SOCIEDAD CHILENA DE INGENIERIA HIDRAULICA

XVIII CONGRESO CHILENO DE HIDRAULICA

USO COOPERATIVO DE MAPAS AUTO-ORGANIZATIVOS Y DIAGRAMAS DE PIPER EN EL ANÁLISIS DE MUESTRAS DE AGUAS SUBTERRÁNEAS

DIEGO RIVERA S.[1]
MARIO LILLO S.[2]
JOSÉ LUIS ARUMÍ R.[3]

RESUMEN

Este artículo presenta una metodología para analizar datos de calidad de aguas subterráneas aplicando una estrategia cooperativa para la caracterización de los datos, que combina las capacidades de visualización e identificación con las potencialidades de los mapas Auto-Organizativos (MAO's) para definir clusters. Un conjunto de datos publicados fue usado como un ejemplo de la aplicabilidad de la metodología propuesta. El uso cooperativo de los Diagramas de Piper y MAO's genera ventajas en la sistematización y actualización de bases de datos y principalmente incrementa la capacidad de discriminación en muestras de agua similares.

[1] Profesor Asistente, Departamento de Recursos Hídricos, Universidad de Concepción, Chillán dirivera@udec.cl
[2] Profesor Asociado, Departamento de Mecanización y Energía, Universidad de Concepción, Chillán
[3] Profesor Asociado, Departamento de Recursos Hídricos, Universidad de Concepción, Chillán

INTRODUCCIÓN

La densificación y extensión de los sistemas de monitoreo generan una gran cantidad de datos que deben ser procesados, jerarquizados y evaluados, de tal manera que sean entradas válidas a los modelos de simulación. Por lo tanto, la definición de metodologías de clasificación y clustering son una importante herramienta en la caracterización de los sistemas hidrológicos (Güler *et al.*, 2002). Entre las herramientas de clustering, los mapas auto-organizativos (MAO) han mostrado ventajas sobre otros métodos, e.g. gráficos y estadísticos, en el análisis y clustering de grandes conjuntos de datos (large data set, LDS) (Vesanto, 2000; Lisched, 2003; Lisched, 2005, Fang *et al.*, 2005).

En estudios de aguas subterráneas, las series de datos toman en cuentacaracterísticas, en escalas espacio temporales superpuestas, que pueden considerarse estáticas en el período a simular o variables en tiempo y espacio. Entre las propiedades estáticas se tienen las características hidráulicas del sistema (conductividad hidráulica, espesor estrato saturado, relaciones tensión-humedad) y respecto a las propiedades variables se pueden mencionar el contenido de humedad, la profundidad del nivel freático y la composición hidrogeoquímica de las aguas. Dado que muestras con similares características físico-químicas generalmente poseen historiales hidrogeológicos similares (Güler *et al.*, 2002), los datos de composición hidrogeoquímica, como el contenido de iones mayores presentes en el agua, han sido utilizados en estudios de origen del agua subterránea, caracterización de las fuentes y flujos, o cambios en los patrones de composición (Domenico & Schwartz, 2001).

Los datos hidrogeoquímicos colectados en una ubicación específica o en un área de estudio, pueden considerarse como vectores en un espacio *n*-dimensional donde existen relaciones que dificultan el uso de métodos determinísticos, estadísticos y gráficos. Por ejemplo, Demirel & Güler (2006) aplicando análisis multivariado, determinaron los principales mecanismos que controlan la química de la aguas subterráneas en Turquía; Ochsenkühn *et al.* (1997) utilizan métodos estadísticos multivariados y dendogramas para identificar clusters y caracterizar el flujo y direcciones predominantes en un acuífero a partir de datos químicos. Güler *et al.* (2002) aplicaron análisis jerárquico de cluster (hierarchical cluster analysis, HCA) en conjunción con Diagramas de Piper para la clasificación de LDS de calidad de las aguas subterráneas, mostrando las potencialidades de integrar métodos gráficos y analíticos.

Debe considerarse que el uso de datos reducidos dimensionalmente es válido si el conjunto reducido es representativo de los datos de entrada (Vesanto, 2000). En este sentido, los métodos gráficos (e.g. Diagramas de Stiff, Diagramas de Collins, Diagramas de Piper) y estadísticos (e.g. matrices de correlación, descriptores estadísticos, identificación de outliers, análisis de componentes principales) se basan principalmente en la reducción dimensional mediante proyecciones que no siempre conservan la topología de los datos, además de las probadas limitaciones en el análisis de conjuntos extensos (Güler *et al.*, 2002). Por otra parte, los dendogramas, que pueden clasificarse como algoritmos jerárquicos donde los datos son sucesivamente agrupados, aún cuando están basados en métricas euclidianas, mantienen un nivel de subjetividad en la definición de los grupos o clases (Güler *et al.*, 2002). Por estas razones, los algoritmos de aprendizaje, como las redes neuronales, son una alternativa de análisis de datos hidrogeoquímicos.

