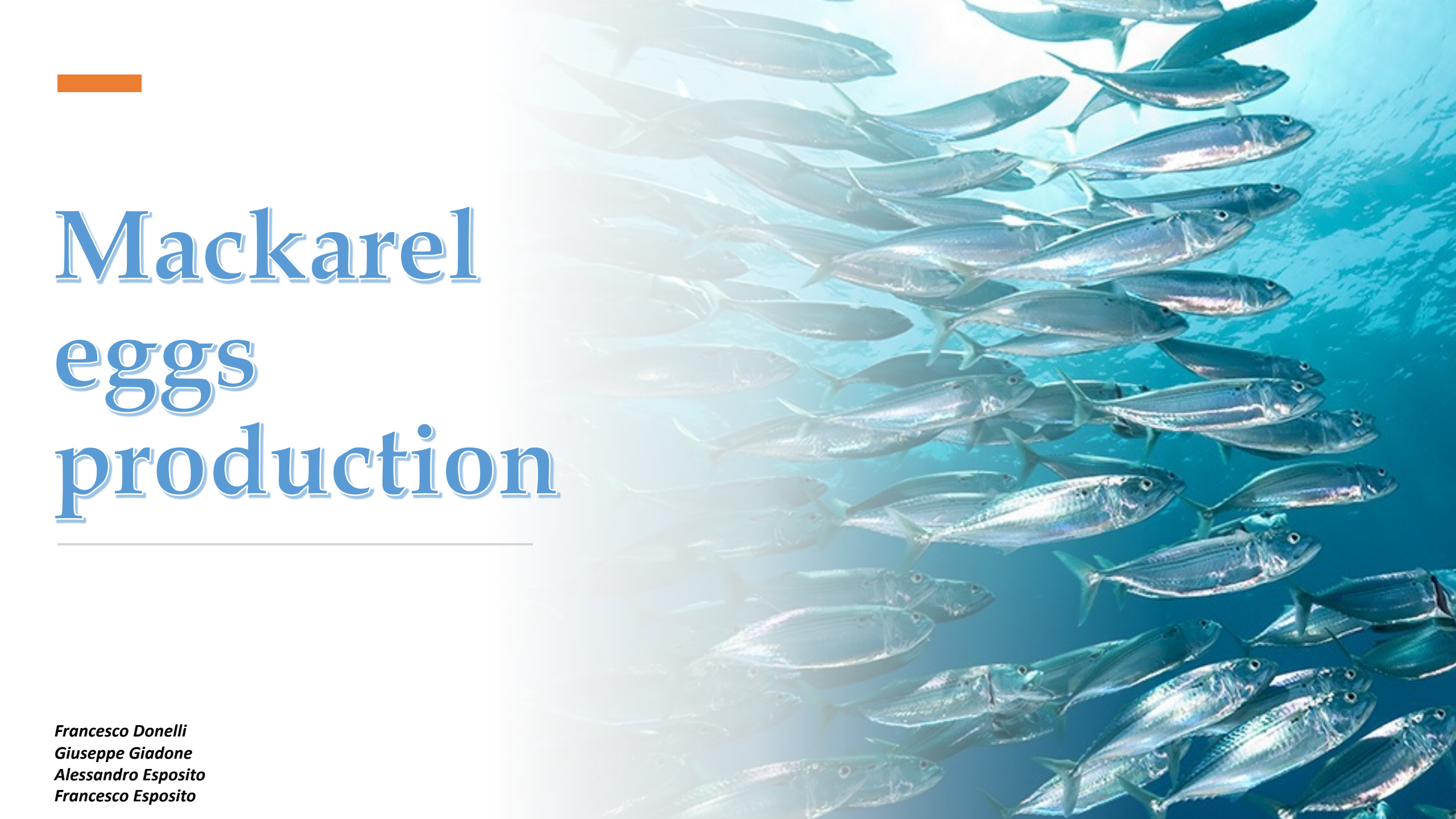
Mackarel eggs production



Francesco Donelli
Giuseppe Giadone
Alessandro Esposito
Francesco Esposito

Outline of the presentation

- Our dataset and goal of the analysis
- Nature of the data and Poisson model
- Zeros inflation and ZIP model
- Model selection and predictive power
- Final considerations

Goal of the analysis and *content* of the dataset

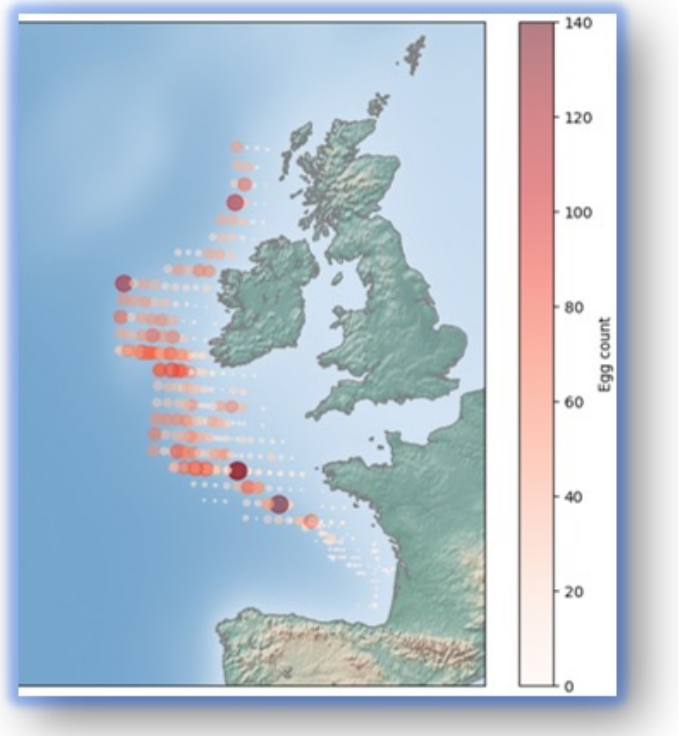
- *Goal:* finding a good model with respect to both ***predictive power*** and ***model validation criteria.***

- *Content:* records about the number of eggs fished through the coasts of England, France and Ireland.

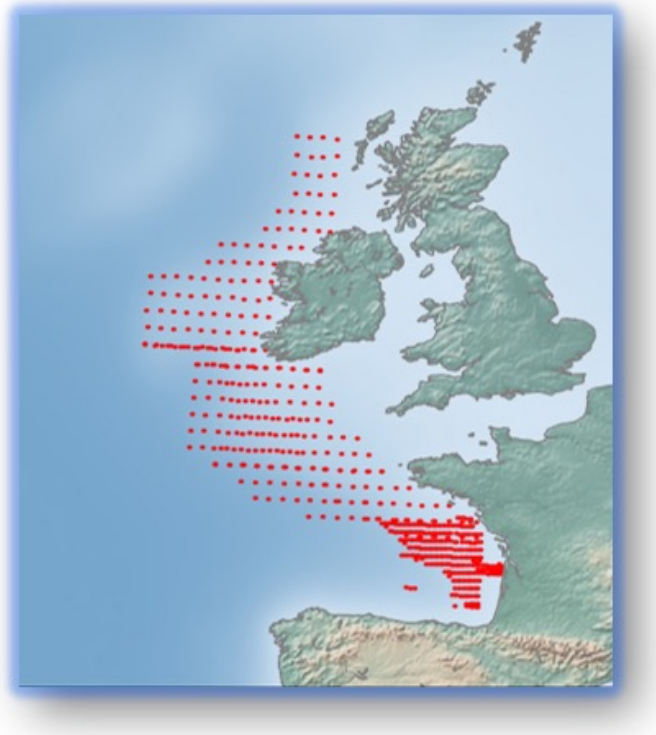
- *Variables:*
  - **Egg count:** number of eggs fished [*response variable*];
  - **Time:** recording time;
  - **Flow:** power of sea currents;
  - **C.dist:** distance from sample to 200m contour (degrees);
  - **Salinity;**
  - **S.depth:** sampling start depth;
  - **Temp.surf:** sea temperature;
  - **Net area:** sampling net area.

Why remove *latitude* and *longitude*?

