Análisis Exploratorio y Estadístico de los Datos

Mortalidad de salmón del Atlántico por causa de bloom y OD

Felipe Tucca

Instituto Tecnológico del Salmón

2022-06-30

Estructura del trabajo exploratorio y estadístico

1) Introducción

- Descripción de la problemática.

2) Análisis exploratorio de los datos

- Histogramas biomasa muerta (toneladas) por causa.
- Boxplot biomasa muerta por causa para los últimos 10 años.

3) Análisis estadísitico de los datos

- Modelos lineales simples y múltiple.
- Comparación de modelos RSS-AIC.

Introducción

1). Descripción de la problemática

- Base de datos presenta registros de mortalidad por causa **bloom de algas y disminución de oxígeno disuelto (OD)**.
- 23 centros de cultivos reportaron la causal de mortalidad en salmón para un barrio de la Región de Los Lagos.
- *Salmo salar* es la especie más cultivada en este barrio.
- Los registros corresponden a reportes de mortalidad entre los años 2011 e incio del 2022 (Total de registros= 1224).
- Variables de estudio: Causa, peso (g), años, mes, semana e identificación de centro de cultivo.

Objetivos del estudio

- Evaluar la causa de mortalidad por bloom de algas y OD sobre la especie *Salmo salar* para un barrio del sur de Chile entre los años 2011 a inicios del 2022.
- Generar un modelo lineal que mejor ajuste la predicción de mortalidad en la biomasa de salmones.

Análisis exploratorio de los datos

- Datos desbalanceados por causa de muertes debido a bloom de algas (n= 360) y OD (n= 864).
- Año 2021 presentó una mayor biomasa muerta entre el 2011 al 2022.
- No existe correlación significativa ($p<0.05$) entre las causas de muerte por bloom de algas y disminución de oxígeno disuelto.
- Entre los periodos 2011 a 2022 existe una mayor biomasa muerta (ton) de *S. salar* en el barrio por causa de la acción de bloom de algas.

