

A - Índice y principal contenido de cada sección.

1 Resumen

El objeto del presente estudio es evaluar la calidad de agua en la costa cuenca del Plata para evaluar si la misma es apta para uso recreativo a través de mediciones en 42 estaciones desde la localidad de Berisso hasta la localidad de Tigre.

En particular, se estudiarán qué variables son las que tienen mayor incidencia sobre la calidad del agua y que zonas son las más afectadas aplicando análisis univariados y análisis de correlación de variables.

A fin de reducir la dimensionalidad del conjunto de datos se aplicará un análisis de componentes principales (PCA).

Se estudiará distintos algoritmos de clasificación supervisada detectando el más performante según la métrica elegida *Accuracy*.

2 Introducción

2.1 Contexto y motivación científica

La calidad de los cuerpos de agua superficiales relacionados con centros urbanos, constituye un elemento básico de la salud ambiental. Resulta incuestionable el impacto del estado del agua sobre indicadores salud ambiental no solo para el uso de sus costas como elemento de recreación e identidad de los ciudadanos sino como la morbilidad de numerosas enfermedades infecto contagiosas, la mortalidad infantil o su influencia en la esperanza de vida de los niños .

Este trabajo se focalizará en el estudio de la costa del Río de la Plata en la ciudad de Buenos Aires (CABA). La ciudad de Buenos Aires utiliza el agua proveniente del Río de la Plata como fuente principal para la generación de agua potable sin explotar su red de agua subterránea. La Ciudad está construida sobre un complejo sistema de desagües pluviales y cloacales que descargan en el río. Es decir que el Río de la Plata es, simultáneamente, principal fuente de agua potable para el área metropolitana y principal receptor de afluentes).

2.2 Objetivos del trabajo

La franja costera sur del Río de la Plata brinda a la sociedad un gran número de beneficios, como ser agua para consumo para los casi 15 millones de personas del Gran Buenos Aires y Gran La Plata (un 47 % de la población argentina), puertos, diversos espacios recreativos utilizados desde siempre por los habitantes de ese aglomerado urbano para realizar deportes acuáticos, nadar o pescar.

El objetivo del presente trabajo es estudiar las mediciones del RIIGLO desde el 2018 en adelante, entendiendo que variables de las medidas son las más relevantes y de mayor impacto en el índice de calidad de agua reportado. Analizar las zonas más afectadas y ver su evolución en los últimos años.

2.3 Organización

El trabajo consta de las siguientes 6 secciones:

1. Datos: se detallarán los datos utilizados, tipos de variables y faltantes.
2. Metodología: descripción de técnicas utilizadas para el manejo de datos

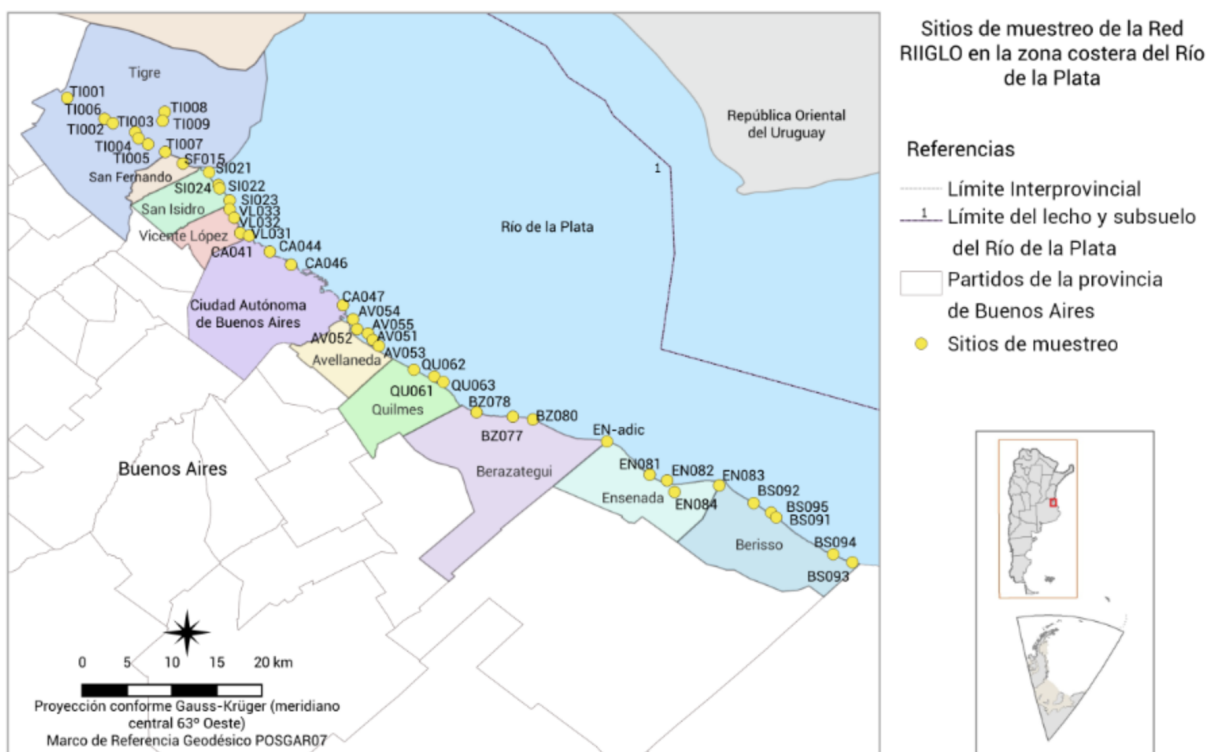
3. Resultados: se detallarán los resultados de los diferentes análisis y la incidencia de los mismos sobre la calidad del agua.
4. Conclusiones.
5. Bibliografía.
6. Anexo.