- ***Density of the quantity fished***



- ***Where the mackerel were fished***



❖COLLINEARITY TEST OF THE VARIABLES

	time	salinity	flow	s.depth	temp.surf	net.area	c.dist
time	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
salinity	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
flow	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
s.depth	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
temp.surf	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
net.area	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
c.dist	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

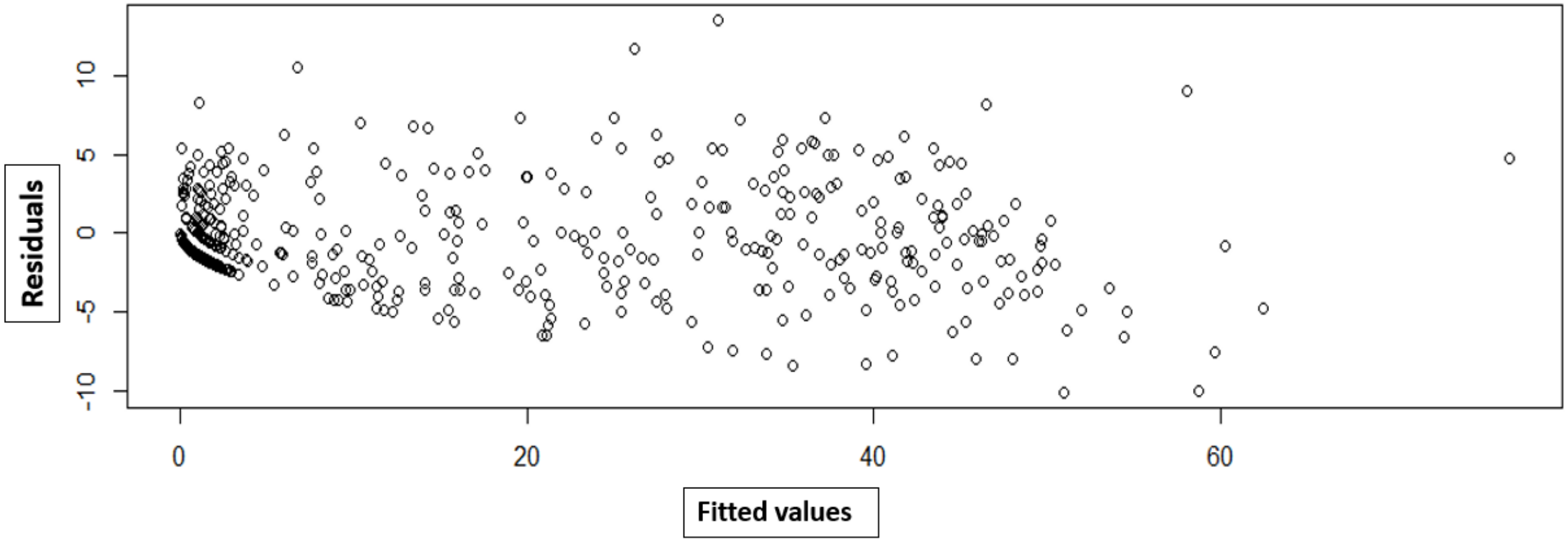
Poisson model

$$p(\lambda, x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

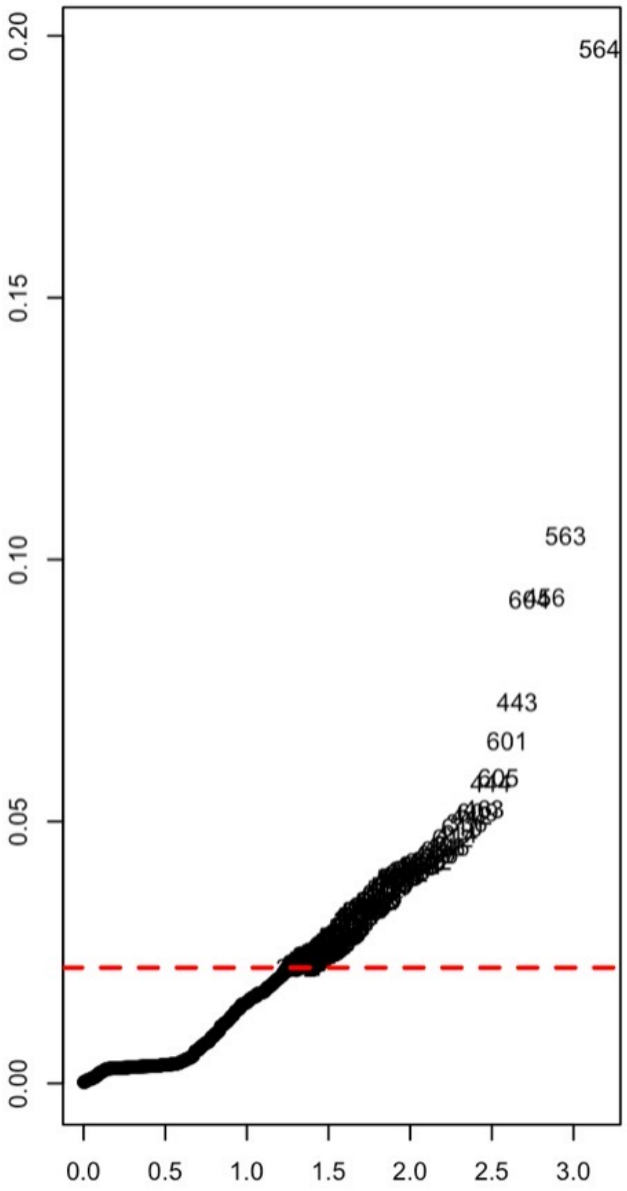
❖SUMMARY OF THE POISSON MODEL

```
Call:
glm(formula = egg.count ~ time + salinity + log(flow) + log(s.depth) +
    temp.surf + c.dist + I(net.area^2), family = poisson, data = mack)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.0970   -1.7710   -1.2092    0.4062   13.5378

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   19.36110    4.69747   4.122 3.76e-05 ***
time           0.28620    0.04582   6.246 4.22e-10 ***
salinity      -0.45288    0.13447  -3.368 0.000758 ***
log(flow)      0.19742    0.01093  18.055  < 2e-16 ***
log(s.depth)   1.98434    0.06277  31.614  < 2e-16 ***
temp.surf     -0.51171    0.01406 -36.403  < 2e-16 ***
c.dist        -0.11533    0.02366  -4.874 1.09e-06 ***
I(net.area^2) 10.97609    1.28619   8.534  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18422.7  on 633  degrees of freedom
Residual deviance:  5476.3  on 626  degrees of freedom
AIC: 7038.1

Number of Fisher Scoring iterations: 6
```