HISTOGRAMA

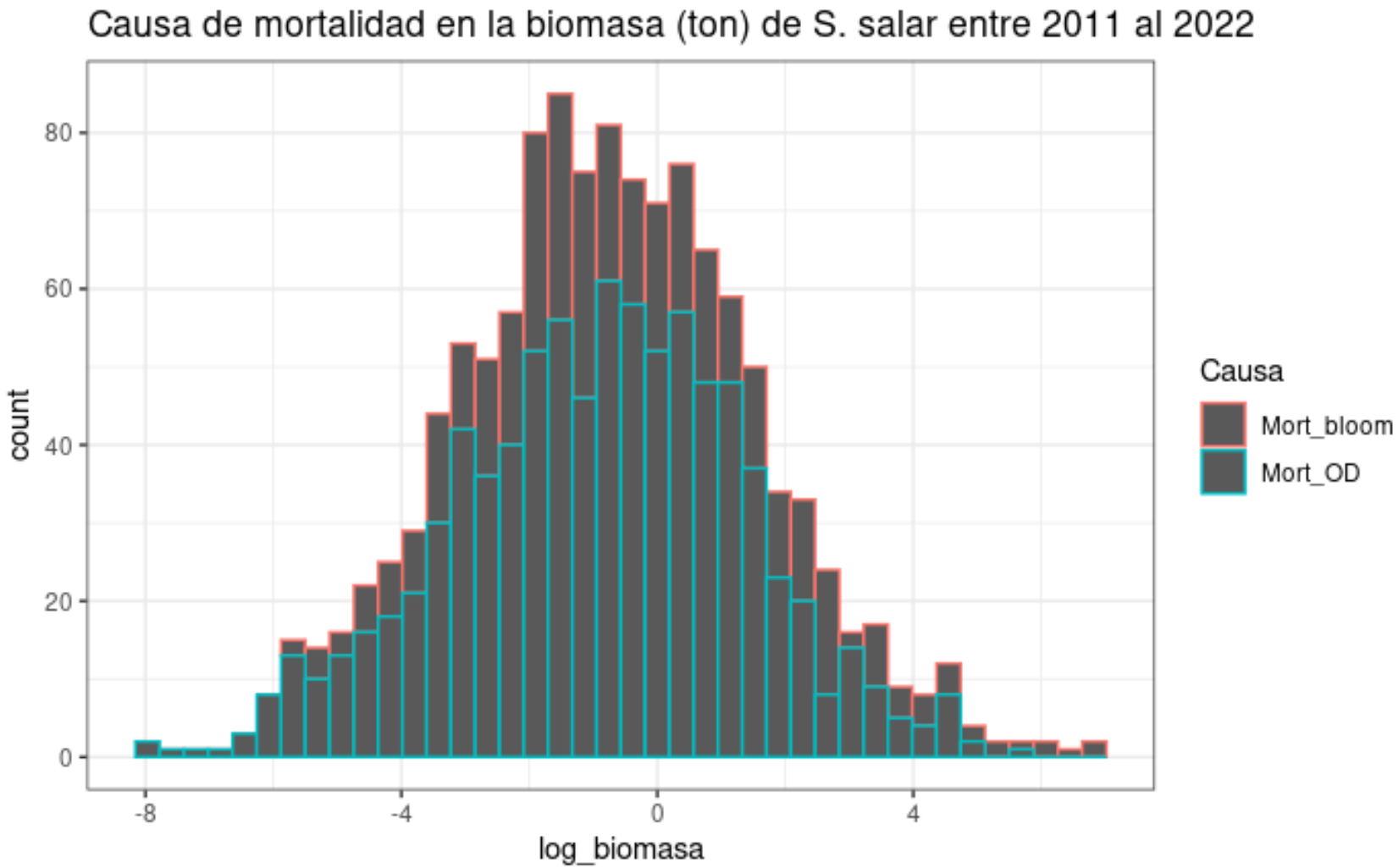

Figure 1: Histograma biomasa muerta (toneladas) por causa

Datos faltantes y datos atípicos: Boxplot

- Boxplot consideró la causa de muerte sobre la biomasa de peces entre el 2011 al 2022.
- Datos faltantes para la causa de mortalidad por bloom entre el periodo 2011 a 2022.
- Valores atípicos presente en las dos causas de mortalidad.