3 Datos

3.1 Información del conjunto de datos

El conjunto de datos seleccionados corresponde al relevamiento trimestral de la calidad del agua de la franja costera (lado argentino) del Río de la Plata realizada por el Ministerio de Ambiente y Desarrollo Sostenible en conjunto con los municipios que integran la Red de Intercambio de Información de los Gobiernos Locales (RIIGLO).

Las mediciones corresponden al período 2013-2022, realizadas en 42 estaciones determinadas dentro de la franja costera, desde la localidad de Berisso hasta Tigre como se observa en el gráfico inferior (ver Figura 1).



(Figura 1) Sitios de muestreo de la Red RIIGLO.

Es esta sección se detallan las variables a utilizar detallando como se calcula el índice de calidad de agua y el nivel de aptitud para uso recreativo con contacto directo (apta, levemente deteriorada, deteriorada, muy deteriorada, extremadamente deteriorada).

3.2 Preprocesamiento y limpieza de los datos

Se realizó como primera medida una limpieza de las bases para dejar los valores resultantes aptos para una correcta integración, unificando nombre de variables, validando unidades de medidas de las misma, formatos de los datos y eliminando el ruido (datos por ejemplo con anotaciones de +- o <).

3.3 Información faltante

Se detallan la información faltante y la decisión final de trabajar con la información del año 2018 en adelante e imputar los registros faltantes con la media de los valores por estación.

3.4 Análisis exploratorio de datos (AED)

3.4.1 Análisis univariado

En esta sección se recorre cada una de las variables que participan en el cálculo del ICA. Se exponen los resultados para la localidad de Tigre a fin de simplificar las visualizaciones.

4 Metodología

Se realizó un análisis univariado de cada una de las variables relevantes para el cálculo del ICA. La variabilidad y la correlación de los datos de calidad del agua se exploraron a través de diferentes enfoques estadísticos univariantes. Se utilizó estadística descriptiva para comparar las variables de los datos categóricos, estos resultados fueron presentados en la sección 3.4.1 Análisis univariado de parámetros.

A fin de analizar la existencia de correlación entre variables se realizaron gráficos de dispersogramas para apreciar la variación conjunta de cada par de variables cuantitativas y se cuantificó las mismas con el cálculo de la covarianza muestral. Adicionalmente se calculó la distancia de Mahalanobis y se efectuó el test de Chi Cuadrado para confirmar significancia estadística.

Se efectuó un análisis de componentes principales (PCA) para reducir la cantidad de variables a un número menor de PC (componentes principales) conservando la mayor cantidad posible de variación de las variables originales. El número de PC extraídas correspondió a aquellas que acumulaban el 65% de la varianza.

Adicionalmente, dado que contábamos con la variable de interés calidad de agua, se procedió a implementar métodos de clasificación supervisada para entender si el algoritmo podría ante una nueva muestra clasificar correctamente a qué grupo de calidad de agua corresponde. A fin de comparar los distintos algoritmos se establece como medida de performance la métrica *accuracy*. Se dividió el dataset entre variables categóricas y variables cuantitativas. Para las variables categóricas como olores, color, material suspendido, calidad del agua se aplicó mecanismos de label encoding. Dividimos el dataset en train and test en una proporción 80% para train y 20% para test. Probamos distintos algoritmos de clasificación y evaluamos su performance con la métrica *accuracy*. Los algoritmos evaluados fueron: *Logistic Regression*, *K Neighbors Classifier*, *Gaussian NB*, *LinearSVC*, *decision Tree classifier* y *Random Forest Classifier*.