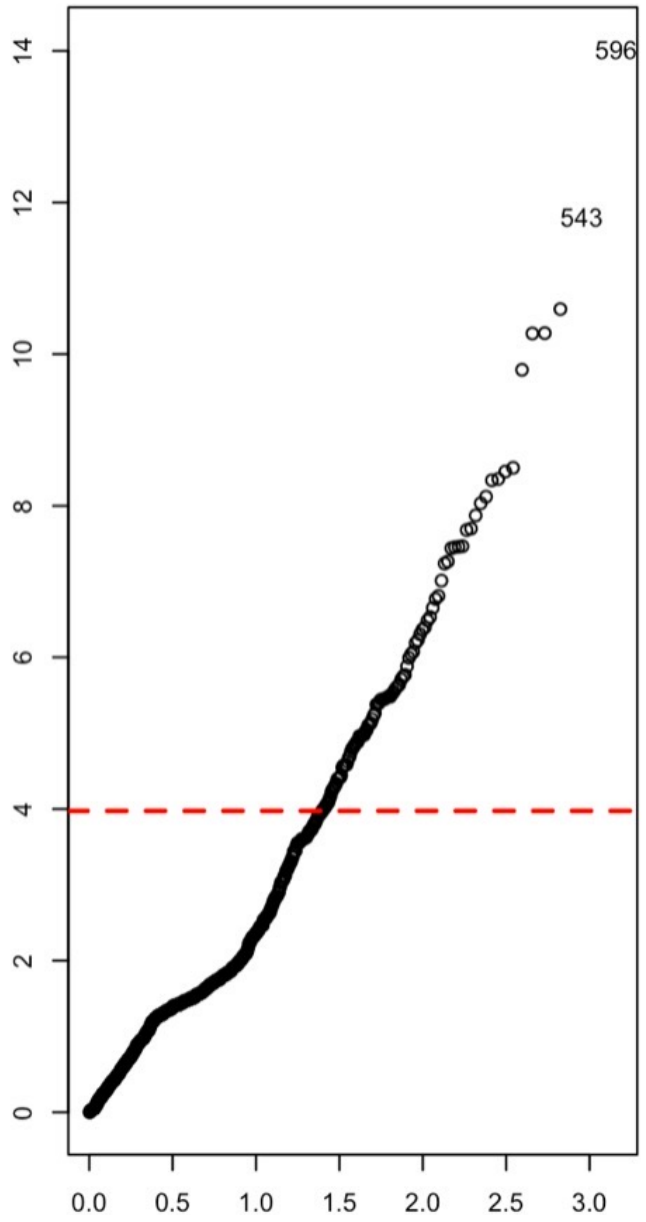
❖RESIDUALS vs FITTED



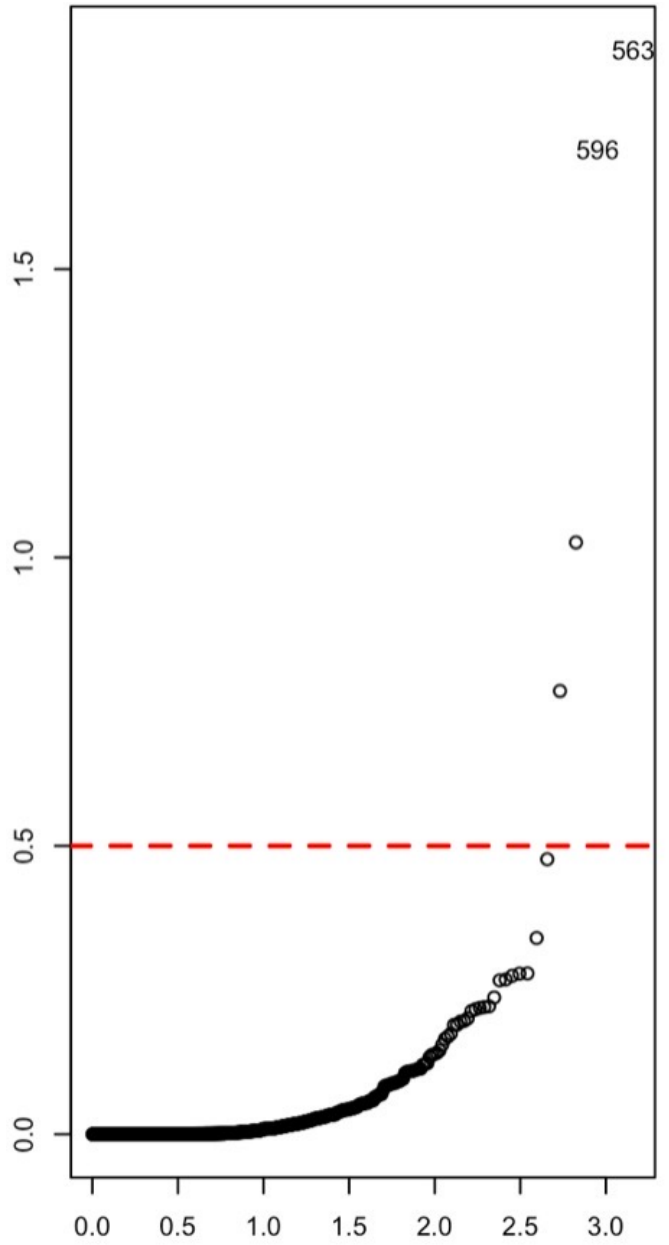
- Check for **Leverages**



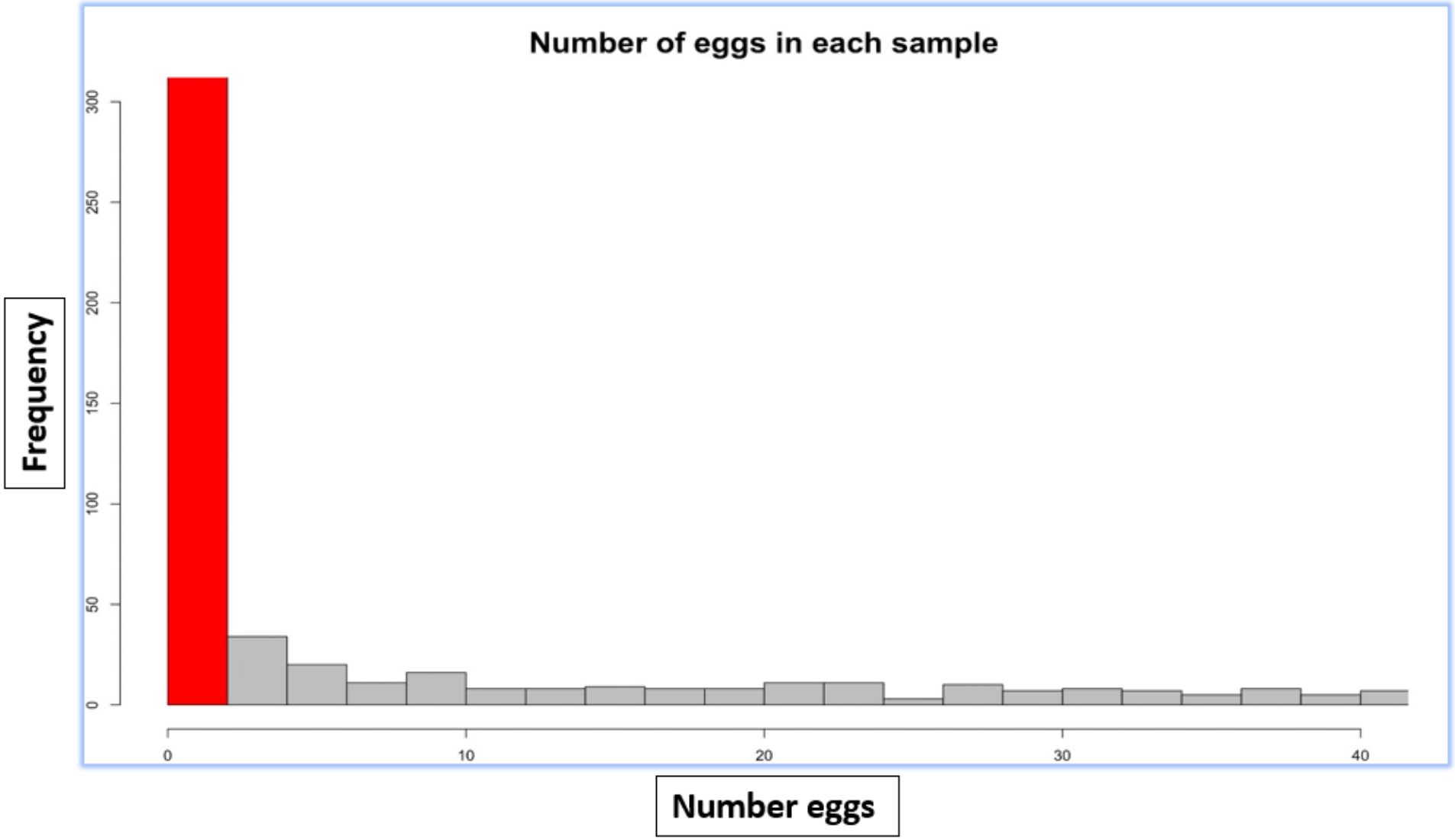
- Check for **Outliers**



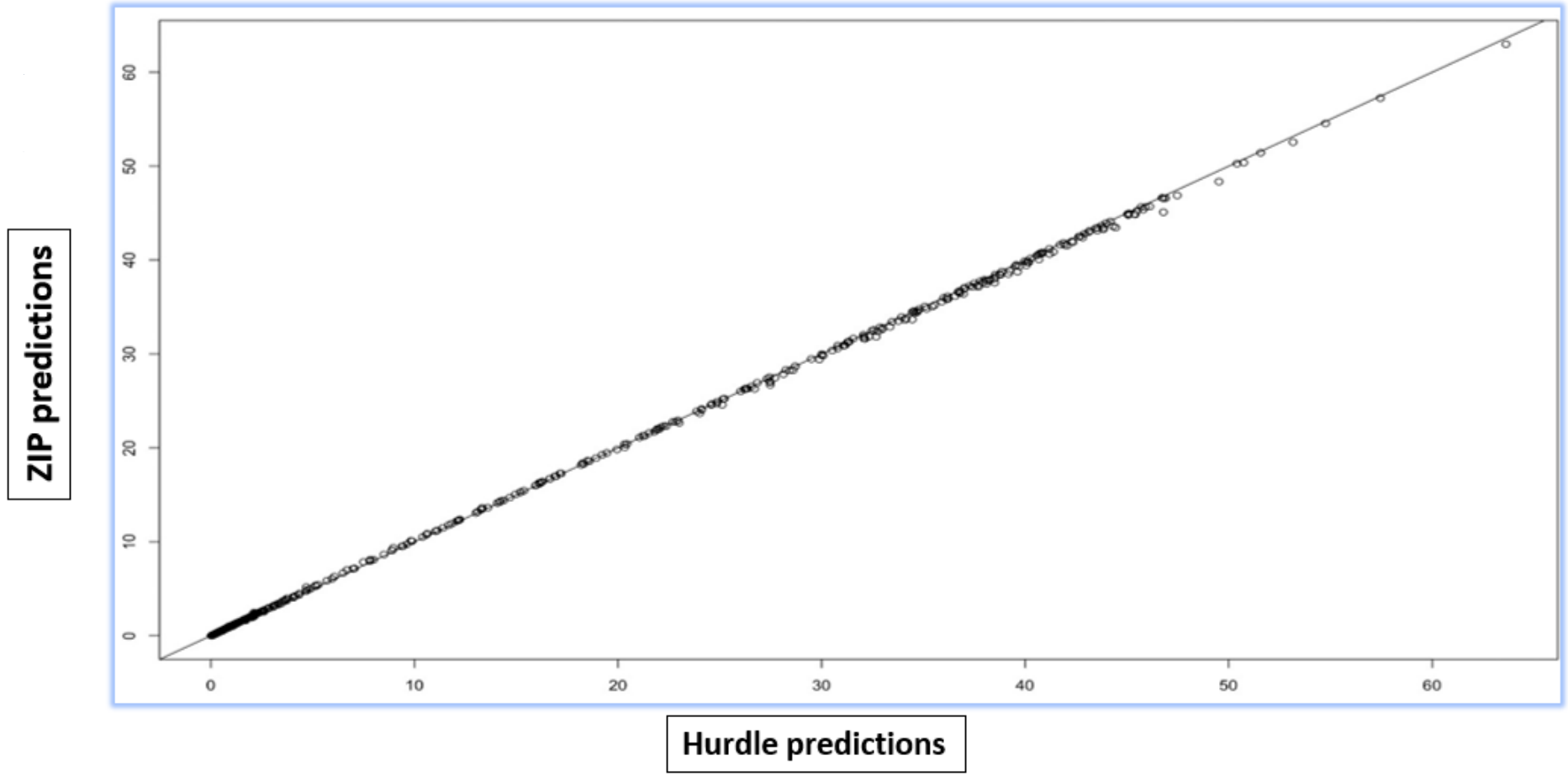
- Check for **Influential Points**



❖HOW MANY ZERO?



