Statistical Inference of Fish Populations from Deep-Learning Data

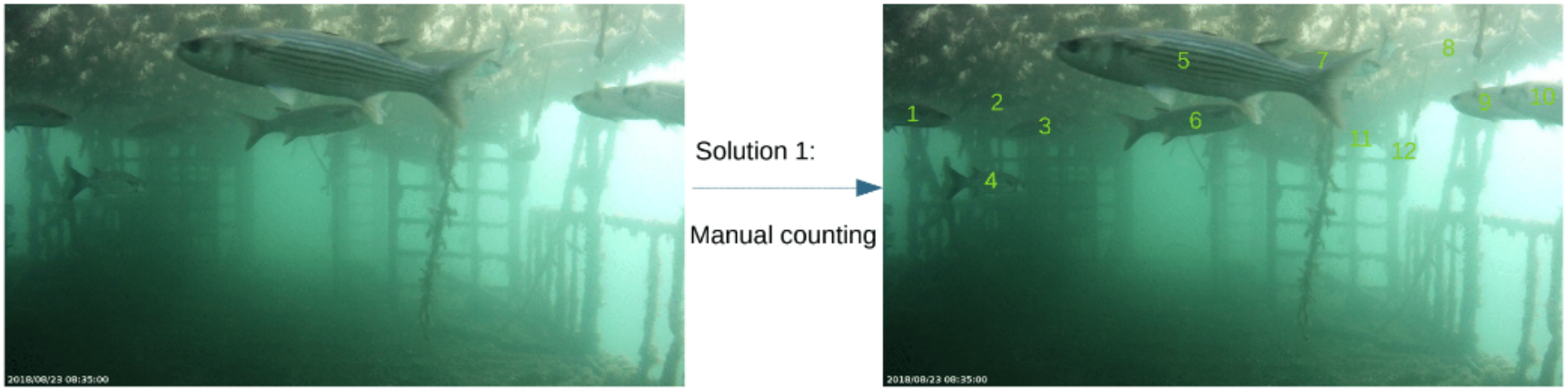
Raúl Almuzara
Pablo Femenía

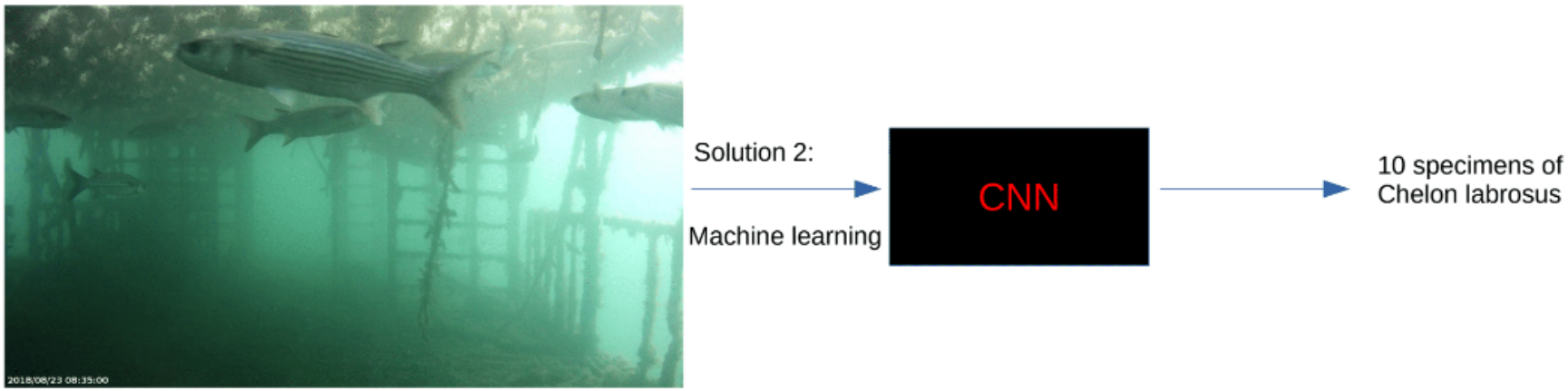
VII Iberian Modelling Week

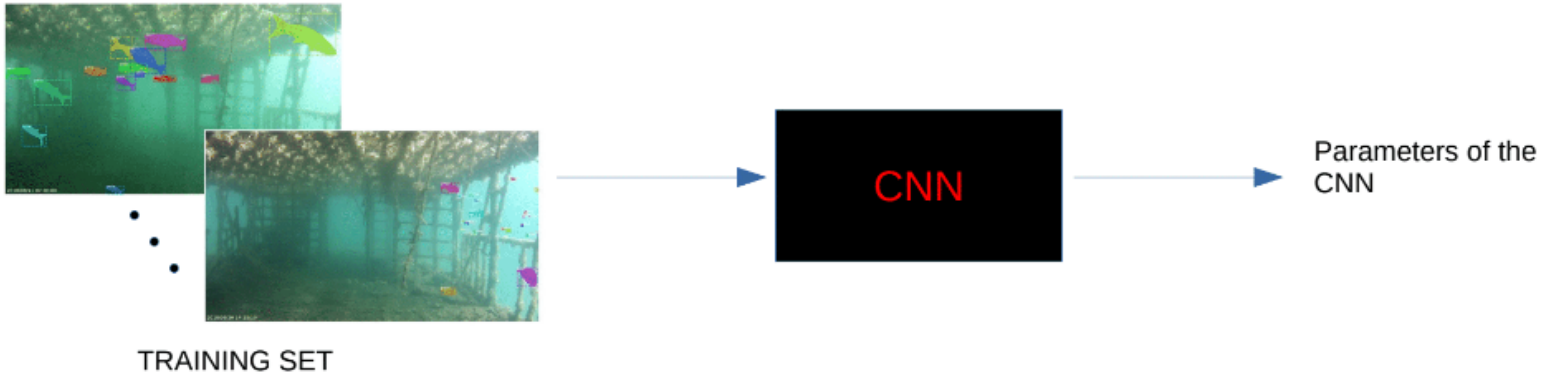
Nov 26 - Dec 1, 2021

Problem statement

