Gestión de Grandes Bases de Datos en el Área de los Recursos Hídricos: Aplicación de Técnicas de Exploración y Preprocesamiento de Datos (Minería de Datos)

Pablo Facundo Andreoni[1], María Inés Rodriguez[1], Marcia Ruiz[1,2], Clarita Dasso[1], Leticia Vicario[1], Laura Colladón[1] y Ana Laura Ruibal Conti[1,2]

[1] Instituto Nacional de Agua. Centro de la Región Semiárida.

[2] Universidad Católica de Córdoba. Facultad de Ciencias Químicas

E-mail: pablofacuandreoni@yahoo.com.ar

RESUMEN

El objetivo de este trabajo es mostrar la aplicabilidad y utilidad de distintas técnicas, en el marco de la minería de datos, para la evaluación de datos en el área de recursos hídricos y su posterior aplicación en la gestión. En particular, los volúmenes grandes de información y la complejidad de los procesos de obtención de datos, que pueden hacerlos vulnerables a errores u omisiones, hacen preciso complementar las técnicas clásicas de la informática con las posibilidades analíticas de la ciencia de datos, para dar respuesta a esta problemática concreta y construir una base sólida de datos para análisis complejos. La importancia de este trabajo radica en ofrecer un marco metodológico de referencia para la exploración y preprocesamiento de datos crudos, paso inicial en la aplicación de minería de datos.

En particular, se presentan y describen metodologías utilizadas para el tratamiento de dos variables: estratificación térmica de la columna de agua del Embalse San Roque y precipitación registrada en su entorno. En concreto, se llevaron a cabo los siguientes tres procedimientos: a) complementación de los criterios del dominio de aplicación (limnología) con criterios estadísticos basados en los datos, b) cálculo de variables accesorias y c) definición de comportamiento esperado de la variable y desviaciones (anomalías), que permiten más acertadamente definir los valores de las variables de estudio.

Como resultado de la aplicación de las metodologías anteriormente mencionadas, pudo evidenciarse cómo, para cada uno de los dos casos de estudio, se consiguió sortear dos de las problemáticas más comunes en lo que respecta al preprocesamiento de los datos: completar valores faltantes (que de otro modo habrían supuesto una disminución en la disponibilidad de una magnitud crítica) y filtrar valores extremos (que habrían sesgado una métrica agregada) a la hora de calcular las variables de interés.

INTRODUCCIÓN

En los ecosistemas acuáticos se producen procesos asociados a la calidad del agua que pueden abarcar desde períodos cortos (horas, días, o estacionales) hasta otros más extensos (quinquenios, décadas). Estos últimos suelen estar asociados a la variabilidad climática de la región (Grimm et *al*., 2013, Malve at *al*., 2012). La evaluación de los procesos a largo plazo, requieren de un monitoreo permanente de variables de calidad de agua, así como de variables meteorológicas e hidrológicas. Desde el año 1998 INA-CIRSA, a través de su programa permanente "*Monitoreo del Embalse San Roque y Gestión de Información de Calidad de Aguas y Cianobacterias"*, monitorea mensualmente el agua del embalse San Roque, y de sus ríos tributarios para estudios que permiten lograr un mayor conocimiento del funcionamiento del sistema. Los datos recabados, los cuales incluyen parámetros de calidad de agua físico-químicos y biológicos se almacenan en una extensa base de datos. Estos parámetros, revisten importancia en términos del seguimiento de la condición eutrófica crónica del lago, que tiene impacto directo en la calidad del agua (INA-CIRSA, 2014). De modo similar el INA-CIRSA también obtiene datos hidrológicos y meteorológicos (precipitación, temperatura, nivel de ríos, entre otras) a través de un sistema telemétrico con estaciones de monitoreo en la cuenca de del embalse San Roque.

