WP5_Edgerton

Meghan Edgerton

2023-02-17

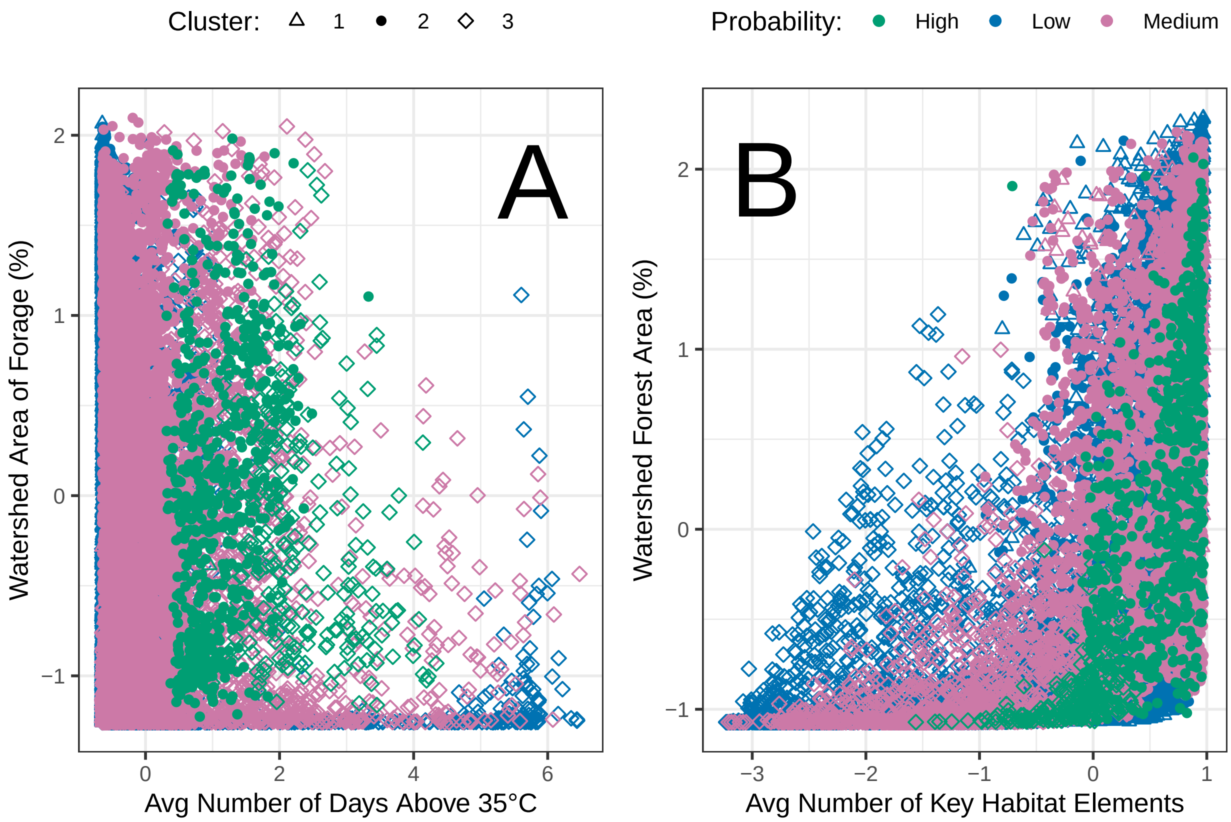

Figure 1. K-means cluster analysis with 3 groups of four continuous variables displaying factors that could possibly increase the chance of seeing a feral pig in the United States. While the plots don't display the best separation of the clusters, Plot B shows slight separation along the y-axis indicating that the clusters are relying more on the percent of watershed area with forest cover. This data was pulled from a 2015 study on occurrences of feral pigs in the United States and the grouping variable (probability) was defined using a relative probability variable from the raw data.

