Universidad del País Vasco Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

INGENIERÍA COMPUTACIONAL Y SISTEMAS INTELIGENTES

EXPLORACIÓN Y ANÁLISIS DE DATOS

Entrega 1

Estudiante: Eneko Perez

Estudiante: Alan García

Curso: 2024-2025

Fecha: 26 de septiembre de 2024

Índice

1. Enunciado

Este trabajo tiene como objetivo trabajar en la visualización de datos y la generación de gráficos con R. Para ello se utilizarán datos públicos de Open Data Euskadi, portal de datos abiertos del Gobierno Vasco. En este caso se trabajará con datos de la calidad de las aguas costeras y medio marino (año2024 y/o año 2023). Elige el/los conjunto/s de datos que más te interese para este trabajo.

Al final del trabajo se deberán entregar 3 gráficos junto con la explicación de cada uno (la pregunta que pretende contestar así como lo que se deduce del gráfico). Además, añadirá un pie de gráfico que incluye estos aspectos: de qué trata el gráfico; detalles específicos del gráfico como por ejemplo cómo se muestran las variables en el mismo; por último, el punto más destacado que muestra el gráfico. Empieza con preguntas simples. Puedes contestar una pregunta con 3 gráficos o realizar 3 gráficos para 3 diferentes preguntas.

Entrega

- Un fichero pdf con:
  - fuente de los datos y breve descripción.
  - los tres gráficos y sus explicaciones.
- Fichero csv con los datos estudiados.
- Fichero con el código R.
- Fecha de entrega: 26 de septiembre.

2. Resolución

Para la elaboración de esta práctica se ha empleado la fuente de datos provista por el portal *Open Data Euskadi* sobre la calidad de las aguas costeras en el año 2024. Concretamente se han utilizado los tres archivos *CSV* presentes: Puntos de muestreo; Mediciones (biológico) y Mediciones (físico/químico). Sin embargo, los datos sobre las mediciones se encuentran separados en varios archivos, por lo que para facilitar su utilización se han concatenado en dos archivos finales. En resúmen, los archivos de datos utilizados son los siguientes:

`samplePoints_2024.csv`

Este fichero contiene información a cerca de los puntos o localizaciones de muestreo en las cuales se miden los datos de los otros dos ficheros. Existen 52 localizaciones registradas y, por cada una de las entradas, se recogen los siguientes datos:

- **Code:** Identificador de la localización de muestreo
- **Sample.Point:** Nombre del litoral
- **Type:** tipo de litoral ('NORMAL' o 'LITORALES (MACROALGAS)')
- **Territory:** Provincia de Euskadi en la que se encuentra
- **Municipality:** Municipio de Euskadi
- **XETRS89:** Corrdenada en X del lugar de muestreo
- **YETRS89:** Corrdenada en Y
- **ZETRS89:** Corrdenada en Z
- **Water.Range:** Zona marítima en la que se encuentra
- **Water.Range.type:** Breve descripción de las aguas
- **Basin:** Cuenca / río de la estación
- **Section:** Sección del río en la que se encuentra
- **Subgroup:** Categoría de la estación de muestreo

Esta información es útil a la hora de relacionar los datos recogidos por las estaciones con sus respectivos nombres.

`measure_fq_2024.csv`

Este archivo recoge datos fisicoquímicos de los diferentes puntos de muestreo. Estas mediciones se dividen en dos subgrupos, el primero es de aguas (fisicoquímica) y el segundo de sedimentos. Del primer subgrupo se han tomado dos mediciones a lo largo del año, generalmente con tres meses de diferencia. En cambio, solo ha habido una medición en el de sedimentos. Estas mediciones no son iguales para todos los parámetros, para algunos parámetros hay diferentes mediciones

variando la profundidad entre 0 y 25 metros. En total el archivo cuenta con 3721 entradas de datos en las que se recogen: *Sample Point Code*, identificador del litoral, *Date*, fecha de muestreo; *Hour*, hora de muestreo; *Type*, tipo de muestreo (en estos datos siempre es 'LITORALES'); *Subgroup*, subgrupo del muestreo (Aguas (Fisicoquímica) o sedimentos); *Parameter*, parámetro medido (en este caso hay varios parámetros, por ejemplo: Benzo(b)fluoranteno, Fluoranteno, fluoreno, etc.); *Species*; *Operator*; *Value*, valor de la medición; *Unit*, unidad de medición (Hay varias unidades de medición, la más común es $\mu g/l$. Aunque, para los sedimentos la más común es mg/kg); *Additional.Information*, información adicional; *Situation*; *Level*, nivel de profundidad (toma los valores 'S' para superficie y 'F' para fondo) y *Depth*, profundidad del muestreo en metros.