El desafío en la evaluación de los procesos a largo plazo es la estructuración adecuada y análisis de grandes bases de datos. Estos repositorios de información, en formato de archivos de texto o planillas de cálculo (con un volumen de información en el orden de los miles de registros en los que refiere a calidad del agua, y de los centenares de miles de registros en lo que refiere a variables hidrológicas y meteorológicas), no están exentos de errores, faltantes y desviaciones. Teniendo en cuenta la complejidad del despliegue montado para el relevamiento de los datos, sumado a las características propias de funcionamiento y reemplazo (ya sea por rotura o sustitución por tecnología más moderna) de los instrumentos de medición, y el obrar humano, hace que los datos no siempre sean correctos. Esta problemática, requiere del uso de tecnología adecuada para acondicionar información tan valiosa, sin que esta manipulación implique imprecisiones e incertezas en los datos. Del mismo modo, el desafío también abarca la organización de los datos para una visualización precisa y a la vez simple para su difusión y entendimiento por diferentes sectores de la sociedad.

Surge así, la necesidad de procesar y combinar las mencionadas bases de datos multivariables. A partir del uso extensivo de las nuevas tecnologías en materia de almacenamiento de información es posible gestionar de una manera rápida y confiable los datos. En este sentido, es que se combinan técnicas de la informática, de uso ampliamente extendido, como lo es el almacenamiento de la información en bases SQL (*Structured Query Language*), por un lado, que permite consultar y realizar cálculos complejos sobre los datos; con técnicas de la ciencia de datos, de más reciente e incipiente aplicación, para el análisis exploratorio, entendido como "el proceso de calcular varios valores agregados (resumidos) y derivados dada una determinada colección de datos", incluyendo actividades tales como: formular preguntas, elegir métodos de análisis, preparar los datos

para la aplicación de dichos métodos, aplicar los métodos a los datos e interpretar y evaluar los resultados obtenidos (Adrienko & Adrienko, 2006).

En este sentido, en el presente estudio se expondrá el tratamiento realizado a dos variables, una correspondiente a la base de datos de calidad de agua (caso 1) y la otra a la base de datos de variables hídricas y meteorológicas (caso 2). En cuanto al primer caso, se detallará el cálculo de la variable de estratificación térmica, que se encuentra potencialmente relacionada con el proceso de eutrofización y de allí su relevancia, junto con sus variables ad-hoc, a partir de variables observadas en campo (un caso de "derivación" según la definición del párrafo anterior). En el segundo caso, se detallará la metodología de procesamiento de datos de precipitación registrada en distintas cuencas para la composición de un único registro diario, a partir de varios registros por día, para cada estación (un caso de "agregación", según la misma definición del párrafo anterior). El presente trabajo refleja la etapa inicial en minería de datos, o preprocesamiento de información, que permitirá finalmente evaluar la progresión temporal de la incidencia de las condiciones meteorológicas sobre la calidad del agua en el embalse. En un sentido más amplio, también se busca desarrollar una plataforma virtual que permita la difusión de la información en un formato sencillo y comprensible y al alcance de diferentes sectores de la sociedad.

METODOLOGÍA

Caso 1: Estratificación térmica

La base de datos de calidad de agua, contiene un registro mensual por cada sitio de muestreo (en adelante *sitio*) y por cada nivel de profundidad definido (en adelante *Z*). El monitoreo de las aguas se lleva a cabo sobre seis sitios del embalse: cuatro puntos en la desembocadura de los ríos tributarios (DCQ, DLM, DLC, DSA), centro (C) y presa (TAC). La figura 1 muestra la distribución espacial de los puntos mencionados y los cursos de agua relacionados.