Hurdle or *ZIP*?



ZIP Model

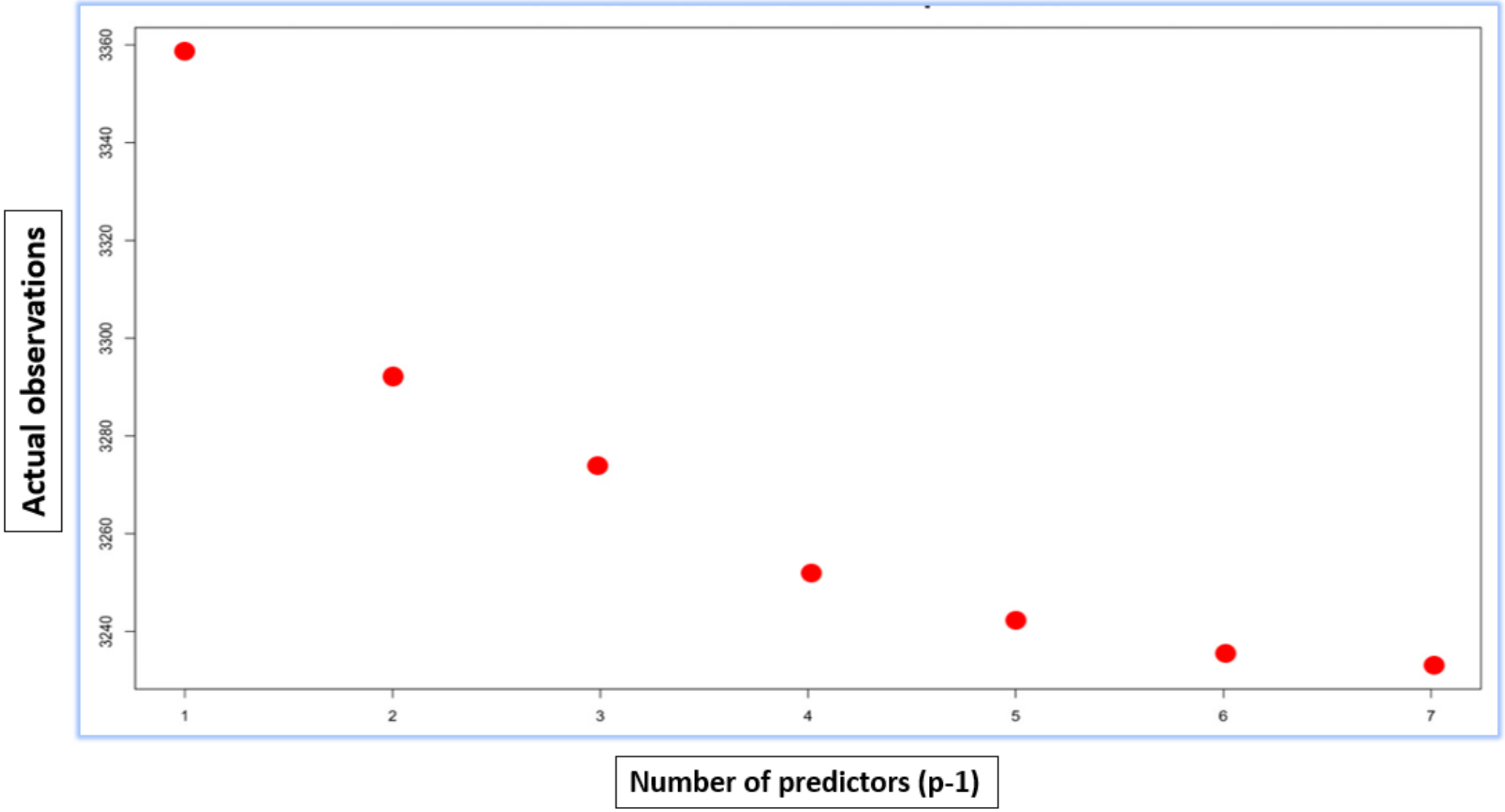
$$\Pr(Y = 0) = \pi + (1 - \pi)e^{-\lambda}$$

$$\Pr(Y = y_i) = (1 - \pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}, \qquad y_i = 1, 2, 3, \ldots$$

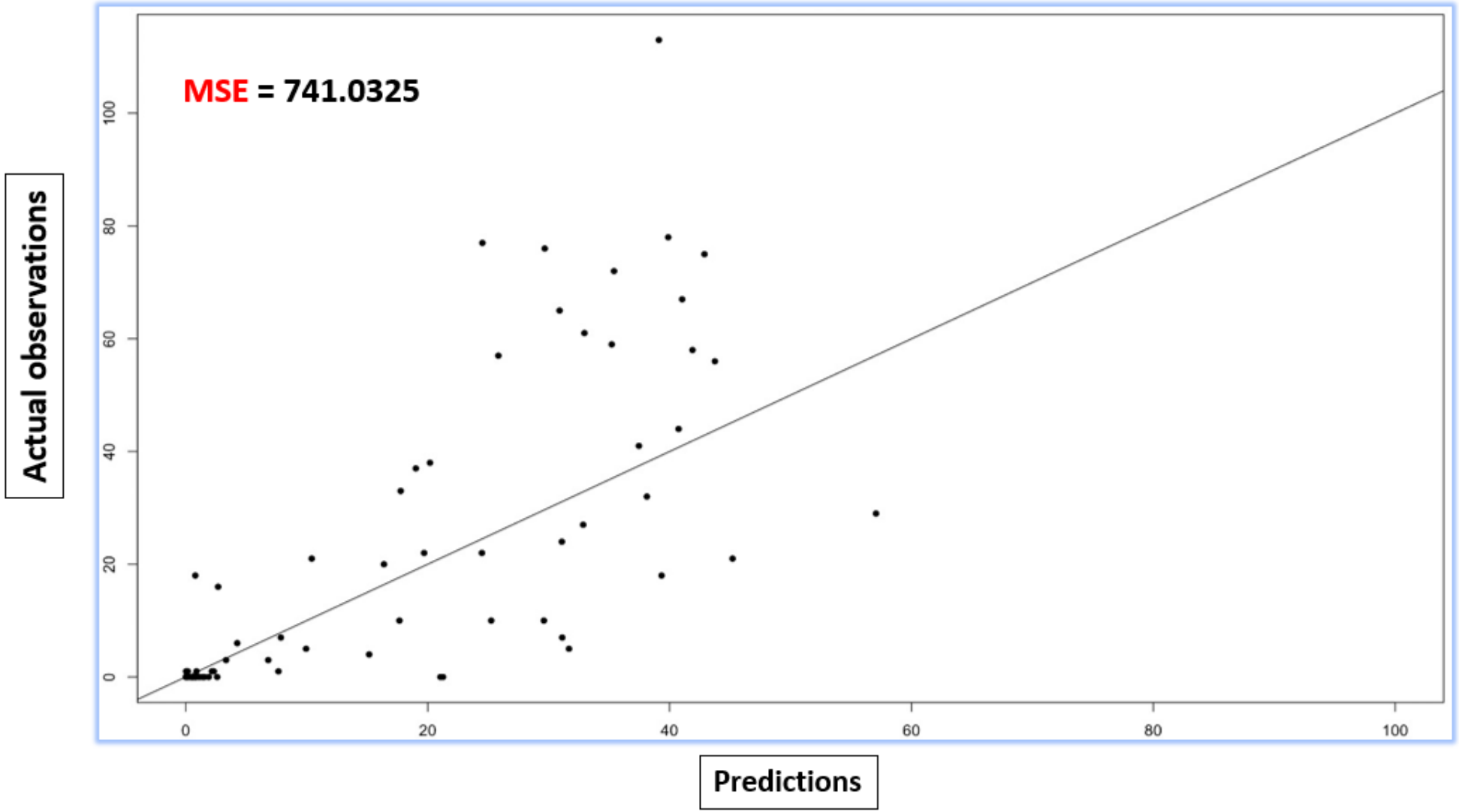
❖SUMMARY OF THE ZIP MODEL

```
Call:
zeroinfl(formula = egg.count ~ temp.surf + time + salinity + I(net.area^2) + log(s.depth) + log(flow) +
    c.dist, data = mack)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-5.5633 -0.7960 -0.4369  0.3302 33.0141

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   20.77889    4.91200   4.230 2.33e-05 ***
temp.surf     -0.33609    0.01555 -21.615  < 2e-16 ***
time           0.26832    0.04554   5.892 3.81e-09 ***
salinity      -0.65206    0.14170  -4.602 4.19e-06 ***
I(net.area^2) 13.29073    1.31689  10.093  < 2e-16 ***
log(s.depth)   1.76735    0.06266  28.204  < 2e-16 ***
log(flow)      0.15549    0.01144  13.593  < 2e-16 ***
c.dist        -0.05893    0.02412  -2.443   0.0146 *

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -20.9981    22.8970  -0.917 0.359108
temp.surf       0.4003     0.1083   3.694 0.000221 ***
time            0.1546     0.4962   0.312 0.755309
salinity        0.2902     0.6481   0.448 0.654289
I(net.area^2) 134.9270    37.1928   3.628 0.000286 ***
log(s.depth)   -0.9003     0.5004  -1.799 0.071969 .
log(flow)       0.2015     0.1747   1.153 0.248865
c.dist          1.5487     0.3531   4.387 1.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 34
Log-likelihood: -3075 on 16 Df
```