BOXPLOT

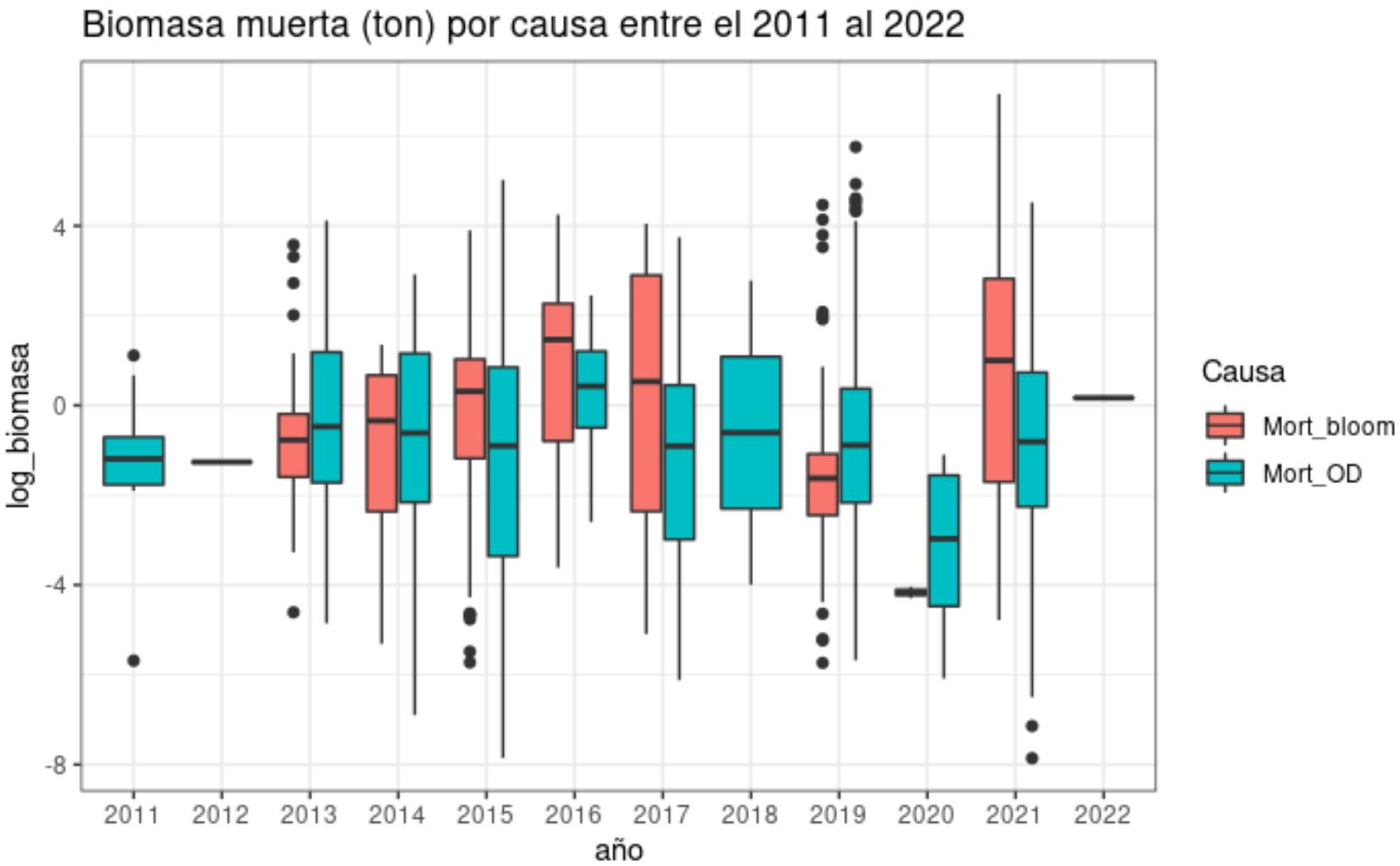

Figure 2: Boxplot biomasa muerta por año y causa

Análisis exploratorio de los datos: Plot

- Se evidencia sobre la ocurrencia de un evento temporal puntual que generó una alta mortalidad en la biomasa de salmones.
- Año 2021 presentó la mayor biomasa de salmones muertos siendo afectada por la **acción de bloom de algas**.
- Año 2021 presentó la mayor mortalidad registrada históricamente en este barrio (log biomasa muerta > 5).