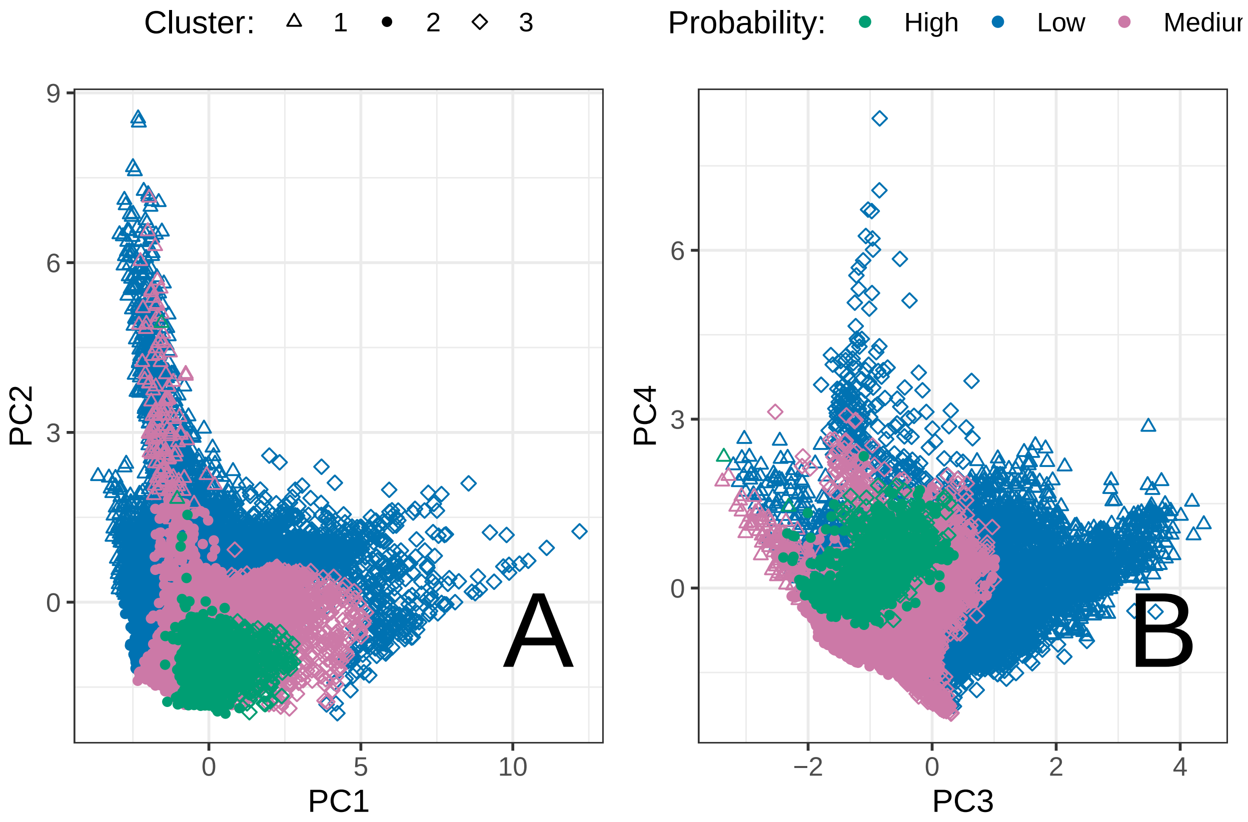

Figure 2. Principal components analysis that compares the probability of encountering a feral pig in the United States to the K-cluster assignments. Plot B shows better separation of the three clusters along PC3, and both plot A and B show distinct separation of the probability groups. Despite this, the clusters do not line up well with the probability groups in either plot.

Interpretation of Results

Overall, it does not seem that the true probability of encountering a feral pig in the United States lines up with the cluster assignments very well. The factors that seem to align the best with the cluster assignments were the average number of days over 35 degrees Celsius, and the average number of key habitat elements nearby. For the principal components analysis the groups did not line up well to the actual probability groups.