- Estimation of fish populations.
- Solutions:







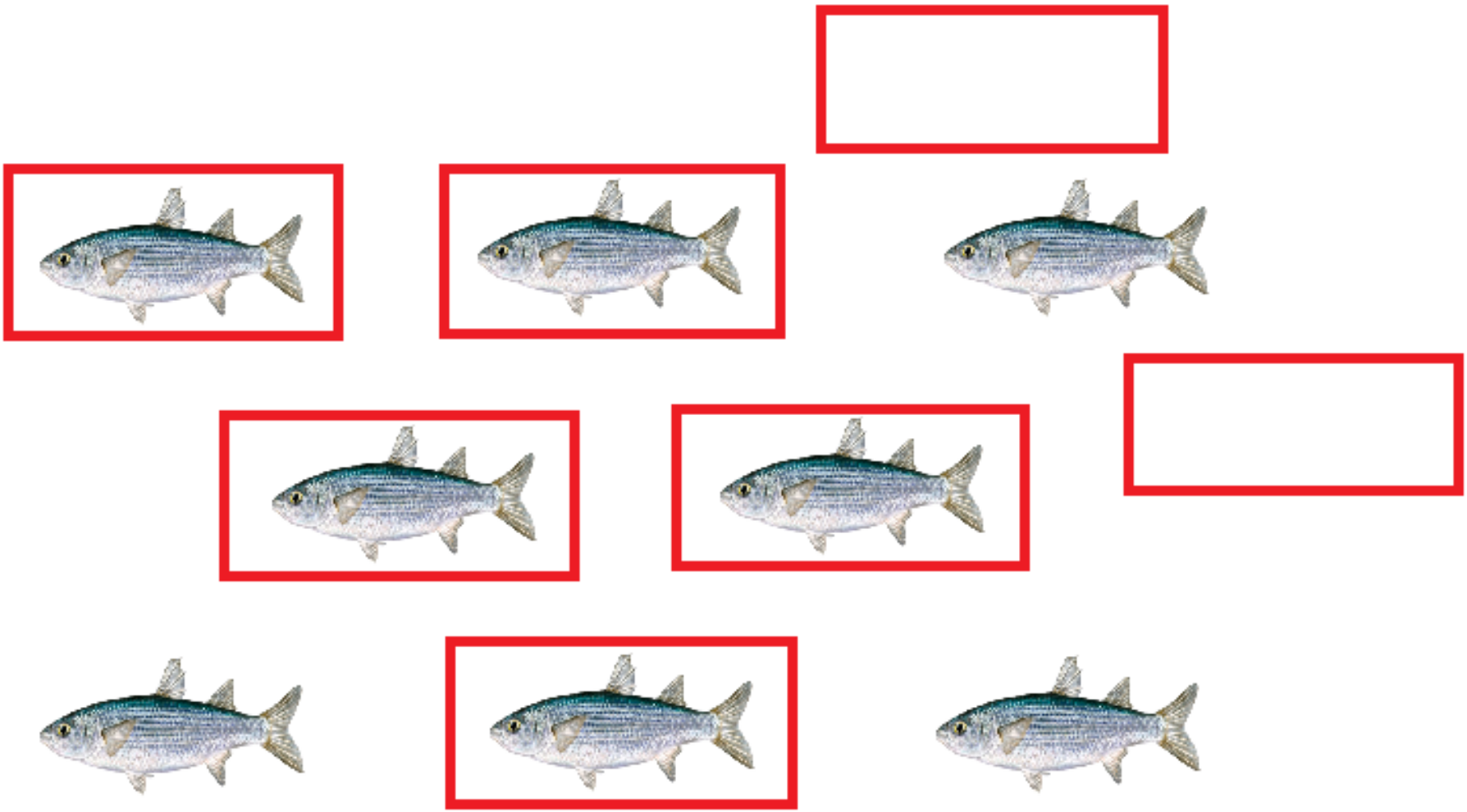
- Objective: Estimate and bound the error of the predictions.
- We will suggest models and methods to quantify prediction errors and estimate the reliability of the information given by the CNN as an output.