`measure_bi_2024.csv`

Este archivo recoge datos biológicos de los puntos de muestreo. En general, cada estación toma dos muestras cada 6 meses: la primera en la superficie del agua y la segunda a 25 metros de profundidad. En total, el archivo cuenta con 72 entradas de datos en las que se recogen: *Code*, identificador de la estación; *Date*, fecha de muestreo; *Hour*, hora de muestreo; *Type*, tipo de muestreo (en estos datos siempre es 'LITORALES'); *Subgroup*, subgrupo del muestreo (en estas entradas siempre es 'Fitoplacton'); *Parameter*, parámetro medido (en este caso siempre es 'Clorofila A' debido a que es el indicador principal de la presencia de Fitoplácton); *Species*; *Operator*; *Value*, valor de la medición; *Unit*, unidad de medición ($\mu g/l$); *Additional.Information*, información adicional; *Situation*; *Level*, nivel de profundidad (toma los valores 'S' para superficie y 'F' para fondo) y *Depth*, profundidad del muestreo en metros.

Revisando estos datos, se ha planteado la siguiente cuestión inicial:

¿Cuál es el lugar con mayor concentración de fitoplancton recientemente en Euskadi?

Para intentar dar respuesta a esta pregunta se plantea el gráfico 1. En él se recogen los datos más recientes sobre la concentración ($\mu g/l$) de clorofila A de cada uno de los puntos de muestreo. Como se puede apreciar, por cada una de las estaciones de muestreo se presentan dos valores distintos: un primer valor sobre la concentración a 25 metros de profundidad y un segundo valor para la concentración en la superficie. Además, las estaciones están ordenadas de mayor a menor en función de las medias aritméticas entre los valores de superficie y fondo.

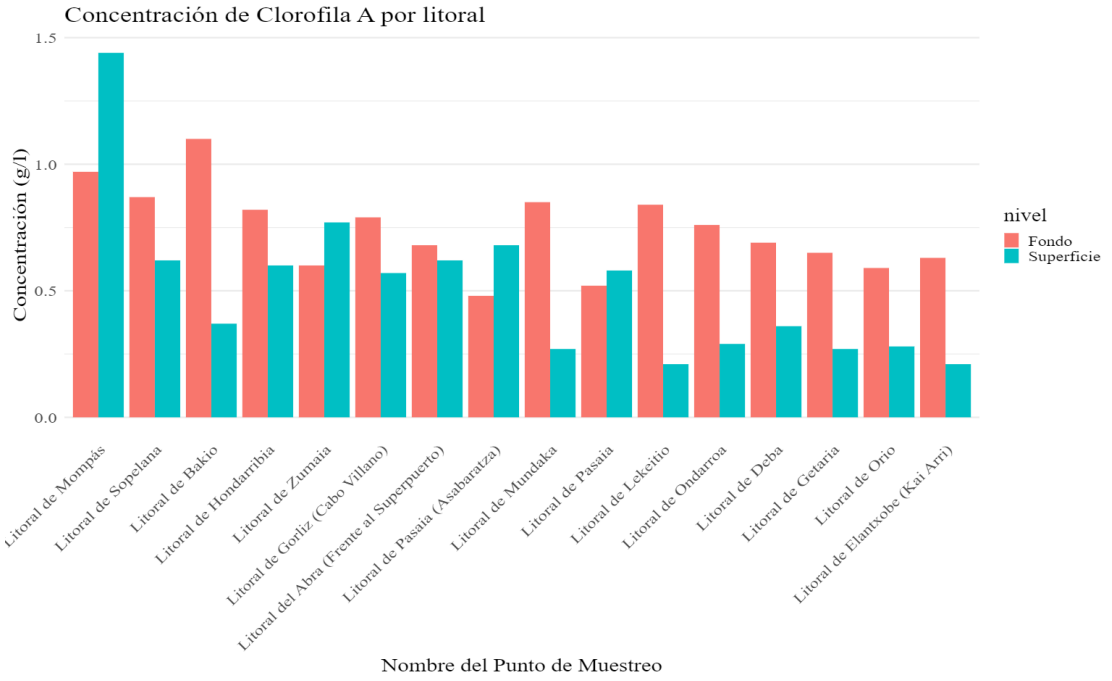

Figura 1: Concentración de Clorofila A por litoral

Con los resultados de esta gráfica podemos decir que en las últimas mediciones registradas, el litoral de Mompás es el que mayor índice de fitoplancton presenta en la superficie; mientras que el litoral de Bakio es el que más concentración tiene a 25 metros. Con esta información se ha planteado una segunda cuestión:

¿Hay alguna estación que tenga un mayor seguimiento?

De esta manera se ha propuesto el gráfico 2. En esta gráfica, se representan cada uno de los puntos de muestreo en el mapa geográfico mediante círculos de radio proporcional al número de muestras tomadas en dicha estación.