Dentro de las redes neuronales, los Mapas Auto-Organizativo (MAO's) permiten desarrollar tareas de clustering y generan una representación visual, dimensionalmente menor de los datos de entrada, a partir de una serie de vectores prototipos que representan las clases identificadas en el algoritmo. Una neurona es la unidad básica de trabajo en redes neuronales que tiene asociado un vector en un espacio *n*-dimensional. En un MAO, se considera un arreglo regular (hexagonal o rectangular) de P neuronas en un espacio 2-dimensional (Fig. 1), que permite realizar tareas clustering, clasificación o visualización de datos de alta dimensión (Vesanto & Alhoniemi, 2000).

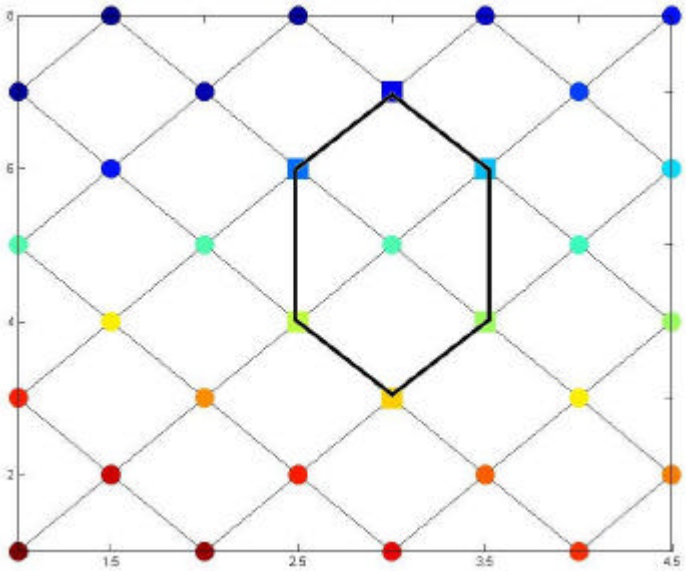

Figura 1 Topología hexagonal para un MAO. Cada neurona se ubica en los vértices de la malla y mediante un código de colores se presentan las relaciones de distancia entre los vectores prototipos de cada neurona. La elipse identifica la vecindad de la neurona central.

En la literatura reciente se han publicado aplicaciones de los MAO's al análisis de datos hidrogeoquímicos. Aguilera *et al.* (2001) a partir de las neuronas de un MAO generaron una forma de visualización en topología rectangular usando mapas de activación, i.e. un arreglo rectangular donde cada celda se corresponde a un vector definido por el MAO's. Las muestras son comparadas con cada celda y la más cercana es activada. Cada vez que la celda es activada el color en una escala de grises se acerca al negro. La clasificación de muestras desconocidas o el agrupamiento de un set de datos a partir, se realiza visualmente por similitud de estructura de los mapas de activación. Este análisis consideró la comparación de los resultados con los obtenidos mediante dendogramas. Sin embargo, en la clasificación mediante mapas de activación, aproxima la estructura topológica, pero no se visualizan las estructuras de distancias.

Sánchez-Martos *et al.* (2002) utilizan la metodología propuesta en Aguilera *et al.* (2001) en el análisis de datos de calidad de las aguas subterráneas. Tran *et al.* (2003) y Zhang *et al.* (2005) combinan las capacidades de los MAO's como algoritmo de clustering y la reducción dimensional Análisis de Componentes Principales (PCA) para clasificar cuencas; además, en esta última investigación, se realiza un agrupamiento de las neuronas siguiendo la metodología propuesta por Vesanto & Alhoniemi (2000). Finalmente, Peeters *et al.* (2006), incluyen las coordenadas espaciales asociadas a las muestras en el análisis de clustering.

