Probabilility of detections and sample occurance

Richard Erickson, Chris Merkes, and Erica Mize

2017

Data source and parameter values

The observation and detection probabilities are based upon "rough" analysis from USGS eDNA samples in Pools 17 and 19. These samples were taken during June, which should be a period with low detentions based upon the biology and seasonal movement of the fish. The lowest observed occurance probabilities were 0.06 and the highest was 0.42. These values are used for the basis of a sample containing DNA. The probabilities of detecting DNA within a sample ranged from 0.3 to 0.4. These values are used for the basis of detecting DNA using an assay for a given sample.

Probability of detecting a species (Occupuancy only)

The first analysis we run estimates the probability of detecting a species at site. This does not allow us to distinguish different densities. Rather it simply informs if a species is present at a site.

We first write a function that estimates the probability of detecting a species assuming different numbers of samples, J; probabilities of samples containing DNA, `theta`; different numbers of assay replicates, K; and different detection probabilities for the assay, `pDetection` (we choose to use `pDetection` rather than `p` to have a variable that was easier to find in our code). The probability of detecting the species at a site in at least one sample a product of detecting the species in sample times the product of not detecting the DNA within a sample using the assay: $(1 - (1 - \theta)^J) \times (1 - (1 - p)^K)$.

```
sampleDetectionOne <- function(
    J = 100,
    K = 8,
    theta = 0.06,
    pDetection  = 0.3){
    return( (1 - (1 - theta) ^ J)  * (1 - (1 - pDetection) ^ K))
    }
```

Next, we explore different sample numbers, $J \in 1, 2, \ldots 100$; different assay detection probabilities, $\theta \in \{0.3, 0.35, 0.4, 0.75\}$; different sample detection probabilities, $p \in \{0.06, 0.24, 0.42, 0.75\}$; and different numbers of molecular replicates $K \in \{2, 4, 8, 16\}$.

We use the `data.table` package for storing and manipulating my data.

```
library(data.table)
results <- data.table(expand.grid(J = 1:100,
                        theta = c(0.06, 0.24, 0.42, 0.75), pDetection = c(0.3, 0.35, 0.4, 0.75),
                        K = c(2, 4, 8, 16)))
results[ , ProbDetect :=
    sampleDetectionOne(J = J, K = K, theta = theta, pDetection = pDetection)]

results[ , thetaPlot := factor(paste0("theta = ", theta))]
results[ , pDetectionPlot := factor(paste0("p = ", pDetection))]
results[ , KPlot := factor(paste0("K = ", K))]
```

Last, we plot the results using `ggplot2`.

```
library(ggplot2)
results[ , KPlot := factor(KPlot, levels = levels(results$KPlot)[c(2, 3, 4, 1)])])]

detectOne <- ggplot(data = results, aes(x = J, y = ProbDetect, color = KPlot)) +
    geom_line() +
    facet_grid( pDetectionPlot ~ thetaPlot) +
    theme_minimal() +
    ylab("Probabiltiy of detecting species at site") +
    xlab("Number of samples per site") +
  scale_color_manual("Number\nof PCR\nreplicates", values = c("wheat4", 'black', 'red', "blue"))
print(detectOne)
```
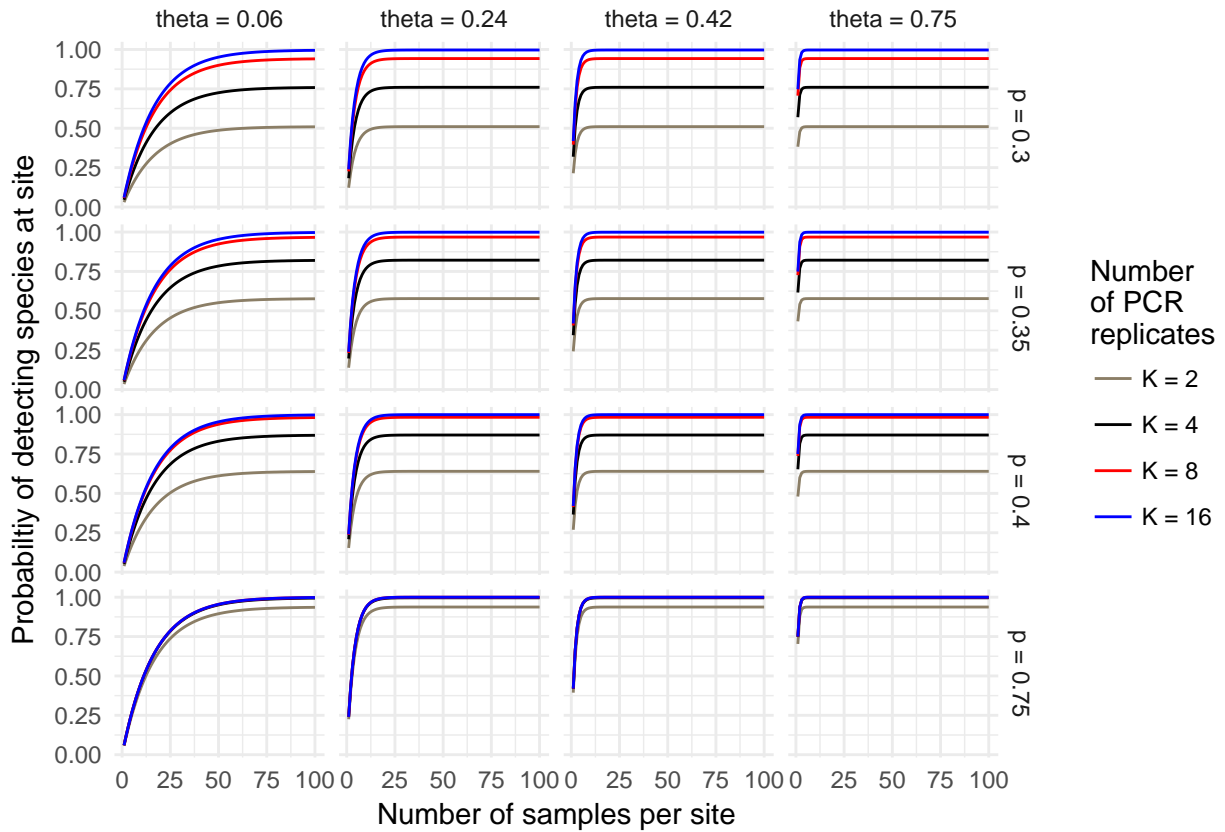