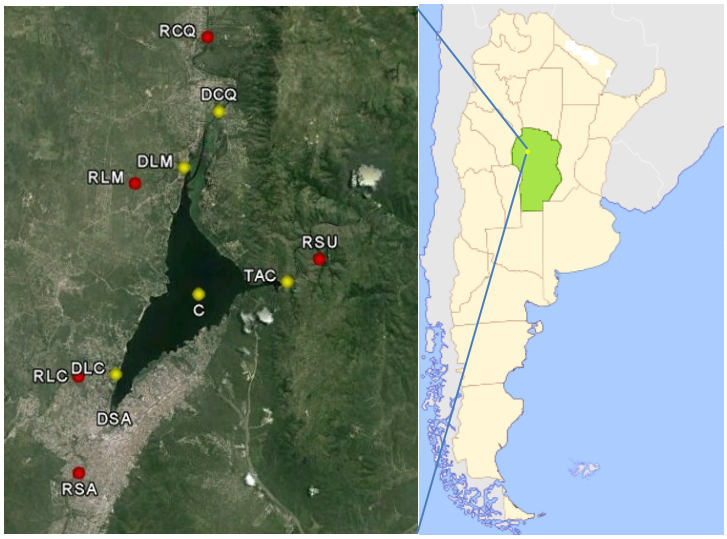

Figura 1.- Puntos de monitoreo (amarillo): DCQ=Desembocadura Cosquín, DLM=Desembocadura Las Mojarras, DLC=Desembocadura Los Chorrillos, DSA=Desembocadura San Antonio, C=centro y TAC=zona próxima a la presa, y cursos de agua (**rojo**): RSA=Rio San Antonio, RLM= Arroyo Las Mojarras, RLC=Arroyo Los Chorrillos, RSU=Rio Suquía, RCQ=Rio Cosquín, sobre el embalse San Roque (punto amarillo) en la provincia de Córdoba (en color verde), República Argentina.

En el caso de C, existen registros de muestras a 0,2m (subsuperficial) y desde 1m hasta 20m, cada 1m; para DCQ, a nivel subsuperficial y desde 1m hasta 4m cada 1m; para DLC y DLM solo a nivel subsuperficial; para DSA a nivel subsuperficial y desde 1m hasta 9m; por último, para TAC, a nivel superficial (0m), subsuperficial, y desde 0m hasta 33m cada 1m. En todos los casos, la cantidad de mediciones que se realizan para una fecha dependen de las condiciones hidrológicas en los puntos y la distancia al fondo.

Teniendo en cuenta la variación de Z y de la temperatura asociada a cada nivel, interesa saber el estado de estratificación térmica del lago para una fecha y un punto de muestreo dados. Dicho estado es en términos generales binario: *estratificación*, cuando existe una diferencia de temperatura tal que el agua más caliente (menos densa) se mantiene en las capas más superficiales, mientras que la más fría se deposita en capas más profundas; y *mezcla*, cuando la diferencia de temperatura entre las capas es menor, produciéndose una variación de temperatura en profundidad mucho más gradual. Existe evidencia de que el cambio de estados estratificación/mezcla potencialmente podría disparar el proceso de eutrofización de los lagos (Yu et *al*), y favorecer el desarrollo de cianobacterias. Por tal motivo cobra importancia la determinación de la estratificación térmica.

En términos de la base de datos, la categorización estratificación/mezcla constituye una variable *derivada* imputada por un experto mediante la observación de otras variables obtenidas directamente en campo. A efectos de sistematizar la determinación del estado de estratificación, se definió el siguiente procedimiento (registrando los resultados en una columna nueva de la tabla):

1) La base de datos completa, originalmente en formato planilla de cálculo (datos revisados y controlados por un experto) se cargó a una tabla "lago" dentro de una base de datos dedicada a este estudio.
2) Se aplicó la siguiente metodología de cálculo, mediante la definición de reglas de aplicación secuencial con formato de consultas al motor de base de datos:
    a. Si la variación de temperatura es mayor o igual a 1 grado (con redondeo a un dígito decimal) entre dos registros con valores de Z consecutivos (con 1m o menos de diferencia, excepto subsuperficial -0.2m-, donde se considera no determinante), entonces se establece que existe *estratificación*. Esta constituye una definición del dominio en el ámbito de la limnología (Cole, 1988; Hutchison, 1957).
    b. Si la variación de temperatura es de 0.9 grados (el redondeo se hace a un dígito decimal, así 0.89 se redondea a 0.9 y 0.84 a 0.8) para una variación en profundidad que no supere el metro (excepto subsuperficial, donde se considera no determinante), se requiere revisión y se indica como T, haciendo alusión a una condición de temperatura que no se cumple. En casos donde la variación de temperatura es próxima a 1 puede decidirse estratificación en función de otras variables a determinar por un experto.
    c. Si la variación de temperatura es superior al grado para una variación de profundidad que supera el metro, se requiere revisión y se indica como Z, haciendo alusión a una condición de profundidad que no se cumple. En caso en que exista variación de temperatura superior al grado con paso de profundidad superior al metro, puede decidirse estratificación en función de otras variables a determinar por un experto.
    d. En cualquier otro caso en que la temperatura no supere el grado se considera *mezcla* (teniendo en cuenta todos los casos en que se tenían al menos dos puntos para comparar temperatura, aunque los demás valores de temperatura faltaran o la distancia en Z fuera superior a 1m).
3) Adicionalmente se crearon dos variables (columnas) ad-hoc: Z-termoclina que registra la profundidad a la que se produce la estratificación (cuando existe) y Delta-temperatura, que registra la variación de temperatura entre profundidad subsuperficial y fondo (que varía según el sitio). Esta última variable, es de utilidad para el experto en la determinación del estado cuando no se cumplen las condiciones generales (casos b y c).