Variables

- TP (True positives): Number of correct detections.
- FP (False positives): Number of wrong detections.
- FN (False negatives): Number of missed detections.
- $TD = TP + FP$: Total number of detections given as an output of the CNN.
- $A = TP + FN$: Actual number of fish considered as ground truth.
- $P = TP/TD$: Precision.
- $R = TP/A$: Recall.

Example

$TP = 5,\ FP = 2,\ FN = 3,\ TD = 7,\ A = 8,\ P = 0.714,\ R = 0.625$



First approach and extension

$$\frac{P}{R} = \frac{\frac{TP}{TD}}{\frac{TP}{A}} = \frac{A}{TD} \implies A = TD \cdot \frac{P}{R}$$

Now, considering errors:

$$\delta A = TD \cdot \delta\left(\frac{P}{R}\right), \qquad f := \frac{P}{R}$$

$$\delta f = \sqrt{\left(\frac{\partial f}{\partial P} \cdot \delta P\right)^2 + \left(\frac{\partial f}{\partial R} \cdot \delta R\right)^2} = \frac{1}{R^2}\sqrt{(R \cdot \delta P)^2 + (P \cdot \delta R)^2} \implies$$

$$\implies \delta A = \frac{TD}{R^2}\sqrt{(R \cdot \delta P)^2 + (P \cdot \delta R)^2}$$

Nonparametric Inference

- The empirical distribution F_n from random variables $X_1, \ldots, X_n$ is the fraction of random variables which are smaller than x.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi_{[X_i, \infty)}(x).$$

- Using the DKW Inequality, it is possible to give a confidence band for an unknown distribution function F with the aid of the empirical distribution.

Theorem

$\mathbb{P}\left(\sup_x |F(x) - F_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$ *gives a confidence band because*
$\mathbb{P}(F_n(x) - \epsilon \leq F(x) \leq F_n(x) + \epsilon) \geq 1 - 2e^{-2n\epsilon^2}, \quad \epsilon = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$

- This can help us give a probability interval for the random variable P/R. Multiplying by the particular TD of our image, we obtain a confidence interval for A.