Este artículo presenta una metodología de análisis de series de datos de aguas subterráneas utilizando de manera cooperativa dos herramientas para la definición de clusters: los Diagramas de Piper y los MAO's. Los primeros permiten una visualización e identificación rápida de datos de contenido iónico similares y de los elementos que definen esta similitud, pero no permiten extraer cuantificadores respecto a la posición de los cluster. Por otra parte, los MAO's son algoritmos de aprendizaje no supervisado ampliamente usados en problemas de clustering y minería de datos (*data mininig*) (Vesanto, 2002; Vesanto 2000, Kaski, 1997, Vesanto, 1997) que conservan la topología y mantienen la dimensión del espacio inicial, permitiendo un análisis cuantitativo de los clusters resultantes. Se analizan además las características vectoriales de los conjuntos de datos hidrogeológicos y se presenta una aplicación de esta metodología a un set de datos publicados por Harvey & Sibray (2001).

ANÁLISIS DE DATOS HIDROGEOQUÍMICOS USANDO MAO's

Mapas Auto-organizativos

El algoritmo básico de entrenamiento considera un conjunto de vectores de entrenamiento y un número determinado de neuronas que se ajustan al conjunto de entrenamiento en un proceso iterativo según (Vesanto, 2000):

1. Cada i-ésima neurona tiene asociado un vector prototipo en el espacio N-dimensional tal que $\mathbf{p}_i = \{p_{1i}, \ldots, p_{Ni}\}$ con i=1…P, donde P es el número de neuronas.
2. En cada paso de entrenamiento un vector **m** es arbitrariamente elegido del conjunto de muestras de entrenamiento.
3. Se calculan las distancias desde cada prototipo **p** al vector de entrenamiento **m**.
4. Se determina the Best-matching Unit (BMU), que es la neurona b cuyo prototipo $\mathbf{p}_b$ es el más cercano a **m**, es decir, $\|\mathbf{m} - \mathbf{p}_b\| = \min_i \{\|\mathbf{m} - \mathbf{p}_i\|\}$.
5. Luego, los vectores prototipo son actualizados, de tal manera que, para la iteración t+1, la BMU y sus vecinos son movidos a un lugar más cercano al vector de entrada según una función de vecindad H_b centrada en la BMU, tal que el prototipo de cada neurona i perteneciente al kernel de H_b es $\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t) \cdot H_b(t) \cdot [\mathbf{m} - \mathbf{p}_i(t)]$.

La principal característica de los MAO's, representado por los vectores prototipo **p**, es la reducción de la cardinalidad del conjunto de muestreo. Los prototipos asociados a cada neurona poseen la misma dimensión que los vectores de entrenamiento, por lo que es posible utilizarlos como representantes de clases o clusters, definiendo un espacio representativo que permite disminuir la complejidad del problema. De esta manera, es posible definir los cluster y estudiar las relaciones topológicas y de distancias de los datos de entrada utilizando los P-prototipos.

Debe considerarse que una vez que los datos han sido agrupados, el proceso de clustering puede ser realizado manualmente observando los mapas de visualización, pero este trabajo puede ser tedioso o inexacto, por lo que es necesario utilizar algún método alternativo. Vesanto & Alhoniemi (2000) presentan una metodología para definir cluster en MAO's aplicable a cualquier conjunto de datos, pero se requiere de programación y entrenamiento del usuario; por el contrario, los Diagramas de Piper y el uso de mapas de visualización de MAO's es cooperativo y no requiere de mayor entrenamiento por parte del usuario.

Para la visualización de los MAO's se utiliza generalmente matrices de distancia como la Matriz Unificada de Distancias (U-matrix), que calcula las distancias usando los vectores en el espacio

de los prototipos y despliega las estructuras de distancia en un arreglo bidimensional para las neuronas, manteniendo la topología y permitiendo la identificación de cluster, límites y neuronas representativas. Las distancia entre neuronas vecinas son representadas por una escala de color (Ultsch, 2003). Generalmente los colores oscuros corresponden a grandes distancias entre neuronas, mientras que colores claros corresponden a pequeñas distancias. Dada esta escala de color, los clusters son representados por áreas claras y los bordes por zonas oscuras. Como un vector de entrenamiento está asociado a una neurona o vector prototipo **p** que cumple con la condición de BMU, para efectos de análisis es posible asociar una etiqueta del vector de entrenamiento a la neurona prototipo.

Además de la U-matrix, es posible visualizar los planos de componentes, que representan la distribución de una variable en el conjunto de datos de entrenamiento. Cada plano componente puede considerarse como un corte en el MAO, desplegando los valores de un componente del vector prototipo en cada neurona con una escala de colores (Ultsch & Herrmann, 2005). Por inspección visual es posible identificar variables con similar distribución, lo que permite detectar posibles correlaciones entre variables (Ultsch & Herrmann, 2005; Vesanto, 1999).

