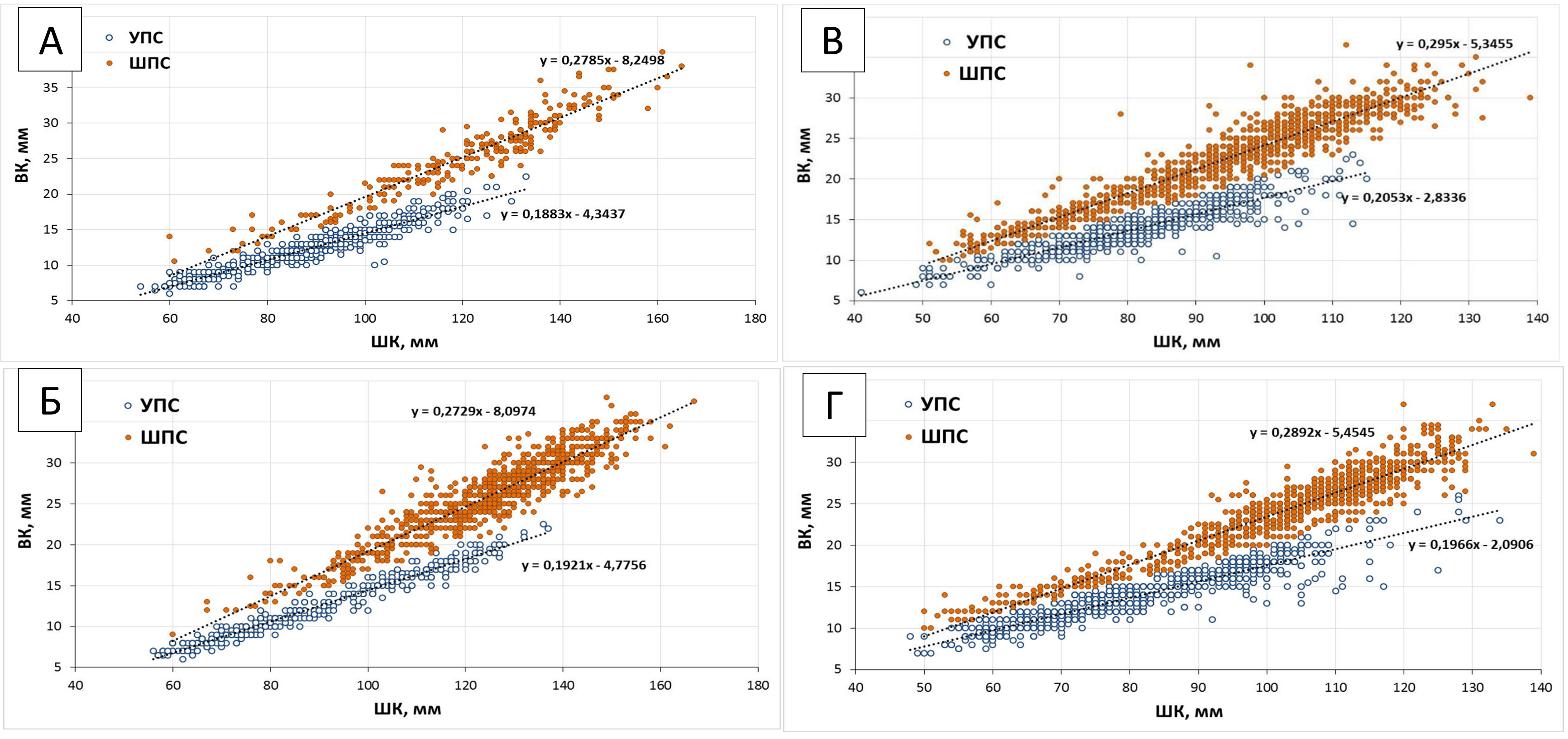


Рисунок 1 – Зависимость ширины карапакса от высоты клешни для УПС и ШПС краба-стригуна бэрди в зоне А, краба-стригуна бэрди в зоне Б, краба-стригуна опилио в зоне В и краба-стригуна опилио зоне Г, рассчитанная с применением нового алгоритма разделения УПС и ШПС по результатам учетной съемки

Таким образом, применение моделей гауссовых смесей, реализованных в пакете mclust для языка программирования R, позволяет в значительной степени автоматизировать и ускорить процедуру разделения самцов крабов-стригунов на УПС и ШПС. Кроме того, данный алгоритм позволяет получить полностью воспроизводимый результат. В дальнейшем планируется выполнить сравнение всех трех существующих методов разделения самцов крабов-стригунов на УПС и ШПС для оценки различия получаемых результатов. Кроме того, с целью дальнейшего совершенствования нового алгоритма, планируется дополнить скрипт на языке программирования R, погрузив в него блок дискриминантного анализа, исключив необходимость применения дополнительного анализа в программе «STATISTICA».