Centro Universitario de Estudios Medioambientales.

Seminario de la reunión semanal del CUEM.

Seminario: 2025-06-02

Expositores: Rigalli, Alfredo

Tema: Manejo de grandes volúmenes de datos

Los datos son los elementos esenciales en una investigación. Sin datos no hay investigación. Un dato es el valor de una variable que puede ser medida, calculada o asignada. Los datos permiten observar de manera objetiva un fenómeno, analizarlo o registrarlo. No todos los datos de una investigación son utilizados, muchos de ellos pueden ser parte de la investigación y permitir el cálculo de otras variables más significativas. En el caso de la base de datos de Atlantis, pongamos de ejemplo tres datos: El código del agua, como ser el caso A529. Este dato importante permite identificar una muestra de agua de manera unívoca, y se trata de un dato asignado. La concentración de arsénico, es otro dato valioso, y se trata de un dato calculado que se obtiene en base a las mediciones de la transmitancia de estándares, qc y muestra. Por último el pH es una variable medida directamente con el pHmetro. La calidad de los datos asegura un análisis posterior y alcanzar conclusiones valederas. Sin calidad de datos no hay investigación seria y menos conclusiones valiosas. Son varias las etapas involucradas en garantizar la calidad de los datos: proceso de medición, almacenamiento y cálculos posteriores. Conocer el error de las mediciones aumenta la calidad de un dato.

Las bases de datos como datosaguas, la base de Atlantis o datoseoloBdI, la base de EoloBdI son bases que se expanden diariamente y en el primer caso ronda los 60000 datos, mientras que EoloBdI supera el 1000000.

Disponer de herramientas que permitan depurar y controlar los datos, es una necesidad si se desea trabajar de manera correcta y lograr que el tiempo invertido conduzca a buenas conclusiones y organice el trabajo de manera correcta. Este proceso de organización de una base de datos se llama curado de la base de datos.

El manejo y curado de los datos de una base como datoseoloBdI puede llevar años a una persona con herramientas convencionales como la observación y el contraste con los registros originales. Por esta razón es necesario el manejo de herramientas informáticas que permitan con rapidez y eficiencia observar los datos de la base.

Existen herramientas de fácil aplicación como son cálculos estadísticos. La observación de una simple gráfica de boxplot de todos los datos de una variable, permitirá fácilmente identificar valores muy alejados de los valores habituales. Los bigotes de esta gráfica muestran los valores extremos y si hay un dato erróneo, muy alejado de lo habitual, será fácilmente identificado. Muchas más conclusiones podemos obtener de estas gráficas. Atlantis tiene una herramienta que permite observar el boxplot, histograma y gráfica de dispersión de los datos de una variable. Además Atlantis dispone de una herramienta que envía un email a director y vicedirector cuando una variable excede un rango preestablecido, siendo un valor raro o no habitual. Dispone también de herramientas que permiten observar la calidad de mediciones a través de análisis multivariado de datos y que acompañan a al informe de un análisis de agua. Esto da al informe mayor confiabilidad y permite hacer ciertos controles antes de su generación.

Entre las herramientas que disponemos para observar grandes volúmenes de datos, tenemos las que nos provee la estadística descriptiva: media, mediana, percentilos, rango, desvío estándar. También disponemos de gráficas inagotables en diseño, por mencionar las más conocidas, boxplot, dispersión e histogramas. Estas herramientas se manejan con códigos sencillos de R o cualquier otro lenguaje de computación. Herramientas más complejas las constituyen funciones de R más sofisticadas como lo son subset, tapply y order, entre las más comunes. Las más sofisticadas son los scripts, que son secuencias de órdenes que pueden incluir a las anteriores o algoritmos más sofisticados. Los scripts son construidos por nosotros en función de las necesidades y permiten el manejo de millones de datos por segundo.

Un conocimiento importante a la hora de garantizar calidad de datos es el conocimiento del funcionamiento de los instrumentos de medición, el tipo de fallas que puede dar y su frecuencia.