Caso 2: Precipitación diaria

La precipitación es una de las variables meteorológicas que se mide en las distintas estaciones distribuidas por las cuencas de los ríos tributarios, perilago y centro del lago, desde principios de los años 90 (y hasta el año 2000 en el caso de la estación del centro del lago). Los datos de precipitación se obtienen a partir de las mediciones que registra un sensor y se trasmiten telemétricamente a los servidores de INA-CIRSA donde se almacenan en archivos de texto plano. Se encuentran registrados a nivel instantáneo y a paso de un milímetro, vale decir que, en condiciones normales de funcionamiento, se espera que exista un registro por cada milímetro

adicional de lluvia que se mida. De esta manera, el valor informado por el sensor consiste de la acumulación de milímetros llovidos hasta el momento y desde la última vez en que se reinició. Sobre este último aspecto, cuando el sensor acumula una cierta cantidad de lluvia (que depende de parámetros de calibración que pueden variar de estación a estación y en función del tiempo), para que pueda seguir operando, se procede al volcado del líquido, y el acumulador vuelve a "cero" (en la práctica puede ser otro valor significativamente menor al valor de acumulación).

En caso de no producirse precipitación, el sensor envía una señal cada doce horas informando el último valor acumulado. De esta manera, estos datos pueden pensarse como una serie temporal de precipitaciones acumuladas instantáneas.

Los datos contenidos en archivos de texto en los servidores, uno por estación y año, registran: fecha, hora y precipitación. Estos archivos son revisados por un experto, y contienen comentarios relativos a determinadas cuestiones observadas como ser: saltos o perturbaciones en la serie de valores debidos a calibración o prueba, valores que se consideran erróneos por su magnitud, detalles acerca de las razones por las que el sensor no funcionó, entre otras.

A efectos de cálculo de la precipitación diaria, como una magnitud agregada de interés para su correlación con otras variables meteorológicas e hidrológicas a nivel diario y asimismo con datos de calidad del agua, se definió la siguiente metodología que se describe a continuación.

1) Los archivos de texto con los datos instantáneos son cargados en una tabla SQL dentro de la base de datos dedicada a este estudio. Los comentarios del experto (dispersos en los archivos) son estructurados en una columna "observaciones".
2) En la figura 2 puede observarse un diagrama esquemático del proceso de cálculo, a partir de reglas de ejecución secuencial a modo de consultas al motor de base de datos, que consta de los siguientes pasos:
    a. En primer término, de modo preliminar se marcan los registros que deben ignorarse (no participan) en el cálculo por tratarse de perturbaciones susceptibles de ser detectadas sistemáticamente, vale decir:

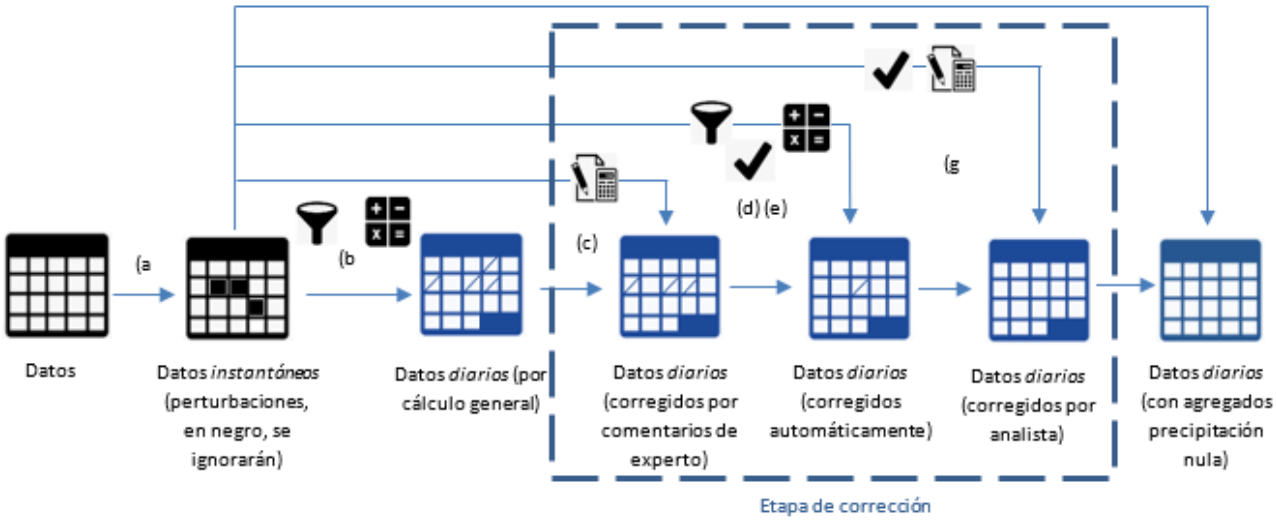

Figura 2.- Detalle del flujo de datos en el cálculo de precipitación diaria.

i. Caída abrupta en la serie temporal: sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: A primero, luego le sigue B y luego C, y además B es menor que A, y B es menor que C, y A es menor o igual que C. Asimismo, A se produce un día antes que B y las fechas de B y C coinciden; o C se produce un día después que B, y las fechas de A y B coinciden; o la precipitación B y C coinciden y las fechas son todas distintas.

ii. Caída aislada a 0 dentro del mismo día (representa un falso reinicio del acumulador en los términos en que se definió antes): sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: A primero, luego le sigue B y luego C, y además B es menor que A, y B es menor que C, y A es menor o igual que C. Asimismo, las fechas de A, B y C coinciden (sucedieron el mismo día), y B es igual a cero.

iii. Subida aislada en la serie temporal: sean A, B, C tres registros de precipitación, donde existe precedencia en el orden de las precipitaciones: A primero, luego le sigue B y luego C, y además B es mayor que A, y B es mayor que C, y A es menor o igual que C. Asimismo, A se produce un día antes que B y las fechas de B y C coinciden; o C se produce un día después que B, y las fechas de A y B coinciden; o la precipitación B y C coinciden y las fechas son todas distintas.

b. Posteriormente se procede al cálculo de precipitación diaria, en general, como diferencia entre el primer y el último registro que se disponga por día, en todos los casos en que dicho registro no corresponda a una perturbación, tal cual se definió en el punto anterior. En el caso particular de un único registro diario, atribuible a una falla del sensor en la transmisión del dato, la diferencia se resuelve cero.