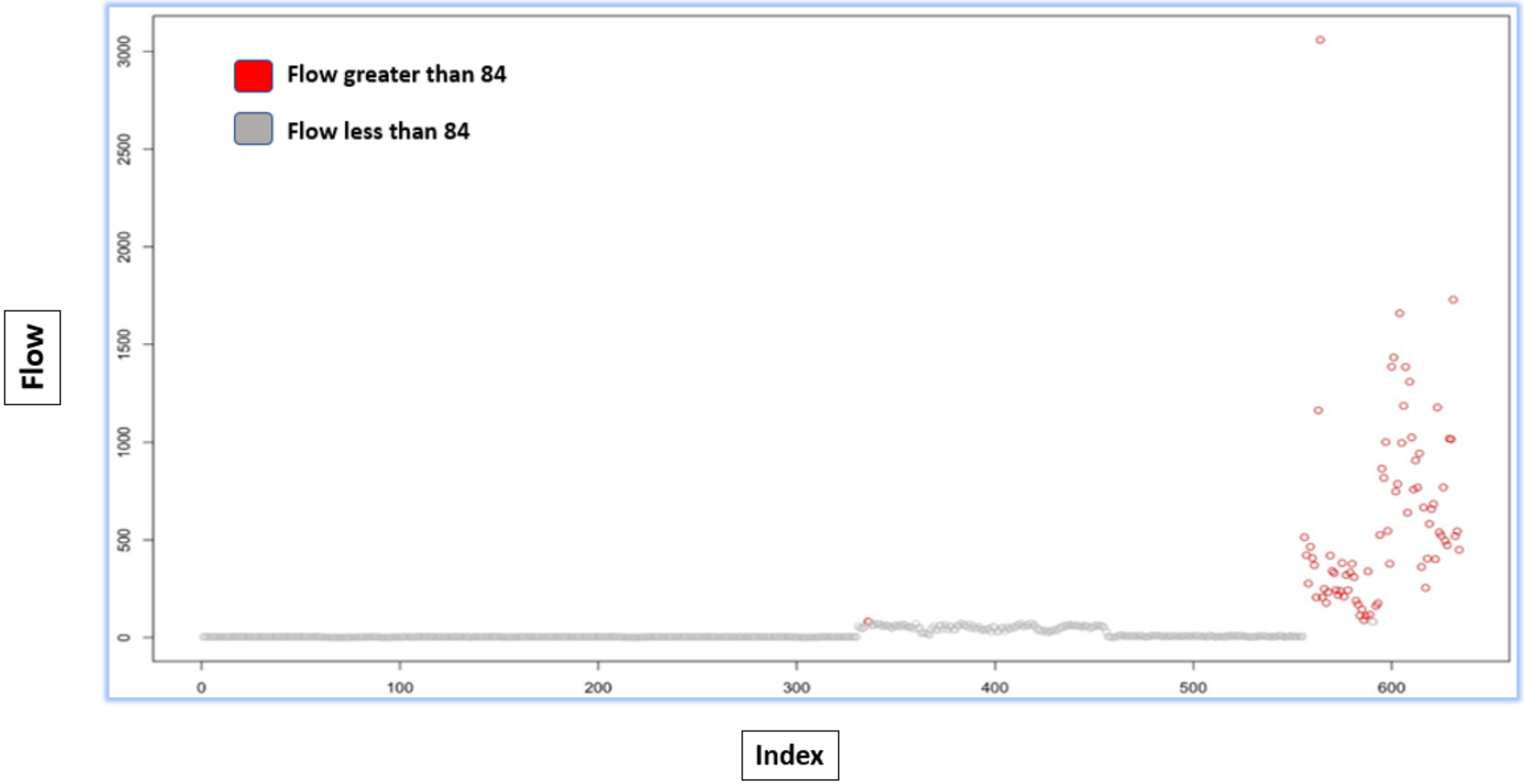
❖ VALUE OF AIC WITH ALL VALUES OF FLOW



❖VALUE OF MSE WITH ALL VALUES OF FLOW



❖EFFECTS OF FLOW

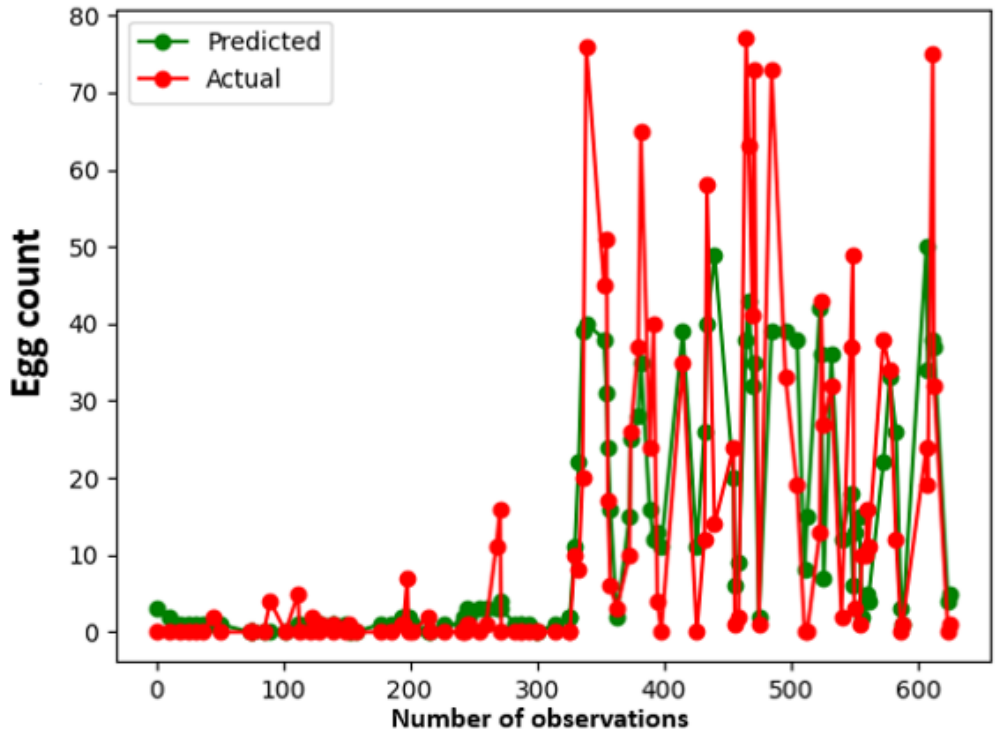


❖NEW SUMMARY OF THE MODEL

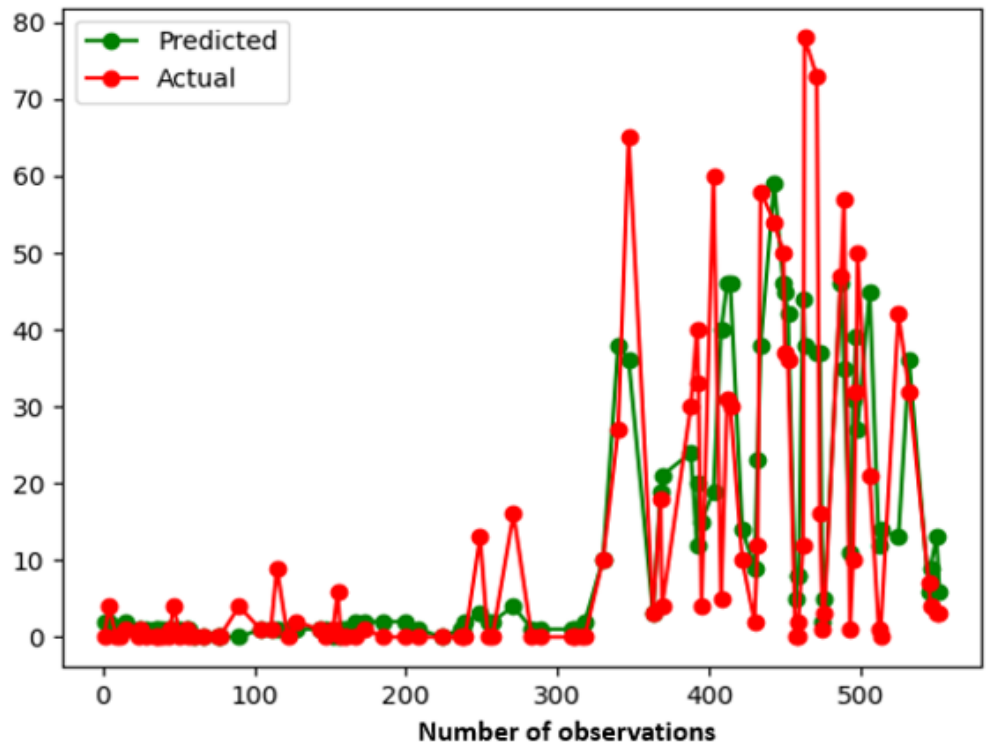
```
Call:
zeroinfl(formula = egg.count ~ temp.surf + time + salinity + I(net.area^2) + log(s.depth) + log(flow) +
    c.dist, data = mack)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-5.30042 -0.73956 -0.42862  0.07985 27.55742

Count model coefficients (poisson with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -17.94253    6.52553  -2.750 0.005967 **
temp.surf      -0.23943    0.01795 -13.340  < 2e-16 ***
time            0.17394    0.05098   3.412 0.000645 ***
salinity        0.35152    0.18634   1.886 0.059231 .
I(net.area^2)   5.74265    5.27038   1.090 0.275886
log(s.depth)    2.23343    0.09942  22.464  < 2e-16 ***
log(flow)       0.03342    0.07335   0.456 0.648610
c.dist          0.18396    0.02683   6.856 7.08e-12 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -47.4773    26.0651  -1.821  0.06853 .
temp.surf       0.3798     0.1359   2.795  0.00519 **
time            0.2608     0.5395   0.483  0.62883
salinity        0.9562     0.7302   1.310  0.19033
I(net.area^2) 226.7024    81.0185   2.798  0.00514 **
log(s.depth)   -1.7950     1.2044  -1.490  0.13613
log(flow)       1.5465     1.1295   1.369  0.17097
c.dist          1.8432     0.4098   4.498 6.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 34
Log-likelihood: -2221 on 16 Df
```

Final conclusions (ZIP model)