Bayesian Methods (I)

- A second method would be to incorporate techniques from the Bayesian Paradigm. Considering A as a random quantity, we can update our beliefs about A after knowing the number of detected fish (TD).
- The core tool is Bayes' Theorem:

$$\mathbb{P}(A = a|\, TD = td) \propto \mathbb{P}(TD = td|A = a)\mathbb{P}(A = a).$$

- The original distribution of A is called the **prior** distribution and the probability $\mathbb{P}(A = a|\, TD = td)$ gives the **posterior** distribution.
- One first approximation could use the empirical distribution to have the distribution of A across all the images and the distribution of TD conditioned to A. This is very general and should give good results.

Bayesian Methods (II)

- Further information could be obtained if we create a more specific model for the variables. We can make assumptions about the distribution they follow:
  - $TP \sim \text{Bin}(A, p)$
  - $FN \sim \text{Bin}(A, 1-p)$
  - $FP \sim \text{Po}(\lambda)$
  - $TD \sim \text{Bin}(A, p) + \text{Po}(\lambda)$
  - $A \sim \text{Po}(\mu)$
- In order to apply the bayesian methods, first, we have estimated the parameters of our random variables with the aid of the MLE estimator.
- We have also found the posterior distribution in this case:

$$\mathbb{P}(A = a | TD = td) \propto e^{-\lambda-\mu} \frac{\mu^a}{a!} \left(\sum_{j=0}^{\min(a,td)} \frac{\lambda^{td-j}}{(td-j)!} \binom{a}{j} p^j (1-p)^{a-j} \right)$$

Simulations with PyMC3

- PyMC3 is a Python package that uses Markov chain Monte Carlo methods to do Bayesian Inference. In particular, to compute the mean and Highest Density Interval of the posterior distribution. We consider the following approximations for our model:
- $A \sim Po(\mu) \rightarrow A \sim N(\mu, \mu)$
- $TD \sim Bin(A, p) + Po(\lambda) \rightarrow N(Ap + \lambda, Ap(1-p) + \lambda)$

Simulations

We made four simulations, always with $\mu = 12$:

- $td = 10, \lambda = 2, p = 0.8$
- $td = 80, \lambda = 2, p = 0.8$
- $td = 10, \lambda = 20, p = 0.1$
- $td = 80, \lambda = 20, p = 0.1$

Results (for $\mu = 12$)