DATOS HIDROGEOQUÍMICOS

La aplicación de metodologías clásicas, al análisis de datos hidrogeoquímicos han sido útiles, pero dadas las herramientas y los volúmenes de datos actualmente disponibles es posible cuantificar de manera más precisa la similitud y clasificación de diferentes muestras de aguas.
Un conjunto de datos de contenido iónico Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, CO_3^{2-}, HCO_3^{-}, SO_4^{2-}, puede ser representado por un vector **m** donde cada una de sus componentes se corresponde con los valores de concentración de los iones de las especies antes mencionadas, tal que $\mathbf{m} = \{m_1,\ldots,m_i,\ldots,m_N\}$, donde N es el número de parámetros. Si se dispone de un conjunto de mediciones en uno o varios puntos de monitoreo, estos forman una serie **M** (temporal y/o espacial), tal que $\mathbf{M} = \{M_s\}$, donde s es el número de muestras tomadas. De lo anterior, es posible establecer la siguiente conjetura:
Conjetura sobre la naturaleza vectorial de datos de monitoreo. Una serie de N datos de contenido iónico (u otro conjunto de parámetros) tomados en un arreglo espacial de puntos de control, son vectores que forman una hiper-superficie de dimensión *p* en un espacio R^{N+1} tal que $p<N+1$, donde $N+1$ considera la agregación de la dimensión temporal.

Los datos de calidad de aguas subterráneas son definidos como vectores en un espacio $(N+1)$-dimensional. Este planteamiento es extensible a cualquier conjunto de datos de monitoreo (e.g. aguas superficiales, subterráneas o de suelos) distribuidos espacial y temporalmente. La dimensión de la hiper-superficie dependerá de las relaciones que puedan establecerse entre las componentes del vector **m**, restringida al dominio impuesto por los límites físicos de las variables medidas (e.g. 0<pH<14). Estas relaciones son generalmente complejas y no lineales, a menos que se disponga de relaciones funcionales conocidas .

El problema de clasificación e identificación de tendencias en el conjunto de datos hidrogeoquímicos, puede considerarse como un problema de aprendizaje no supervisado o clustering, cuyo objetivo es agrupar objetos según un criterio determinado (Vesanto, 2000). El proceso de clustering es la división de un conjunto **Q** en C subconjuntos o clusters q_i , i=1,…,C (Vesanto & Alhoniemi, 2000). Existen variados algoritmos de clustering y formas de visualizar

los resultados (e.g. dendrogramas); o bien algoritmos de partición como el k-means que dividen el conjunto de datos en k-clusters, generalmente minimizando una función de error.

Por lo tanto, si se consideran las ventajas en cuanto a visualización, cuantificación de distancias, identificación de cluster y generación de elementos representativos de un cluster específico, es válido considerar que el uso de MAO's permite mejorar las tareas de clasificación de métodos gráficos como los diagramas de Piper, especialmente en conjuntos de datos donde la dependencia entre los componentes no es conocida o no se tiene supuestos iniciales respecto al origen y similitud de las muestras.

METODOLOGÍA PROPUESTA

La metodología considera la aplicación e integración de las propuestas en el Principio de Similitud expuesto por Rivera (2006), Sánchez-Martos *et al.* (2002), Vesanto (2000), Vesanto (1997), Vesanto & Alhoniemi (2000), Kaski (1997) y Vesanto (2002) en un proceso iterativo y continuo que es parte integrante del proceso de conceptualización y entendimiento del fenómeno estudiado.

El proceso de análisis de muestras de aguas subterráneas (raw data) comienza con el pre-procesamiento de los datos, eliminando valores fuera de rango físico, muestreos incompletos o valores bajo los límites de detección de los métodos químico-analíticos. Después de esta etapa, se dispone de una matriz de $R = [R]_{s\times N}$ (filas×columnas), donde s es el número de muestras y N es el número de parámetros.

Debido a que los valores de los parámetros de entrada pueden tener órdenes de magnitud distintos y que el ajuste del espacio de la neuronas respecto al espacio de los datos de entrada considera la norma euclidiana para calcular las distancias, es necesario escalar o normalizar los datos, de tal manera que los valores en la matriz de entrada sean comparables. La elección de los métodos de normalización para obtener la matriz de datos normalizada $\tilde{R} = [\tilde{R}]_{s\times N}$ influye en el desempeño y ajuste de los MAO's. La matriz $\tilde{R}$ es utilizada como entrada para la construcción de los MAO's, obteniéndose una matriz $M=[M]_{P\times N}$, donde P es el número de neuronas (prototipos) del MAO. Esta matriz es utilizada para calcular la U-matrix que permite visualizar e identificar los posibles clusters.