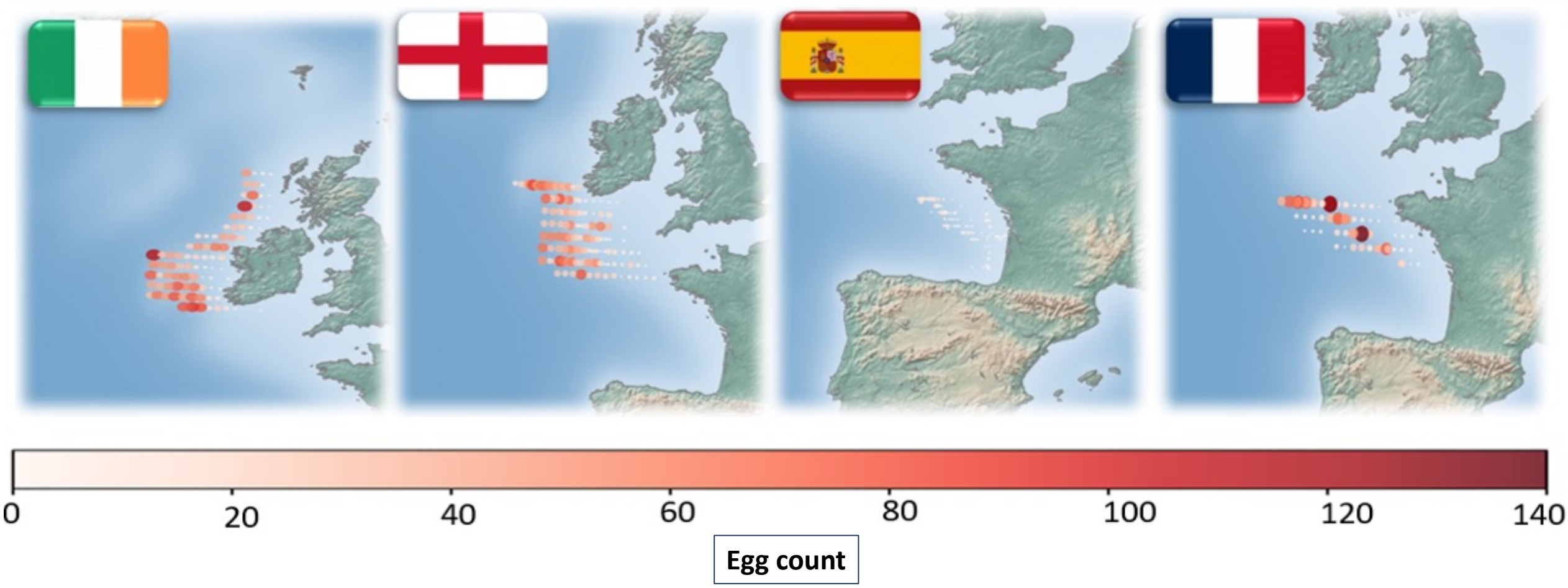
❖ High flow values included



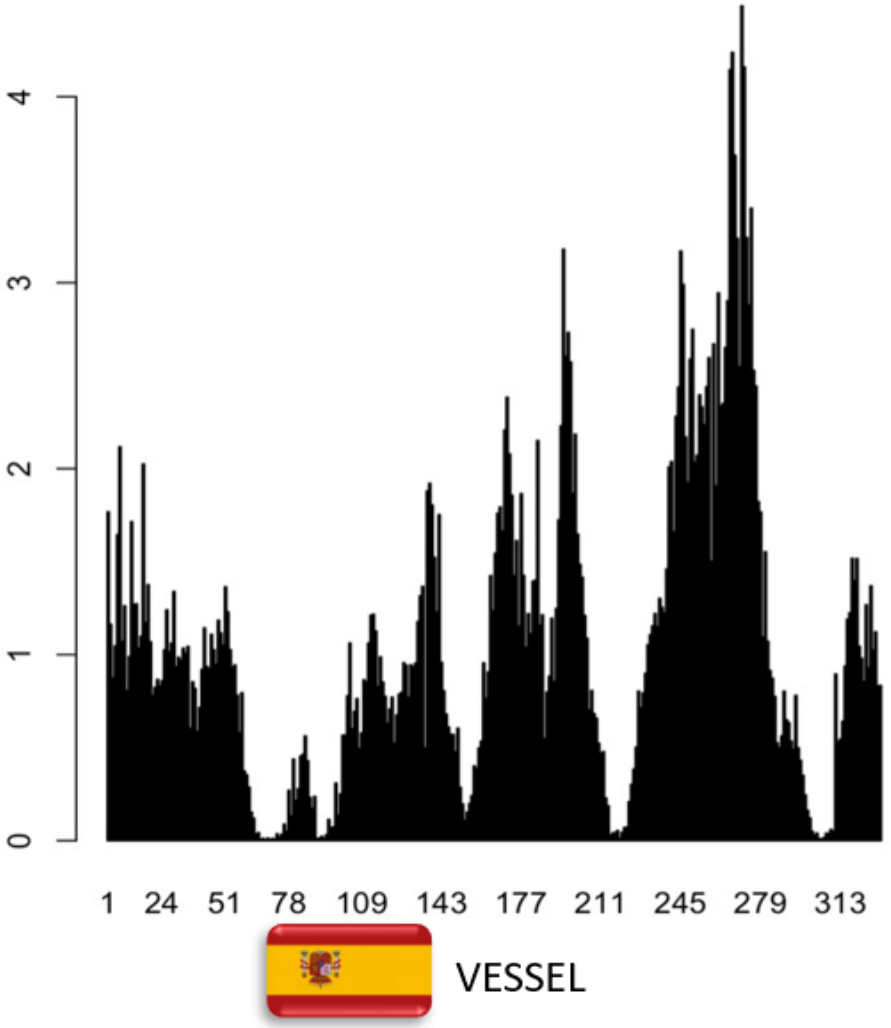
❖ High flow values excluded

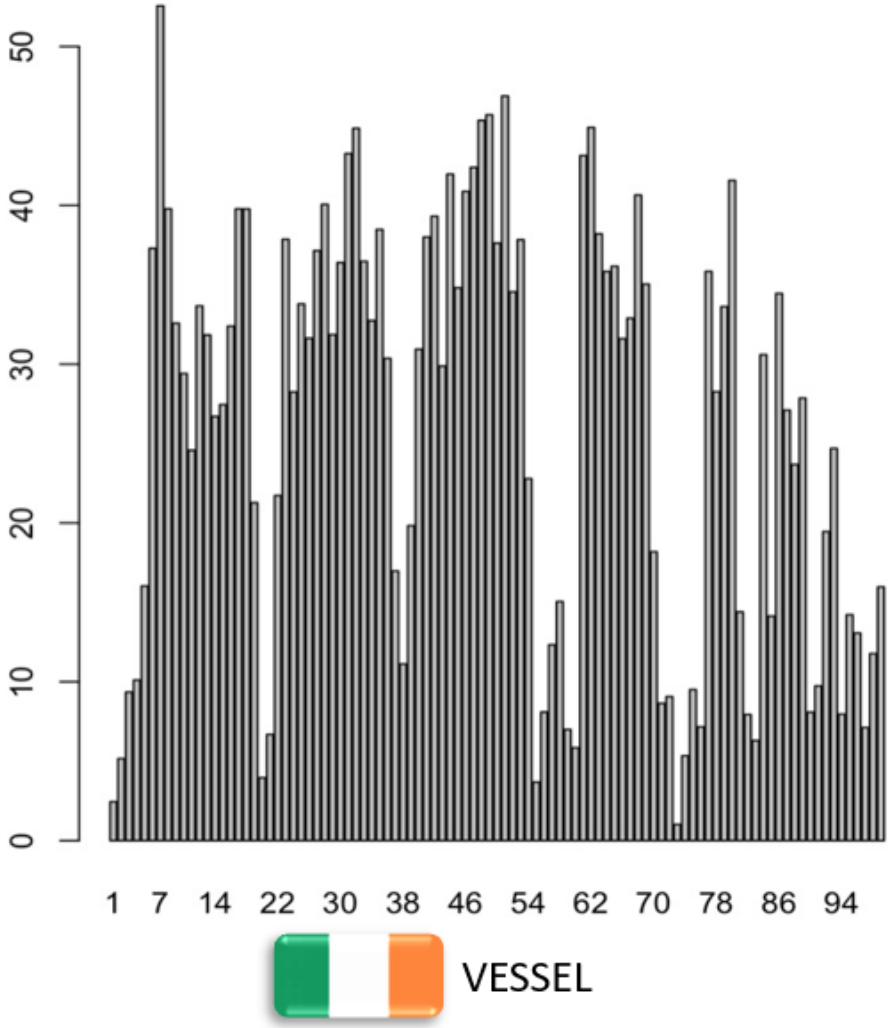


❖WHO NEEDS OUR MODEL?



❖EGG COUNT PREDICTED FOR IRLAND AND SPAIN





THANKS FOR ATTENTION

❖APPENDIX

Vuong test (zip is model 1)

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -4.958636 model2 > model1 3.5495e-07
AIC-corrected        -4.867776 model2 > model1 5.6431e-07
BIC-corrected        -4.665517 model2 > model1 1.5392e-06
```

```
> anova(modp, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: egg.count

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          633    18422.7
time          1      0.1       632    18422.6  0.785743
salinity      1    201.3       631    18221.3 < 2.2e-16 ***
log(flow)     1   3504.9       630    14716.4 < 2.2e-16 ***
log(s.depth)  1   7764.8       629     6951.5 < 2.2e-16 ***
I(net.area^2) 1     18.3       628     6933.2 1.863e-05 ***
c.dist        1      9.1       627     6924.1  0.002599 **
temp.surf     1   1456.8       626     5467.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> drop1(modp, test="Chisq")
Single term deletions

Model:
egg.count ~ time + salinity + log(flow) + log(s.depth) + I(net.area^2) +
    c.dist + temp.surf
             Df Deviance    AIC     LRT  Pr(>Chi)
<none>            5467.3 7029.2
time          1   5515.4 7075.2   48.05 4.159e-12 ***
salinity      1   5479.2 7039.0   11.86 0.0005747 ***
log(flow)     1   5766.2 7326.0  298.89 < 2.2e-16 ***
log(s.depth)  1   6914.1 8473.9 1446.78 < 2.2e-16 ***
I(net.area^2) 1   5539.3 7099.1   71.93 < 2.2e-16 ***
c.dist        1   5493.0 7052.8   25.65 4.083e-07 ***
temp.surf     1   6924.1 8484.0 1456.81 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 0.1 ' ' 1
```

Ocount: actual zeros
Pcount: predicted zeros

```
          0           1           2           3           4           5           6
136.31722311 104.98060584  65.19307336  34.58671537  17.87726135  10.30645260   7.27048798
          7           8           9          10          11          12          13
  6.19078204   5.85926287   5.76171733   5.70294073   5.62408326   5.51816043   5.39395388
         14          15          16          17          18          19          20
  5.26274456   5.13465853   5.01814134   4.91986724   4.84442226   4.79398923   4.76841661