5 Resultados

A fin de facilitar la exposición de resultados y como se mencionó anteriormente, se muestran los resultados de la localidad de Tigre y se mencionan los resultados a nivel global de la cuenca. En esta sesión se muestran los resultados por cada una de las técnicas aplicadas.

5.1 Correlación entre variables

5.2 Análisis de componente principales

5.3 Clasificación supervisada

6 Conclusión

A través del análisis univariado y de correlaciones, se identificó un grupo reducido de variables con mayor incidencia en la calidad del agua en la cuenca del plata, las mismas son la turbiedad, el OD y escherichia coli.

Cuando analizamos la turbiedad por localidad, se observa que los valores más altos corresponden a la localidad de Tigre, seguido por CABA y Berisso. Respecto al OD las localidades que registran los valores más altos corresponden a Berazategui, Ensenada y Quilmes. Para la escherichia coli las localidades con mayor presencia son Tigre, Berisso, y Vicente Lopez .

Analizando las localidades más afectadas, las mismas son Vicente Lopez, Berazategui, San Fernando y Berisso, siendo la estación BS095 la que se encuentra extremadamente deteriorada.

B - ChatGPT

A continuación se presentan las preguntas y respuestas a ChatGPT.

Pregunta: which data science methodology apply to water pollution analysis

Respuesta: ChatGPT

When it comes to analyzing water pollution, data science methodologies can be applied to gain insights and make informed decisions. Here are some common data science methodologies that can be used for water pollution analysis:

1. **Data Collection:** Gather relevant data on water quality parameters such as pH, dissolved oxygen levels, chemical concentrations, turbidity, and bacterial counts. This data can be collected from field measurements, sensor networks, satellite imagery, and public databases.
2. **Data Cleaning and Preprocessing:** Clean the collected data to remove errors, outliers, and missing values. Preprocess the data by normalizing, scaling, and transforming it as needed. This step ensures that the data is in a suitable format for analysis.
3. **Exploratory Data Analysis (EDA):** Perform exploratory data analysis to understand the distribution, patterns, and relationships within the water pollution data. **Visualizations, statistical techniques, and descriptive statistics can be employed to identify trends, correlations, and anomalies.**
4. **Feature Engineering:** Derive new features or transform existing features to enhance the predictive power of the data. For example, you might calculate additional metrics such as water quality index, pollution index, or average pollutant concentration over a specific time period.
5. **Machine Learning Modeling:** Apply various machine learning **techniques such as regression, classification, clustering, or anomaly detection to build predictive models.** These models can help in predicting water pollution levels, identifying pollution sources, or classifying water samples based on pollution levels.
6. **Model Evaluation and Validation:** Assess the performance of the developed models using appropriate evaluation metrics, such as accuracy, precision, recall, or root mean square error (RMSE). Validate the models using cross-validation techniques to ensure their robustness and generalizability.
7. **Visualization and Reporting:** Communicate the results of the analysis effectively using visualizations like charts, maps, and graphs. Create reports or dashboards to summarize the findings, highlight key insights, and provide actionable recommendations.

It's important to note that the specific data science methodology used for water pollution analysis may vary depending on the available data, research objectives, and the complexity of the problem. Domain knowledge and expertise in water quality and pollution dynamics are also crucial for accurate interpretation and meaningful analysis of the results.

Pregunta: which are the best visualization for water pollution

Respuesta: ChatGPT

There are several effective visualization techniques that can be used for water pollution analysis. The choice of visualization depends on the type of data and the specific analysis goals. Here are some commonly used visualizations for water pollution:

1. **Line Charts:** Line charts are useful for showing trends and variations in water pollution parameters over time. For example, you can plot the changes in pollutant concentrations or water quality indices over different time periods.
2. **Scatter Plots:** Scatter plots can be used to visualize the relationship between two variables, such as pollutant concentration and a specific water quality parameter (e.g., dissolved oxygen). Each data point represents a water sample, and the scatter plot helps identify correlations or patterns.
3. **Heat Maps:** Heat maps provide a visual representation of pollutant concentrations across different geographical locations. They can be particularly useful for identifying pollution hotspots or areas with high pollution levels.
4. **Chloropleth Maps:** Chloropleth maps use color shading to represent pollution levels in different regions. This type of map is effective for showing spatial variations in water pollution, especially when data is available at a regional or country level.
5. **Box Plots:** Box plots summarize the distribution of pollutant concentrations and provide information about the median, quartiles, and outliers. They can be used to compare pollution levels across different locations or time periods.
6. **Bar Charts:** Bar charts are useful for comparing pollutant concentrations or pollution levels between different categories. For example, you can use a bar chart to compare the concentration of different pollutants in different water sources.
7. **Geographic Information Systems (GIS):** GIS tools allow you to overlay water pollution data on maps, enabling you to visualize pollution levels in relation to geographical features such as rivers, lakes, or industrial areas. This can provide valuable insights into pollution sources and potential impacts.
8. **Animated Visualizations:** Animated visualizations can be used to show the temporal changes in water pollution parameters. For example, you can create an animation to visualize the spread of pollution over time or the movement of pollutant plumes in a river.

Remember that the choice of visualization technique should be guided by the specific requirements of your analysis and the story you want to convey. It's important to create clear, informative, and visually appealing visualizations that effectively communicate the insights derived from the water pollution data.

Pregunta: Is it worthy to proceed with PCA water consumption analysis

Respuesta: ChatGPT

Principal Component Analysis (PCA) can be a valuable technique for analyzing water consumption data, as it helps in identifying underlying patterns and reducing the dimensionality of the data. Here are some considerations to determine if PCA is suitable for your water consumption analysis:

1. **High Dimensionality:** If you have a high number of variables related to water consumption (such as different types of water usage, temporal factors, geographical factors, etc.), PCA can be useful in reducing the dimensionality of the data. It can help you identify the most significant components that explain the majority of the variance in the data.

2. **Multicollinearity:** If your water consumption variables are highly correlated with each other, PCA can help in capturing the shared variance among them. It can create linear combinations of variables (principal components) that are uncorrelated with each other, making it easier to interpret and analyze the data.
3. **Interpretability:** PCA provides a way to transform the original variables into a new set of orthogonal variables (principal components) that are ranked based on their importance in explaining the variance in the data. This can aid in identifying the most influential factors contributing to water consumption patterns.
4. **Visualization:** PCA can be used to visualize the data in a lower-dimensional space. By plotting the data based on the principal components, you can identify clusters or patterns that may not be evident in the original high-dimensional data.

However, it's important to consider the limitations and assumptions of PCA before proceeding:

1. **Linearity:** PCA assumes a linear relationship between variables. If the relationship between water consumption variables is highly nonlinear, PCA may not capture all the relevant information.
2. **Interpretation of Components:** While PCA provides a way to reduce the dimensionality of the data, interpreting the meaning of the resulting principal components can sometimes be challenging. It requires careful analysis and domain knowledge to interpret and assign meaningful interpretations to the components.
3. **Loss of Information:** PCA involves reducing the dimensionality of the data, which means there is some loss of information. You should carefully assess whether the loss of information is acceptable for your specific analysis objectives.

In conclusion, if you have a high-dimensional dataset with correlated variables and you are primarily interested in identifying patterns, reducing dimensionality, and exploring the most influential factors affecting water consumption, then PCA can be a valuable technique. However, it's important to assess the specific characteristics of your data and the objectives of your analysis before deciding to proceed with PCA.

Pregunta: which is the level of pollution in cuenca del plata Argentina

Respuesta: ChatGPT

As an AI language model, I don't have access to real-time data. However, I can provide some general information about water pollution in the Cuenca del Plata region in Argentina.

The Cuenca del Plata is a vast watershed that encompasses several countries, including Argentina, Brazil, Bolivia, Paraguay, and Uruguay. It is home to the Paraguay, Paraná, and Uruguay rivers, which are major water bodies in the region.

Water pollution in the Cuenca del Plata can stem from various sources, including industrial activities, agriculture, urban runoff, and untreated sewage. Specific pollutants can include heavy metals, pesticides, fertilizers, organic waste, and sedimentation.

The pollution levels in the Cuenca del Plata can vary depending on factors such as location, proximity to pollution sources, seasonal variations, and local management practices. Different stretches of the rivers may exhibit varying degrees of pollution due to differences in industrial and agricultural activities along their courses.

To obtain up-to-date and accurate information on the current level of pollution in the Cuenca del Plata, it is recommended to refer to local environmental agencies, water management organizations, or research institutions that monitor and report on water quality in the region. These organizations may conduct regular monitoring and provide reports or data on pollution levels in specific areas of the Cuenca del Plata.