c. Seguidamente se procede a la rectificación (si correspondiera) de los valores calculados. En primera instancia, se revisan las observaciones de los registros originales (a nivel instantáneo), y se determina:
    i. Si no se requiere tomar ninguna acción (el comentario no afecta el cálculo realizado tal cual se describe en el punto anterior). Por ejemplo: "se apagó el sensor por tormenta".
    ii. Se requiere una acción, porque de considerar el valor en el cálculo se tendría un resultado erróneo. Por ejemplo: "prueba" o "calibración", y caben dos posibilidades: Si el cálculo manual ignorando los valores no afecta el cálculo de los días adyacentes, se coloca como observación del registro diario: "Según comentario" o si, por el contrario, el cálculo manual afecta el cálculo de días adyacentes, para preservar la corrección se coloca como observación del registro diario: Según comentario: "Perturbación", de manera tal de que sea ignorado en los cálculos que se detallan a continuación.
d. Con posterioridad se realiza la corrección por reinicio del acumulador, denominada por diferencia negativa: sean A y B dos precipitaciones consecutivas que se producen el mismo día y en ese orden, si la diferencia de B y A es negativa, y siendo que el primer ni el último registro diario corresponden a perturbaciones, y la diferencia entre el último y el primero es menor que cero o bien B es cero, y siendo por último que ese día no fue objeto de corrección manual identificada como perturbación (como se detalla en el segundo inciso del punto anterior), entonces se procede al cálculo por partes. Es decir, por un lado, se calcula la diferencia entre A y el primer registro diario, y a eso se suma la diferencia entre el último registro diario y B.
e. Una particularidad que puede darse es que se haya registrado precipitación interdiaria, por la madrugada. Vale decir, cuando el último registro de un día, no coincide con el primero del día siguiente, se asume que llovió entre medio. Debido al principio de funcionamiento del sensor, si la diferencia entre uno y otro de los registros mencionados es de un milímetro y además el tiempo transcurrido entre los dos es igual o inferior a las doce horas (como se indicó anteriormente el sensor debe transmitir al menos cada doce horas), entonces se está en condiciones normales de operación y el milímetro llovido corresponde al segundo día. Debe cumplirse además que ninguno de los registros intervinientes sea considerado una perturbación ni tampoco se haya hecho una corrección manual por perturbación para el día en cuestión, según comentario. El milímetro se adiciona siempre al valor calculado para el día en cuestión.
f. En caso que no se cumpla alguna de las dos condiciones principales mencionadas, se procede al cálculo por repartición. Es decir, en situación donde la diferencia de precipitaciones es

mayor a un milímetro o bien han transcurrido más de doce horas, se asume mal funcionamiento del sensor o pérdida de información. En estas circunstancias no es posible determinar inequívocamente a qué día corresponde sumar la diferencia, por lo que se reparte entre el primer y el segundo día. Nuevamente, los registros intervinientes no deben estar considerados como perturbaciones ni tampoco ninguno de los días haber sido objeto de cálculo manual por perturbación atribuible a un comentario. Asimismo, se tiene en cuenta que el registro siguiente al primero del segundo día contenga una precipitación distinta que la del último registro del primer día (en tal caso se estaría ante una irregularidad, vale decir perturbación, de la serie y se ignora).

g. Por último, se realiza un control de valores diarios calculados:
    i. Cuando sean negativos, se los considera inválidos y se realiza la corrección manual. Estos casos constituyen situaciones excepcionales que no fueron contempladas en las reglas precedentes.
    ii. Cuando superen los 100 milímetros (menos del 0.5% de los casos), puede que el cálculo sea correcto y en tal caso no se interviene o bien puede suceder que, como en el punto anterior, sea producto de un caso no previsto y se corrija manualmente.

3) Un párrafo aparte corresponde a la revisión de los días con valores faltantes. En aquellos casos en que, el último registro de precipitación previo al(los) día(s) faltante(s), coincide con el correspondiente al primer registro siguiente, entonces significa que no se registró precipitación en ese período (siempre y cuando no exista un comentario que indique lo contrario), y por ende corresponde imputar cero. Esta imputación, está sujeta sin embargo a un análisis pormenorizado de la cantidad de días consecutivos con faltante de datos.

RESULTADOS Y DISCUSIÓN

Caso 1: Estratificación térmica

Siguiendo la metodología propuesta, de un total de 2.473 registros en la base, uno por fecha, sitio y profundidad de muestreo (téngase en cuenta que la condición de estratificación es única por fecha y sitio), se tiene que:

1) Originalmente, la condición térmica se había especificado para 1.700 registros (69%).
2) A partir de la aplicación de la metodología propuesta, se pudo calcular la condición térmica en 2.429 registros (98%), con la siguiente distribución: 458 “estratificación” y 1.771 “mezcla” (entre ambos 90%), además de 80 con “diferencia mínima de temperatura” y 120 con “diferencia en Z superior a 1m” para su revisión por un experto. Los restantes 314 registros no pudieron completarse debido a la

falta de datos. De entre los casos originalmente catalogados por el experto (1.700), 1.516 coincidieron exactamente con los calculados (89%), los restantes 184 son los sujetos a revisión.