         21          22          23          24          25          26          27
  4.76582155   4.78356471   4.81924498   4.87136592   4.93947950   5.02381691   5.12458230
         28          29          30          31          32          33          34
  5.24116262   5.37149268   5.51173617   5.65633711   5.79839796   5.93027295   6.04423814
         35          36          37          38          39          40          41
  6.13311161   6.19073534   6.21228014   6.19438061   6.13513815   6.03404256   5.89185809
         42          43          44          45          46          47          48
  5.71050421   5.49294264   5.24306592   4.96557385   4.66582348   4.34964355   4.02311436
         49          50          51          52          53          54          55
  3.69232323   3.36311399   3.04085191   2.73022540   2.43510091   2.15844079   1.90228587
         56          57          58          59          60          61          62
  1.66779742   1.45534720   1.26464093   1.09485948   0.94480288   0.81302500   0.69795000
         63          64          65          66          67          68          69
  0.59796573   0.51149250   0.43702876   0.37317692   0.31865397   0.27229156   0.23302998
         70          71          72          73          74          75
  0.19990955   0.17206197   0.14870311   0.12912786   0.11270680   0.09888422
> ocount

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23
265  63  27  19  15  11   9   5   6   4  12   3   5   4   4   4   5   3   5   4   4   5   6   5
 24  26  27  28  29  30  31  32  33  34  35  36  37  38  40  41  42  43  44  45  46  47  49  50
  6   3   6   4   3   4   3   5   4   3   4   1   6   2   5   4   3   5   2   2   4   2   2   4
 51  52  53  54  55  56  57  58  59  60  61  62  63  65  66  67  69  70  72  73  75  76  77  78
  3   2   5   1   1   2   4   2   2   1   1   1   1   2   2   1   1   1   2   2   2   2   2   2
 81  84  87  90
  1   1   1   1
>
```

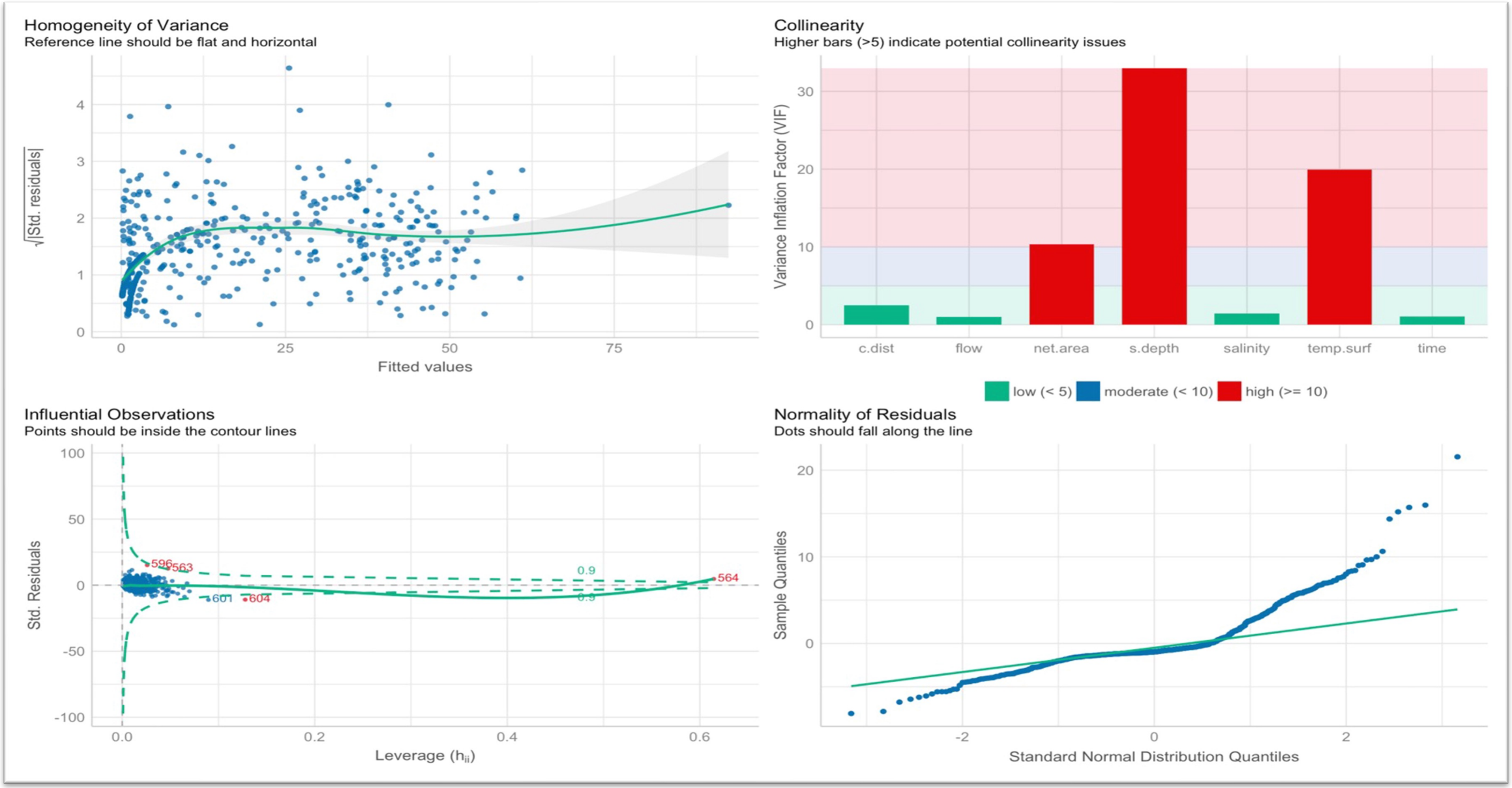
```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap
import pandas as pd
coordinates1 = pd.read_csv("/Users/alessandroesposito/cordinates_sp")
egg = pd.read_csv("/Users/alessandroesposito/egg_sp")


egg1= egg["x"].values
lat=coordinates1["lat"].values
lon=coordinates1["lon"].values
margin = 3.5
lat_min = min(lat) - margin
lat_max = max(lat) + margin
long_min = min(lon) - margin
long_max = max(lon) + margin