Las funciones de normalización consideran un valor x que se trasforma en un valor x'=f(x), con $x \in [0,1]$. Dentro de estos métodos de normalización (Vesanto, 2000) se tiene (1) Transformación lineal, que asigna valores 1 y 0, al máximo y mínimo valor de la serie de datos para el parámetro (2) Normalización según varianza, aplicando una transformación lineal que asigna valor promedio $\bar{x} = 0$ y varianza $\sigma_x = 1$ (3) Transformación logarítmica, aplicable datos con variación exponencial (4) Transformación logística aplicando la función logística a la variable normalizada según varianza.

Luego de entrenar un MAO, es importante conocer cómo este se ha adaptado a los datos de entrenamiento, por lo que es necesario aplicar distintos métodos de normalización y elegir aquel

que entregue los mejores valores de los indicadores de ajuste. La calidad de un mapa es generalmente evaluada usando las siguientes medidas:
(a) Precisión de mapeo, que describe cuan acertadamente las neuronas responden al conjunto de datos de entrada. Por ejemplo si el vector prototipo p_c de la BMU calculado para un vector de prueba m_i es exactamente el vector im_i, el error de presición es 0. Una medida común es el error medio de cuantización ε_q sobre el conjunto de datos:

$$\varepsilon_q = \frac{1}{s}\sum_{i=1}^{s}\left\|m_i - p_c\right\| \quad (1)$$

(b) Preservación de topología. El error topográfico es una medida que describe cuan bien el MAO preserva la topología del conjunto de datos. Un métdo simple para calcular el error topográfico ε_t aplica una función $u(x_i)$ que es 1 si la primera y segunda BMU de x_i no están cerca una de otra; en otro caso es cero:

$$\varepsilon_t = \frac{1}{s}\sum_{i=1}^{s}u(x_i) \quad (2)$$

Una vez definido el método de normalización, el mapa resultante puede ser visualizado utilizando la matriz unificada de distancias. Sin embargo, para analizar estos mapas es necesario conocer la correspondencia entre los vectores prototipo de cada neurona y los vectores de entrada. Para ello se utiliza un procedimiento de etiquetado que consiste en asignar a la neurona BMU la etiqueta del vector de entrenamiento.

De esta manera, cada vector de entrenamiento m_i asociado a la etiqueta l_i es comparado con cada vector prototipo **p**. Una vez determinado el vector prototipo p_b asociado a la neurona BMU b se construye el vector $\mathbf{L}=\{L_k\}$. Debe considerarse que en un mapa no necesariamente cada neurona es BMU de un único vector de entrenamiento, ya que el mapa se ajusta al espacio de los datos de entrada y no a los valores de entrenamiento. Por lo anterior, algunas neuronas serán BMU de varios vectores m_i. Generalmente de consideran tres procedimientos de etiquetado: (i) Se asigna la etiqueta que más veces ha sido ‘votada’, es decir, sólo una etiqueta por neurona (ii) Se asignan a cada neurona las etiquetas de cada vector m_i, para el cual la neurona es BMU (iii) Se etiqueta con un código numérico que refleje las frecuencias relativas de cada etiqueta.

Finalmente, el análisis de la información cualitativa contenida en las matrices de distancias y de prototipos, y la información visual del mapa, permite al modelador ajustar o ratificar las hipótesis iniciales y discriminar de manera más fina y objetiva la diferencia entre muestras de tal manera de mejorar el modelo conceptual asociado al fenómeno descrito por el set de datos (Vesanto, 2000). Estos ajustes en la conceptualización permiten a su vez mejorar las técnicas de muestreo (Lisched, 2005), tanto en extensión espacio temporal, como en aumento o disminución del número de muestras y parámetros.

ESTUDIO DE CASO

La metodología propuesta fue aplicada al conjunto de datos publicado por Harvey & Sibray (2001). Estos datos corresponden a monitoreos de contenido iónico de una serie de pozos de observación, a distintas profundidades, en las Great Plains, Nebraska.