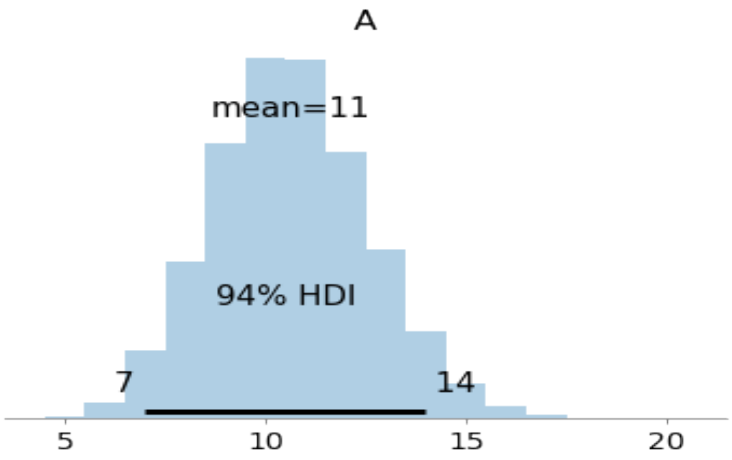


Figure: $td = 10, \lambda = 2, p = 0.8$

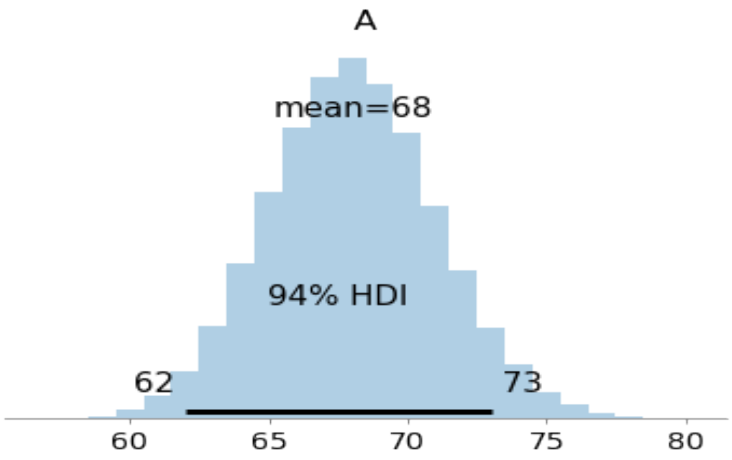


Figure: $td = 80, \lambda = 2, p = 0.8$

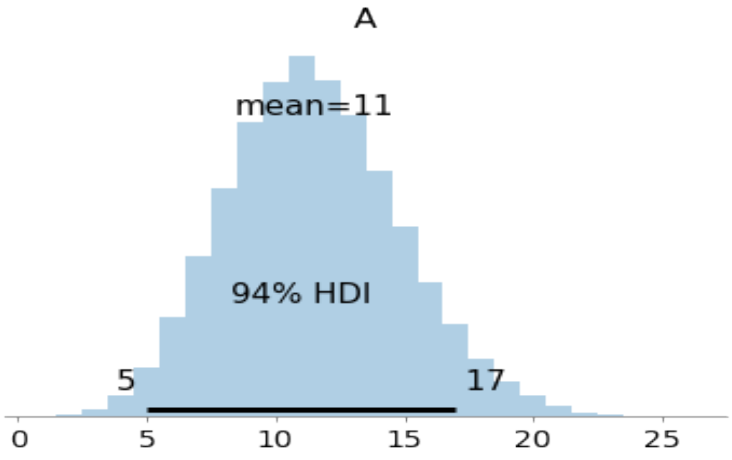


Figure: $td = 10, \lambda = 20, p = 0.1$

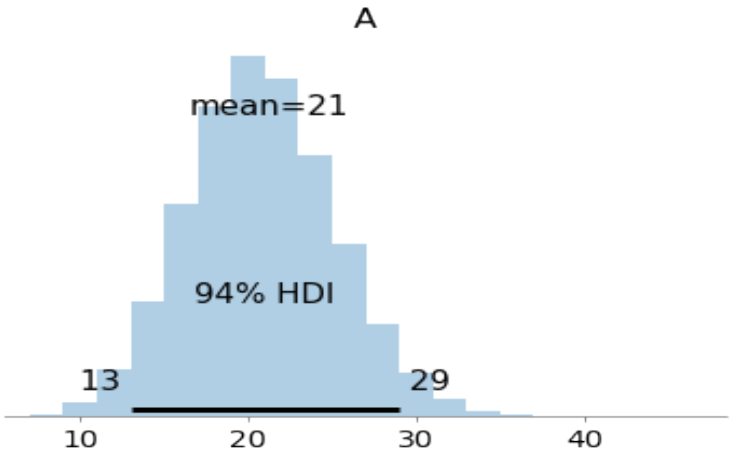


Figure: $td = 80, \lambda = 20, p = 0.1$

Conclusions from the simulations

When the network behaves "badly" according to the validation set, our model ignores the result of the network more than if it behaves "well".

- Bayesian methods provide a good framework to correct the results from a Neural Network.
- The Bayesian Framework proves useful as we can update our previous beliefs about the actual number of fish and give a reasonable HDI.
- The estimation of the performance, either done by nonparametric or parametric methods, is crucial to quantify the strength of the correction.

Further work

- Test the model with actual data from the *Deep-Ecomar* project.
- Compare the performance between the nonparametric and parametric inference methods.
- Take into account the effects of labeling errors.

References


Abdulla, W. (2017). *Mask R-CNN for Object Detection and Segmentation*. https://github.com/matterport/Mask_RCNN


Casella, G., Berger, R. (2002). *Statistical Inference*. Duxbury.


Davidson-Pilon, C. (2013). *Probabilistic Programming & Bayesian Methods for Hackers*. https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/


Itzikovitch, R. (2019). *Are We Confident Our Model's Recall is Precise?*. https://towardsdatascience.com/are-we-confident-our-models-recall-is-precise-133112a6c407


Khan, S. S. (2020). *An Introduction to Classification Using Mislabeled Data*. https://towardsdatascience.com/an-introduction-to-classification-using-mislabeled-data-581a6c09f9f5


VII Iberian Modelling Week website. https://www.spm.pt/PT-MATHS-IN/7imw/english/index.html