Caso 2: Precipitación diaria

La cantidad de registros de precipitación instantánea es de 528.152 para un total de 27 estaciones a lo largo de 16 años en promedio cada una. A partir del seguimiento de la metodología propuesta pudieron calcularse un total de 135.749 registros diarios de precipitación. En este sentido, corresponde indicar el criterio adoptado para la imputación indicada en el punto 3) de la sección previa, basada en los resultados que se detallan a continuación.

La tabla 1 muestra la cantidad de días consecutivos sin dato junto con su frecuencia correspondiente, considerando el período completo de datos de todas las estaciones.

Tabla 1.- Cantidad de días consecutivos con precipitación faltante y su frecuencia

Nº dias	1	2	3	4	5	6	7	8	9	10	11	13	14	15	17	21	22	23	30	31	61	66	89	92	356
Frecuencia	911	251	85	42	26	11	5	5	4	6	3	3	4	2	1	1	1	1	1	8	1	1	1	1	1

Analizando la tabla 1, puede observarse que la distribución presenta una cola a derecha, concentrándose el 90% de los casos en el rango de 1 a 3 días, con tendencia decreciente en lo sucesivo. El valor aislado de 8 casos para 31 días, corresponde a una disfunción común del sensor, donde no se puede obtener el archivo completo para un mes entero (junto con el caso adicional previo de 30 días). A continuación, se muestra en la tabla 2, la frecuencia de ocurrencia de meses completos sin precipitación por estación, junto con la cantidad total de meses con dato y el total de meses completos sin dato normalizado por este último valor. Asimismo, se observa la distribución de esta última variable en la figura 3.

Tabla 2.- Cantidad de meses completos sin precipitación por estación y mes (NA=sin dato)

Estacion	May	Jun	Jul	Ago	Sep	Total Meses	con dato	Total Normalizado
100	0	3	5	2	1	11	118	0.09322034
200	0	1	1	4	1	7	269	0.02602230
300	0	4	3	3	1	11	277	0.03971119
400	0	4	3	3	2	12	266	0.04511278
500	0	2	1	3	1	7	289	0.02422145
600	1	6	6	10	3	26	290	0.08965517
700	1	3	5	6	1	16	280	0.05714286
900	1	4	3	3	1	12	277	0.04332130
1000	1	0	2	3	1	7	164	0.04268293
1010	0	2	3	2	1	8	138	0.05797101
1100	1	5	5	6	2	19	294	0.06462585
1200	0	5	6	5	0	16	296	0.05405405
1400	0	2	2	0	0	4	60	0.06666667
1700	0	1	2	3	0	6	102	0.05882353
1800	1	3	5	5	0	14	280	0.05000000
2000	0	0	1	2	0	3	101	0.02970297
2100	0	2	1	0	1	4	98	0.04081633
2200	0	3	2	1	1	7	79	0.08860759
2300	0	2	2	1	0	5	63	0.07936508
2400	0	3	4	6	1	14	260	0.05384615
2700	2	5	7	8	4	26	256	0.10156250
2900	0	3	3	1	0	7	72	0.09722222
3051	NA	NA	NA	NA	NA	NA	52	NA
3900	0	2	4	0	0	6	71	0.08450704
4050	0	1	2	1	0	4	57	0.07017544
4400	0	2	2	0	0	4	54	0.07407407
4800	0	0	0	1	0	1	56	0.01785714

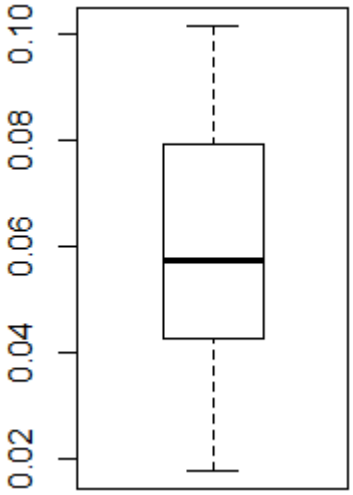

Figura 3.- Distribución del *Total Normalizado*.