El set de datos considera 49 muestras de contenido iónico, en mg L^{-1}, para $Ca2^{+}$, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, HCO_3^{-}, SO_4^{2-}, además de pH, T (ºC) y balance de cargas CB (%). De este conjunto se tomaron 40 muestras para entrenar el MAO y las restantes se utilizaron para analizar las capacidades de clasificación. La matriz de datos de entrada $R_{40\times10}$ tiene 40 filas correspondientes con el número de muestras y 10 columnas, correspondientes al número de parámetros considerados. Una característica importante del set de datos elegido es que cada muestra está etiquetada, es decir, es conocida su procedencia, profundidad de muestreo y fecha de muestreo. Con la información anterior se construyó un vector de etiquetas $\mathbf{l} = \{l_1,\ldots,l_s\}$ en el cual se almacena la información de procedencia de la muestra: R, para muestras del acuífero regional; D, para muestras del acuífero profundo ; M, para muestras tomadas en la parte media del acuífero y S para muestras tomadas en la parte somera del acuífero.

La Tabla 1 muestra la evaluación de los mapas generados a partir de 5 funciones de normalización. La importancia de normalizar los datos es clara respecto a los valores de las medidas de calidad de los MAO's. En efecto, el menor average quantization error es diez veces menor una vez que los datos han sido escalados mediante una transformación lineal aún cuando el topographic error es nulo.

Dado que el conjunto de vectores prototipos o *codebook* del mapa es dependiente de los vectores de entrenamiento, cambios en la cardinalidad y en los vectores cambiarán la calidad del mapa. En el caso extremo, si se dispone de un conjunto de *j* elementos y se toman *k* elementos de este conjunto se tendrán C_k^j combinaciones posibles. Sin embargo, en la práctica se elige aleatoriamente un porcentaje de elementos del conjunto de entrenamiento como elementos de verificación.

Tabla 1 Medidas de ajuste para los SOM's entrenados aplicando diferentes funciones de normalización.

Función normalización	Números de neuronas	Error medio de cuantización
Raw data	32	28.5083
Transformación lineal	32	0.2883
Varianza	32	1.289
Transformación logarítmica	30	0.9801

RESULTADOS & DISCUSIÓN

Los resultados de la metodología propuesta pueden dividirse en cuatro tópicos: (i) Efecto de la normalización de los datos (ii) Potencialidades de visualización (iii) Clasificación de muestras desconocidas, y (iv) Uso cooperativo con Diagramas de Piper.

La Fig 2 muestra el Diagrama de Piper para los datos analizados. Domenico & Schwartz (1997) y Chadha (1999) señalan que el agrupamiento de datos similares debe realizarse gráficamente en el diamante central del Diagrama de Piper. Como se observa en la figura, las muestras tomadas en profundidades medias y shallow tienden a agruparse en el centro del diagrama. Las muestras profundas tiene un comportamiento disperso, pero es reconocible el grupo que conforman. Asimismo, se observa una mezcla en la zona central entre muestras provenientes de

profundidades medias y profundas. Debe considerarse que estos diagramas no incluyen la temperatura del agua ni el balance de cargas, pero implícitamente se considera el pH, lado que existe una relación de equilibrio entre las concentraciones de HCO_3^- y CO_3^{-2}.

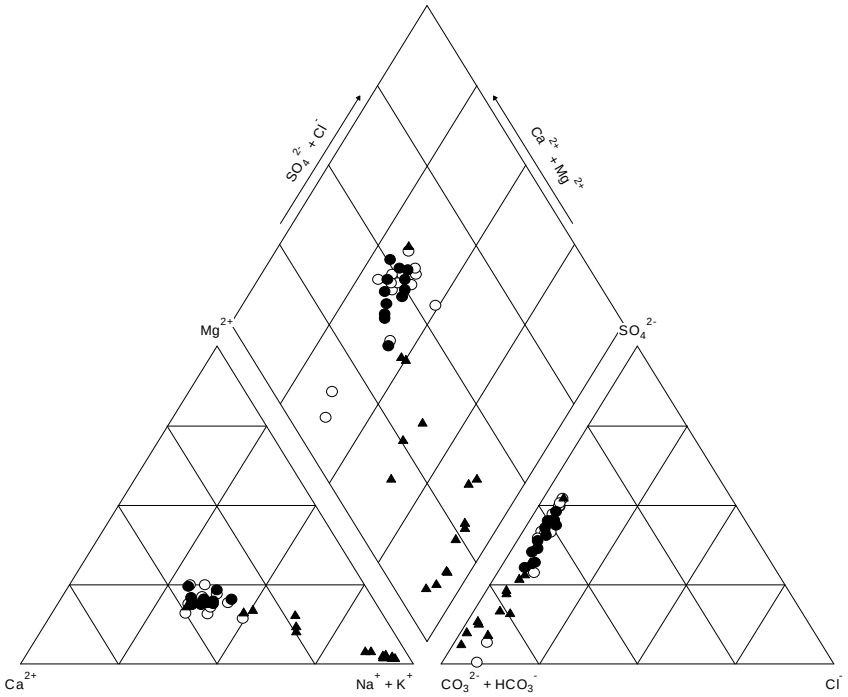

Figura 2 Diagrama de Piper para el set de datos de entrenamiento. ? Medium wells, ? Shallow wells, ? Deep/Regional wells.