Figure 1: Probability of detecting a species in at least one sample at a site given different assay and sample detection probabilities.

```
ggsave(filename = "detectingOne.pdf", detectOne, width = 6, height = 4)
ggsave(filename = "detectingOne.jpg", detectOne, width = 6, height = 4)
```

Probability of having different observable sample occurancies

A more interesting question than simply detecting species at a site using eDNA is "Can eDNA detect different levels of sample occurance at sites?". To do this, we conduct a simulation study.

First, we draw samples J from a site with the probability θ of any sample containing DNA from a Bernoulli distribution Bernoulli(J, θ) (Note that is a special case of the binomial distribution with size = 1.). Next, we

re-sample the positive samples with K assay replicates with the probability p that an assay detects DNA from a Binomial distribution Binomail(K, p).

```r
samplesDetect <- function(
    nSims = 2,
    J = c(10, 100),
    theta = c(0.06, 0.42),
    K = 8,
    pDetection = c(0.3, 0.4)){
        results <- data.table(expand.grid(simulation = 1:nSims,
                        J = J, theta = theta,
                        K = K, pDetection = pDetection))
    for(index in 1:dim(results)[1]){
        results[ index, nPositive :=
            length(which(
                rbinom(n = length(which(rbinom( n = J, size = 1, prob = theta) >0)),
                    size = K, prob = pDetection) > 0))]
        }
    results[ , pPositive := nPositive/J]
    results[ , Samples := factor(paste0("J = ", J))]
    results[ , SamplesPlot  := factor(J)]

    results[ , thetaPlot := factor(paste0("theta = ", theta))]
  results[ , thetaPlot2 := factor( theta)]

  results[ , pDetectionPlot := factor(paste0("p = ", pDetection))]

    factorOrder <- order(as.numeric(gsub("J = ", "", levels(results$Samples))), decreasing = FALSE)
  results[ , Samples := factor(Samples, levels = levels(Samples)[factorOrder]) ]

  results[ , KPlot := paste0("K = ", K)]
  KOrder <- unique(results$KPlot)[order(as.numeric(gsub("K = ", "", unique(results$KPlot))), decreasing
  results[ , KPlot := factor( KPlot, levels = KOrder)]

  results[ , KPlot2 := factor(gsub("K = ", "", KPlot))]
  KOrder2 <- unique(results$KPlot2)[order(as.numeric(unique(results$KPlot)), decreasing = FALSE)]
  results[ , KPlot2 := factor( KPlot2, levels = KOrder2)]


  return(results)
  }
```