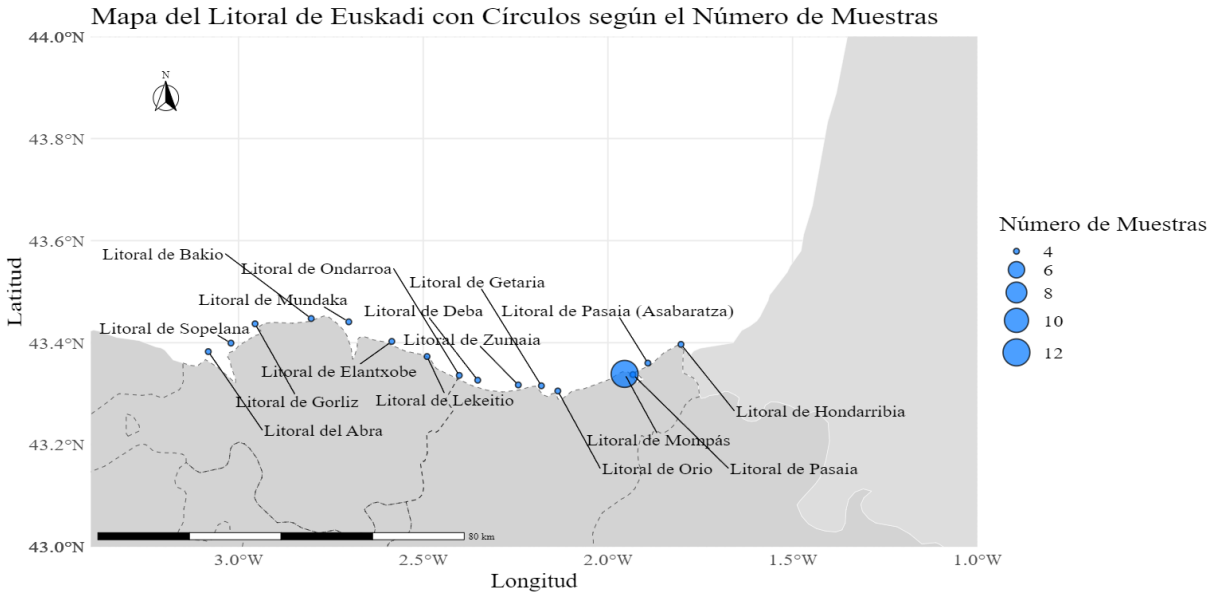

Figura 2: Mapa de numero de muestras. PROVISIONAL

Como se puede observar, el litoral de Mompás tiene un mayor seguimiento en cuanto al número de muestras que el resto de estaciones. Además, se ha elegido esta manera de representar los datos debido a que se planteaba que la cercanía geográfica pudiera influir en el número de muestras tomadas en regiones próximas a litorales con una atención alta. Sin embargo, esta hipótesis no se ha corroborado por ninguna de las dos gráficas mostradas.

De los lugares con más concentración de fitoplancton, ¿existe alguna relación entre la concentración de parámetros físico/químicos y de Clorofila A?

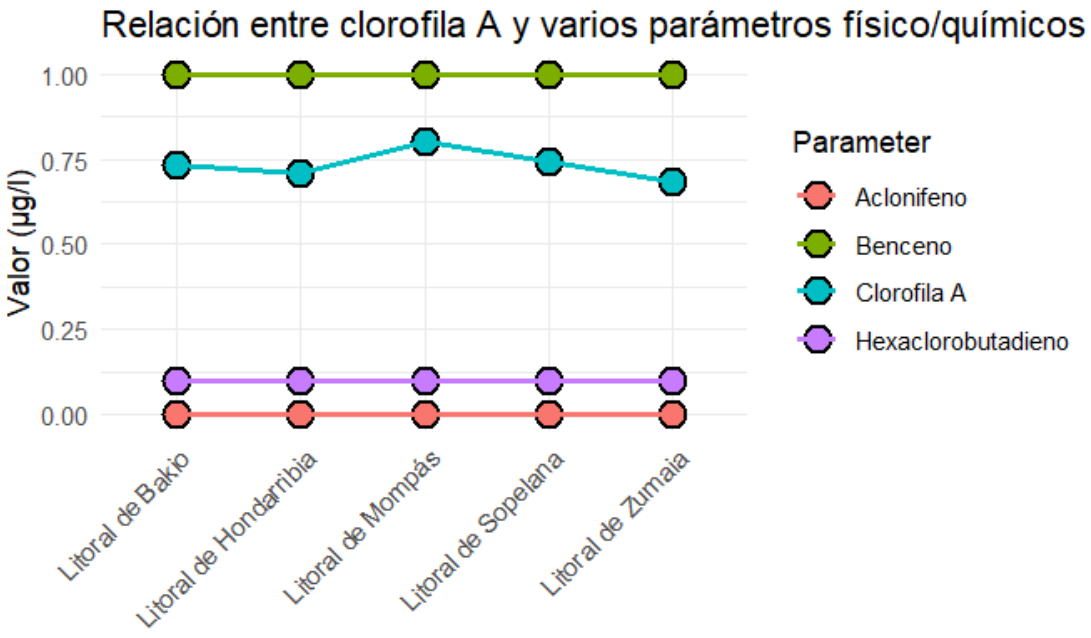

Figura 3: Relación entre clorofila A y varios parámetros físico/químicos

En la gráfica se puede observar como el valor medio de las mediciones de Clorofila A varia en los diferentes litorales. En cambio, todos los demás parámetros se mantienen constantes. El Benceno se mantiene alrededor de $1\mu g/l$ y el Aclonifeno y el Hexaclorobutadieno están alrededor del $0\mu g/l$. Por lo tanto, no existe una relación directa entre estos compuestos fisicoquímicos y la clorofila A.