Como se puede observar, la mayoría de las estaciones tienen al menos un mes en que no se ha registrado precipitación alguna. Al mismo tiempo se puede notar que los meses con mayor frecuencia corresponden a la temporada invernal y en menor medida fin del otoño. En lo que respecta a la distribución, se observa simétrica con media en 0.06 (alrededor de 6% de meses enteros con precipitación cero) no existiendo suficiente evidencia para rechazar la hipótesis de normalidad (test de Shapiro-Wilks: p-valor 0.5996, con 95% de confianza). Considerando el patrón mencionado anteriormente, se analizan los 8 casos de meses completos faltantes (frecuencia correspondiente a 31 días en la tabla 1), de los cuales 7 responden al patrón y 1 corresponde a la temporada estival, donde no existe evidencia de que no se haya producido precipitación alguna vez, y por ende se desestima. De los casos de menos de 30 días, existe evidencia, basada en frecuencias, suficiente y representativa de todas las longitudes. Con respecto a los casos de más de 31 días, se observa asimismo evidencia empírica suficiente para 61 y 66 días sin dato, con lo cual se completan los casos con precipitación cero. No así para el caso de 89 días consecutivos sin precipitación (que abarcan la temporada estival, donde no existe evidencia de tal cantidad de días sin lluvia), 92 días (donde se observa escasa representatividad: sólo ocurrió una única vez en 6 de las 27 estaciones) ni mucho menos para el caso de 356 días (considerado una anomalía del sensor).

CONCLUSIONES

Por medio del desarrollo realizado pudo observarse cómo el empleo de tecnologías de procesamiento de datos combinado con el análisis exploratorio permitió dar una respuesta satisfactoria a las necesidades de tratamiento de volúmenes grandes de información como es el caso de los datos de calidad de agua y meteorológicos. La metodología presentada puede servir como marco de referencia en lo que respecta a la derivación de nuevas variables calculadas tendientes a lograr una mayor comprensión de los fenómenos a largo plazo en los cuerpos de agua.

Así pudieron sortearse dos de las problemáticas más comunes en el ámbito de la minería de datos: imputación de valores faltantes (caso 1) como la detección y tratamiento de valores extremos (caso 2), además de perfeccionar y exponer las reglas de cálculo de variables secundarias y agregadas, respectivamente.

REFERENCIAS

Adrienko G., Adrienko N., 2006. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*, Springer, Berlin, Alemania, pp 1-2.

Cole, G. A., 1988. *Manual de limnología*, Hemisferio Sur, Buenos Aires, Argentina.

Grim, N. B., Chapin, S. F., Bierwagen, B., González, P., Groffman, P. M., Luo, Y., Melton, F., Nadelhoffer, K., Pairis, A., Raymond, P. A., Schimel, J., y Williamson, C. E., 2013. *The impacts of climate change on ecosystem structure and function*. University of Texas at Dallas.

Hutchison, E.G., 1957. *A treatise on limnology, Volumen I, Part 1*, Wiley Interscience Publication, USA.

INA-CIRSA, 2014. *Actividad permanente de monitoreo del lago San Roque y su cuenca*, Informe periódico Nº 2.

Malve, O., Jeppesen, E., Kernan, M., Goldsmith, B., Bennion, H., Huttula, T., Duel, H., Harezlak, V., Penning, E., Moe, J., Liukko, N., Kotamäki, N., Taskinen, A., 1/3/2009 –29/2/2012. *Deliverable D5.2-6: Synthesis paper on options for lake management to improve ecological status – Resistance to climate change in focus*, Finnish Environment Institute, 7th Framework Programme, pp 10-11.

Yu Z., Yang j., Amalfitano S., Yu X. & Liu L., 2014. *Effects of water stratification and mixing on microbial community structure in a subtropical deep reservoir*.