# 1. Draw the map background
fig = plt.figure(figsize=(8, 8))
m = Basemap(llcrnrlon=long_min,
            llcrnrlat=lat_min,
            urcrnrlon=long_max,
            urcrnrlat=lat_max,
            lat_0=(lat_max - lat_min)/2,
            lon_0=(long_max-long_min)/2,
            projection='lcc',
            resolution = 'h',
            #area_thresh=10000.,
            )
m.shadedrelief()
m.drawcoastlines(color='gray')
m.drawcountries(color='gray')
m.drawstates(color='gray')

# 2. scatter egg data, with color and size reflecting area
m.scatter(lon, lat, latlon=True, c=egg1, s=egg1,

          cmap='Reds', alpha=0.8)

# 3. create colorbar and legend
plt.colorbar(label=r'Egg count')
plt.clim(0,140)

plt.show()
```

- **Python code used to implement the maps**

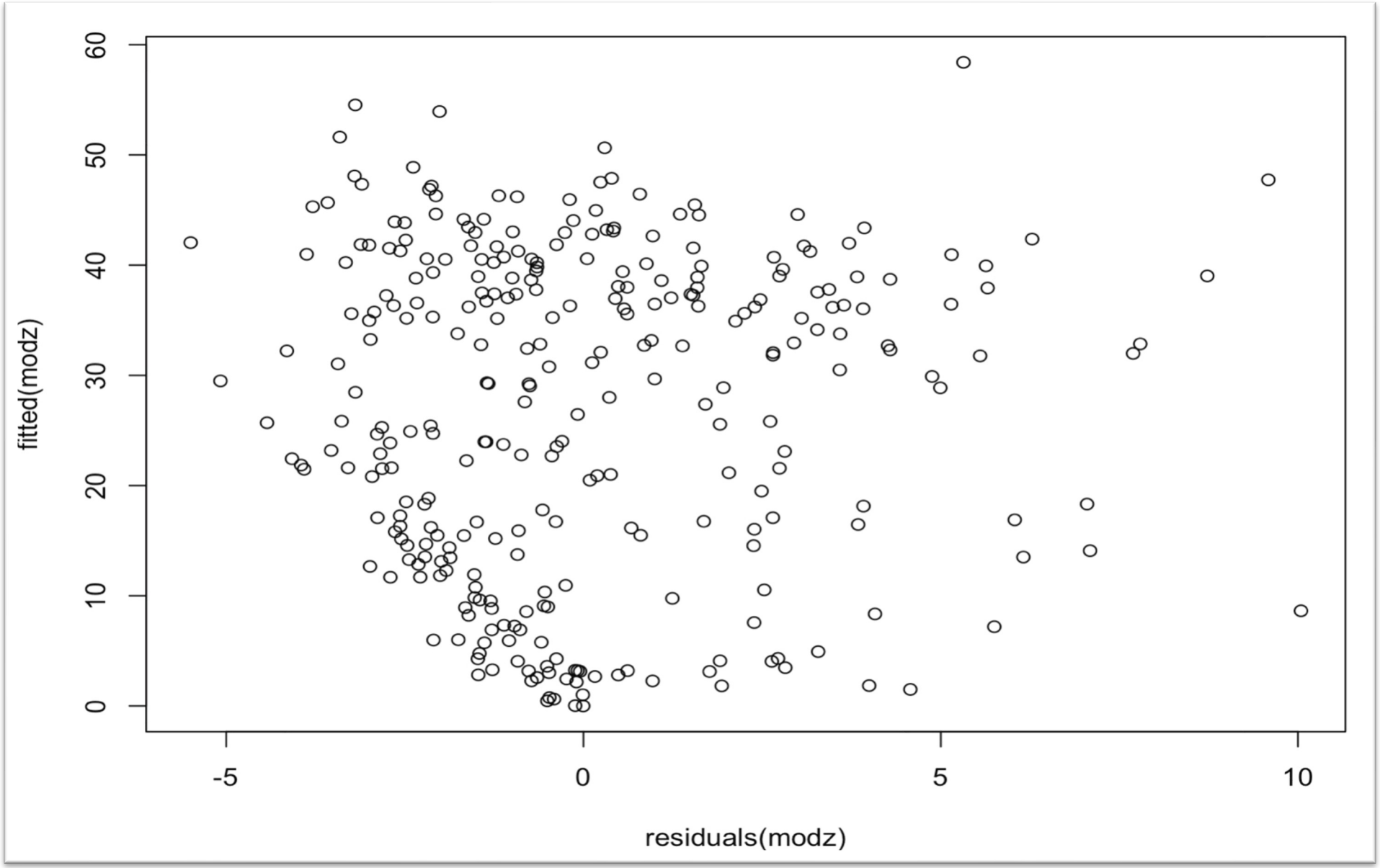
Python code used to create graphics in Final conclusion slides

```
import pandas as pd
from patsy import dmatrices
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

#data set imported with predictors already adjusted
df = pd.read_csv("/Users/alessandroesposito/Downloads/eggs", header=0)

for i in df:
    print(i)

print(df.head())

df.groupby('egg_count').count()

#Create the training and test data sets. Note that for now, we are not doing a stratified random split:
mask = np.random.rand(len(df)) < 0.8
df_train = df[mask]
df_test = df[~mask]
print('Training data set length=' +str(len(df_train)))
print('Testing data set length=' +str(len(df_test)))

#Setup the regression expression in Patsy notation. We are telling Patsy that FISH_COUNT is our dependent variable y and it
#depends on the regression variables Time + Salinity + Flow + S_depth + Temp_surf + Net_area + C_dist:

for i in df:
    print(i)

expr= 'egg_count ~ time +salinity + flow + s_depth + temp_surf + net_area + c_dist'

#Let's use Patsy to carve out the X and y matrices for the training and testing data sets.
y_train, X_train = dmatrices(expr, df_train, return_type='dataframe')
y_test, X_test = dmatrices(expr, df_test, return_type='dataframe')


#Using statsmodels's ZeroInflatedPoisson class, let's build and train a ZIP regression model on the training data set.

zip_training_results = sm.ZeroInflatedPoisson(endog=y_train, exog=X_train, exog_infl=X_train, inflation='logit').fit()
print(zip_training_results.summary())

zip_predictions = zip_training_results.predict(X_test,exog_infl=X_test)
predicted_counts=np.round(zip_predictions)
actual_counts = y_test["egg_count"]
print('ZIP RMSE='+str(np.sqrt(np.sum(np.power(np.subtract(predicted_counts,actual_counts),2)))))

fig = plt.figure()
fig.suptitle('Predicted versus actual counts using the ZIP model')
predicted, = plt.plot(X_test.index, predicted_counts, 'go-', label='Predicted')
actual, = plt.plot(X_test.index, actual_counts, 'ro-', label='Actual')
plt.legend(handles=[predicted, actual])
plt.show()
```

Poisson model residuals



ZIP model residuals vs fitted



Hurdle model

```
Call:
hurdle(formula = egg.count ~ time + salinity + log(flow) + log(s.depth) + temp.surf +
    I(net.area^2), data = mack)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.8809 -0.7111 -0.4703  0.3475 14.3802

Count model coefficients (truncated poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   20.69174    4.65947   4.441 8.96e-06 ***
time           0.30018    0.04557   6.587 4.50e-11 ***
salinity      -0.53543    0.13421  -3.990 6.62e-05 ***
log(flow)      0.15187    0.01138  13.340  < 2e-16 ***
log(s.depth)   1.76486    0.06243  28.268  < 2e-16 ***
temp.surf     -0.33690    0.01566 -21.518  < 2e-16 ***
I(net.area^2) 13.06633    1.32099   9.891  < 2e-16 ***
Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.25056   21.45672   0.058 0.953523
time          -0.51844    0.46576  -1.113 0.265662
salinity       0.16323    0.60983   0.268 0.788961
log(flow)     -0.05837    0.16349  -0.357 0.721096
log(s.depth)   1.53020    0.40465   3.782 0.000156 ***
temp.surf     -0.34137    0.09699  -3.520 0.000432 ***
I(net.area^2) -89.77781   32.46777  -2.765 0.005690 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 17
Log-likelihood: -3078 on 14 Df
```