NORMALIZACIÓN DE LOS DATOS

La figura 3 muestra las matrices unificadas de distancia y el etiquetado para el conjunto de datos presentado en la Tabla 2, sin normalizar los datos de entrada y aplicando una normalización lineal que genere un escalamiento de los datos en el intervalo [0,1].

La U-matrix resultante de los datos sin normalizar muestra una frontera definida entre las neuronas en la zona superior izquierda, pero el etiquetado de las neuronas no es consistente con los resultados utilizando el Diagrama de Piper (Fig. 2). La mejor definición en la frontera se debe al hecho que el rango de valores para algunos parámetros es dos órdenes de magnitud superior al resto (e.g. HCO_3^- y pH), lo que implica una distorsión en las distancias. Por otra parte, la U-matrix para los datos normalizados no muestra claras fronteras que definan clusters, pero el etiquetado es consistente con el diagrama de Piper.

Como se observa en la Fig. 3b, las etiquetas muestran una suave continuidad en el espacio de las neuronas, apreciándose como las muestras etiquetadas como S, se mezclan con las etiquetas M y estas últimas con las etiquetas D. La suavidad en estas transiciones se debe a las características del conjunto de muestras, ya que de no existir condiciones de confinamiento, se tendrá un perfil de transición en la composición del agua a diferentes profundidades.

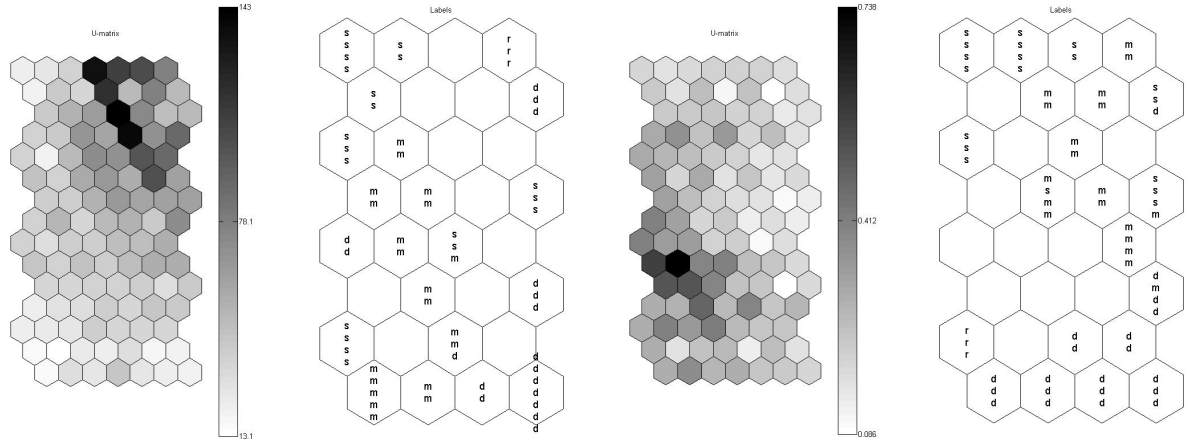

Figura 3 Matriz Unificada de Distancias U-matrix visualizada mediante una escala de grises y etiquetado de las neuronas en una topología hexagonal para: (a) datos sin normalizar (b) datos normalizados según transformación lineal.

POTENCIALIDADES DE VISUALIZACIÓN

La visualización de la relaciones de distancias entre las distintas neuronas y la preservación de la topología de los datos, son algunas de las ventajas del uso de MAO's en la identificación de clusters. Otra herramienta importante es la visualización de planos de los componentes. Establecidos los clusters, esta herramienta permite mostrar los valores de cada variable en los clusters y, mediante inspección visual buscar correlaciones entre variables, expresadas en patrones similares (Vesanto, 2000; Peeters *et al*., 2006).