Relación biomasa muerta y las semanas que se reportó mortalidad

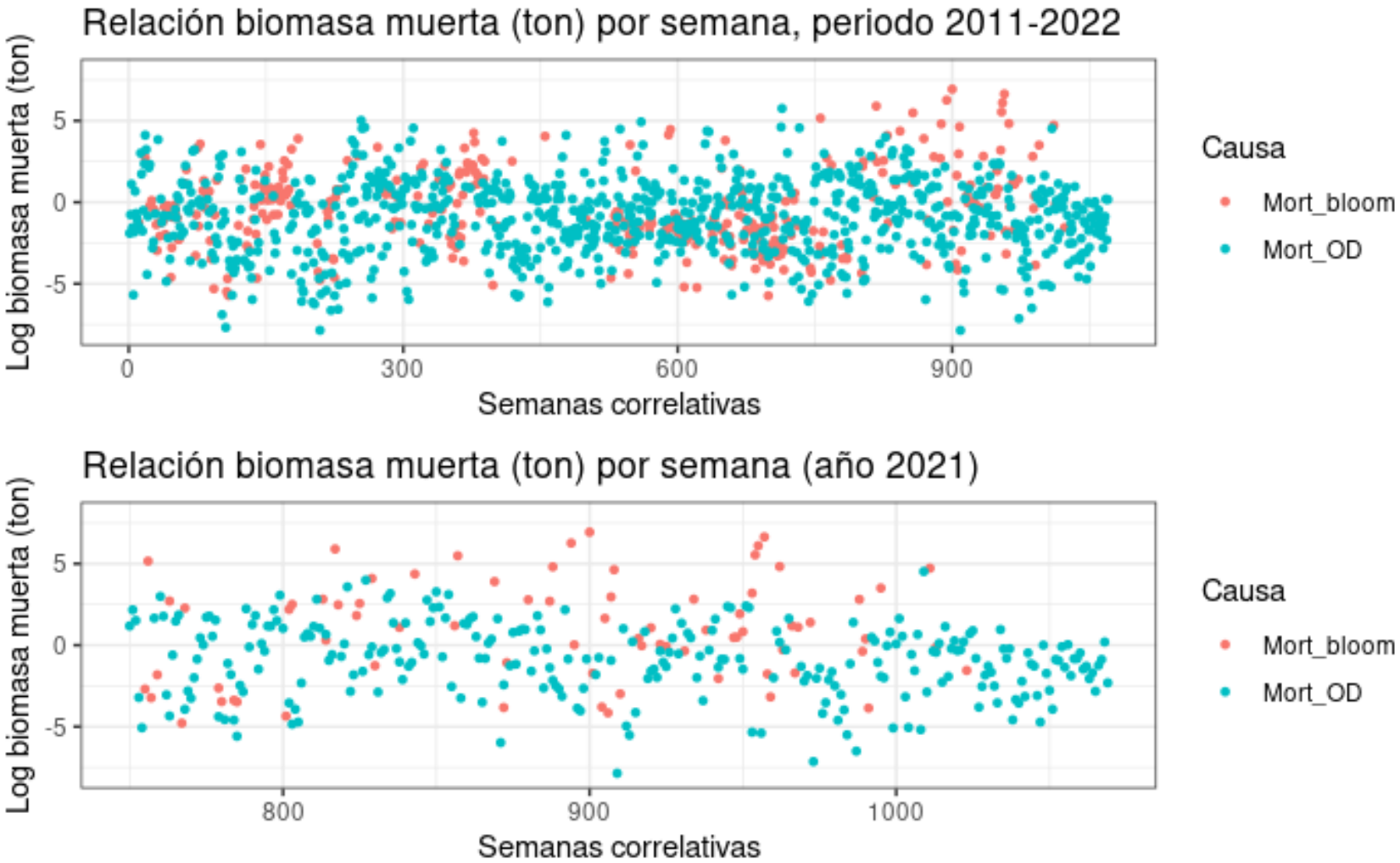

Figure 2: Biomasa muerta por semana y causa

Análisis exploratorio de los datos: Mortalidad bloom vs OD

- Mayor biomasa muerta es por causa de bloom de algas, alcanzando 16 toneledas en los ultimos 10 años.
- Mortalidad por OD alcanza las 4.1 toneladas.
- Peso promedio de los salmnes muertos fue de 3.1 kilogramos.

Tabla resumen de la biomasa muerta por causa

Table 1: Resumen de la biomasa muerta (toneladas) para la especie S. salar por causa entre los años 2011 al 2021

Causa	N	Promedio	DE	Mediana	Mín	
Mort_bloom	360	16.184949	82.38176	0.4478442	0.0032292	1031
Mort_OD	864	4.071008	16.81336	0.4349179	0.0003860	316

Análisis estadísitico de los datos: Modelo lineal simple

- **Modelo de regresión lineal simple** con los factores centro, semanas y años.
- Modelos de regresión lineal simple con los factores fueron estadísticamente significativos ($p < 0.05$), pero con un bajo ajuste o R^2 ajustado menor al 7%.

Hipótesis modelo lineal simple

- Basado en estos modelos de regresión simple se rechaza hipótesis nula que postulaba:

Hipótesis nula (H0): Existe similitud en la biomasa muerta entre centros/semanas/años.

Hipótesis alternativa (H1): No existe similitud en la biomasa muerta entre centros/semanas/años.

Hipótesis modelo lineal múltiple

- Para el **modelo de regresión múltiple** se postularon las siguientes hipótesis:

H0:

$$\beta_j = 0; j = 1, 2, ..., k$$

H1:

$$\beta_j \neq 0; j = 1, 2, ..., k$$

- El modelo cumplió con los tres supuestos: linealidad, homogeneidad de varianza y normalidad.

Análisis estadísitico de los datos: Ajuste modelo lineal múltiple

- La modelación integró los factores causa, centro, año, mes y la inetracción entre causa y año.
- Modelo nos estrega como resultado coeficientes distintos de cero, por lo tanto, se rechaza la HO (valores p menores al 5%).
- El R^2 ajustado de esta modelación múltiple correspondió al 23%.

Análisis de varianza (ANOVA)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Causa	1	63.15504	63.155039	14.624314	0.0001381
año	11	248.97448	22.634044	5.241187	0.0000000
centro_id	22	615.45852	27.975387	6.478040	0.0000000
mes	11	665.02969	60.457244	13.999607	0.0000000
Causa:año	7	175.41762	25.059660	5.802868	0.0000012
Residuals	1171	5056.95851	4.318496	NA	NA

Comparación de modelos por RSS y AIC

- Crietrios de anova de residuales (RSS) y Akaike Information Criterion (AIC)
- Ambos criterios sugieren al modelo lineal múltiple con la mejor predicción y ajuste (23%).

Quitting from lines 201-202 (Presentacion_Felipe_Tucca.Rmd) Error in anova(modelo1_anova1_centro, modelo2_anova2_mes, modelo3_anova3_año, : object 'modelo1_anova1_centro' not found Calls: ... eval_with_user_handlers -> eval -> eval -> %>% -> pander -> anova

Quitting from lines 205-207 (Presentacion_Felipe_Tucca.Rmd) Error in AIC(modelo1_anova1_centro, modelo2_anova2_mes, modelo3_anova3_año, : object 'modelo1_anova1_centro' not found Calls: ... eval_with_user_handlers -> eval -> eval -> %>% -> pander -> AIC

Interpretación y y conclusiones del trabajo

- Análisis exploratorio muestra mayor mortalidad de la biomasa de peces en el barrio por bloom.
- Mortalidad por baja de OD presenta mayor frecuencia.
- Se realizó ANOVA con un vía de criterio de clasificación para los factores centro de cultivo, semanas y años con ajustes menores al 7%.
- Modelo lineal múltiple agrupó todas los factores mostrando un significancia menor al 5%.
- El ajuste de la predicción de la variable biomasa muerta fue de un 23% (modelo regresión múltiple)
- Análisis comparativo por RSS y AIC entre modelos determinó que la **regresión lineal múltiple representa una mejor predicción**