R Code Used To Make These Plots

```
library(dplyr)
library(ggplot2)
library(rmarkdown)
library(readxl)
library(reshape2)
library(fclust)
library(GGally)
library(gridExtra)

options(scipen = 9999)

path <- "/Users/meghanedgerton/Documents/DataVizINFO526/wildpig_sdm_data.xlsx"

data <- read_xlsx(path, .name_repair = "universal")
```

```
pigs <- data %>% select(Probability, zAvgAb35, zDistH2O,
                        zPercForest, zForage, zHeterog,
                        zSnowDep, zAvgBeln4) %>% na.omit()


pigs2 <- pigs %>% select(-Probability)

# K-means Clustering & adding clusters to data set

pigs_k <- kmeans(pigs2, centers = 3)

pigs$KCluster <- pigs_k$cluster
```

```
# Plot 1 - Hot Weather & Forage
g1 <- ggplot(pigs, aes(x = zAvgAb35, y = zForage, shape = as.factor(KCluster),
                 col = Probability)) +
  geom_point() +
  scale_color_manual(values = c("#009E73", "#0072B2", "#CC79A7"),
                     name = "Probability") +
  scale_shape_manual(values = c(2,16,5), name = "Cluster:") +
  theme_bw() +
  labs(x = "Avg Number of Days Above 35°C",
       y = "Watershed Area of Forage (%)") +
  theme(legend.position = "top", text = element_text(size = 10)) +
  guides(col = "none") +
  annotate(geom = 'text', label = 'B ', x = 6, y = 1.75, size = 14)
```

```r
# Plot 2 - Heterogeneity & Percent Forest
g2 <- ggplot(pigs, aes(x = zHeterog, y = zPercForest, shape = as.factor(KCluster),
                 col = Probability)) +
  geom_point() +
  scale_color_manual(values = c("#009E73", "#0072B2", "#CC79A7"),
                     name = "Probability:") +
  scale_shape_manual(values = c(2,16,5), name = "Cluster:") +
  theme_bw() +
  labs(x = "Avg Number of Key Habitat Elements",
      y = "Watershed Forest Area (%)") +
  theme(legend.position = "top", text = element_text(size = 10)) +
  guides(shape = "none") +
  annotate(geom = 'text', label = 'B ', x = -2.75, y = 1.95, size = 14)

grid.arrange(g1,g2, ncol = 2)
```

```r
pigs_table <- pigs %>% group_by(Probability, KCluster) %>%
  tally() %>% dcast(Probability~KCluster, fill = "")
```

```r
# PCA

pca <- prcomp(pigs2, scale. = TRUE)

pca.output <- data.frame(Kcluster = pigs_k$cluster,
                         Probability = pigs$Probability,
                         pca$x)
```

```r
p1 <- ggplot(pca.output, aes(x = PC1, y = PC2,
                             shape = as.factor(Kcluster),
                             col = Probability ))+
  geom_point() +
  scale_color_manual(values = c("#009E73", "#0072B2", "#CC79A7"), name = "Probability:")+
  scale_shape_manual(values = c(2, 16, 5), name = "Cluster:")+
  theme_bw()+
  labs(x = "PC1", y = "PC2")+
  theme(legend.position = "top", text = element_text(size = 12))+
  guides(col = "none") +
  annotate(geom = 'text', label = 'A ', x = Inf, y = -Inf,
           hjust = 1, vjust = -1, size =14)
```

```r
p2 <- ggplot(pca.output, aes(x = PC3, y = PC4,
                             shape = as.factor(Kcluster),
                             col = Probability ))+
  geom_point() +
  scale_color_manual(values = c("#009E73", "#0072B2", "#CC79A7"), name = "Probability:")+
  scale_shape_manual(values = c(2, 16, 5), name = "Cluster:")+
  theme_bw()+
  labs(x = "PC3", y = "PC4")+
  theme(legend.position = "top", text = element_text(size = 10))+
  guides(shape = "none") +
  annotate(geom = 'text', label = 'B ', x = Inf, y = -Inf,
           hjust = 1, vjust = -1, size =14)
```

```r
grid.arrange(p1,p2, ncol = 2)

rot <- data.frame(round(pca$rotation, 5)) %>% mutate(var = row.names(.))

PC1 <- rot %>%
  slice(which.max(abs(PC1)))
```