Next, we explore different sample numbers, $n \in \{20, 50, 100\}$; different assay detection probabilities, $p \in \{0.3, 0.35, 0.4, 0.75\}$; and different sample detection probabilities, $\theta \in \{0.06, 0.24, 0.42, 0.75\}$ by running 4,000 simulations.

```r
sampleResults <- samplesDetect(nSims = 4000,
                        theta = c(0.06, 0.24, 0.42, 0.75),
                        pDetection = c(0.3, 0.35, 0.4, 0.75),
                        J = c(20, 50, 100),
                        K = c(2, 4, 8, 16))
```

Last, we plot the results using `ggplot2`

```r
compareSites <- ggplot(sampleResults, aes(x = KPlot2, y = pPositive, fill = thetaPlot2)) +
    geom_boxplot(outlier.size = 0.5) +
```

```
      facet_grid( pDetectionPlot ~ Samples) +
    theme_minimal() +
      ylab("Propotion of positive samples") +
    xlab("PCR replicates per sample") +
    scale_fill_manual(expression(theta), values = c("orange", 'seagreen', 'red', "blue"))

print(compareSites)
```
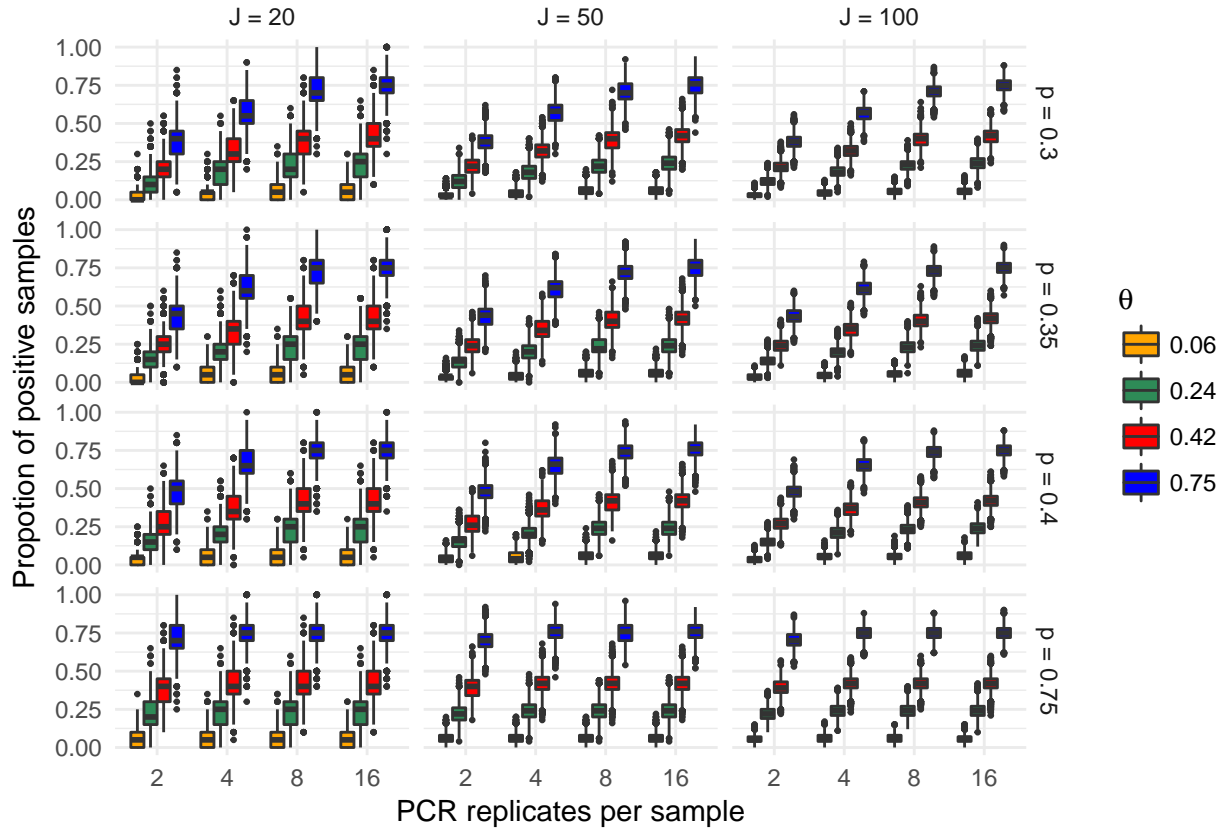

Figure 2: Proportion of samples per site (Sample occurance) that are positive based upon sample size and the assay's probability of detection.

```
ggsave(filename = "compareSites.pdf", compareSites, width = 9, height = 6)
ggsave(filename = "compareSites.jpg", compareSites, width = 9, height = 6)
```