La figura 4 muestra los planos de componentes para las variables analizadas. Para la visualización de estos planos se aplicó un algoritmo de similitud. En este procedimiento de coloring, el mapa es dividido en 2 o mas colores. La correspondencia de cada neurona a cada color se determina por una relación de métrica mínima. Para efectos de análisis, es posible definir mayores niveles de discretización, pero debe considerare un compromiso entre la agrupación de las neuronas y la capacidad de discriminación de patrones. Para el caso de aplicación, dado que se conocen las etiquetas de los datos, se consideraron tres clusters, ya que las muestras etiquetadas como R y D son semejantes. Por ejemplo, las variables Ca^+ y Mg^+ muestran patrones similares, lo que implica un nivel de correlación entre ellas; por otra parte, los patrones entre pH y K^+ son complementarios, lo que implica una correlación negativa o proporcionalidad inversa.

CLASIFICACIÓN DE MUESTRAS DESCONOCIDAS

Para estudiar las capacidades de clasificación de muestras desconocidas, se tomaron las 9 muestras marcadas CLASS en la tabla 1. Para cada una de las muestras se calculó la BMU del espacio de los vectores prototipos y se compararon las etiquetas (Tabla 2). De las 9 muestras, 6 fueron clasificadas correctamente, 2 correspondían a neuronas sin etiquetar y 1 muestra no fue correctamente clasificada. Los resultados ratifican la importancia del etiquetado en la fase de entrenamiento del mapa. En efecto, si se considerará clasificar muestras comparándolas con un etiquetado único (se asigna la etiqueta con mayor cantidad de coincidencias a la neurona), solamente 3 muestras se considerarían correctamente clasificadas.

La clasificación usando etiquetado por adición o frecuencia permite ubicar las muestras desconocidas en zonas del espacio vectorial de los datos de entrada que no están etiquetadas (muestras 5 y 8) o bien ubicarlas en zonas de transición o mezcla (muestras 3, 4 y 6). Estas muestras a clasificar permitirán aumentar el número de vectores de entrada, mejorando el entrenamiento, y por ende las capacidades de clasificación, ya que las neuronas que no son etiquetadas en el proceso de entrenamiento corresponden a neuronas que se ubican en zonas del espacio que no están representadas por los vectores de entrenamiento.

Tabla 2 Clasificación de las muestras del conjunto CLASS definido en la tabla 2.

Número muestra	Etiqueta	Etiqueta de BMU			
1	S	S	S	S	S
2	S	R	R	R	
3	S	M	S	M	M
4	M	S	S	S	M
5	M	-	-	-	-
6	M	D	M	D	D
7	D	D	D	D	
8	D	-	-	-	-
9	D	D	D		

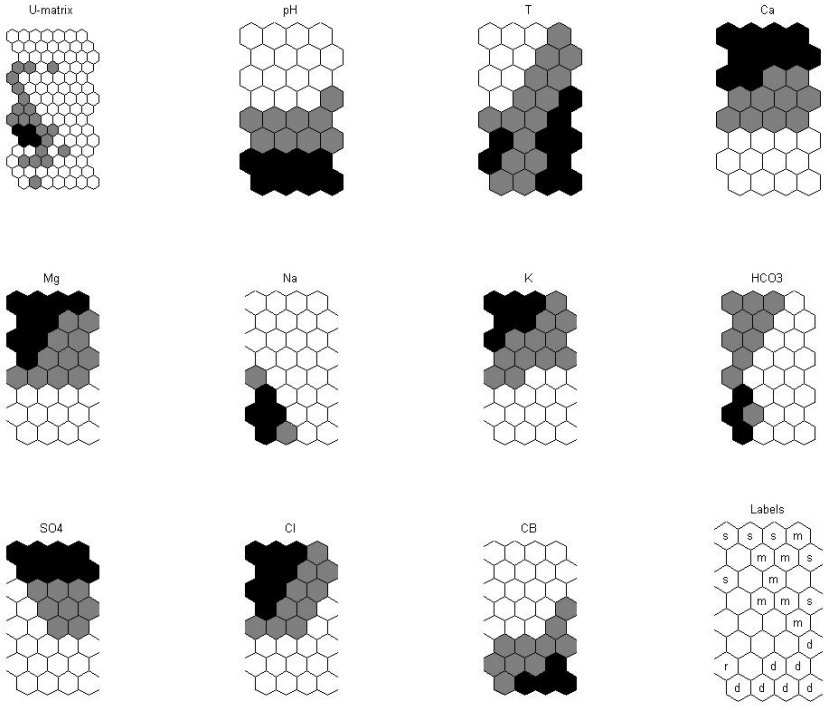

Figura 4 Matriz Unificada de Distancias, planos de componentes y etiquetado para el mapa entrenado con los datos de la Tabla 1.

USO COOPERATIVO CON DIAGRAMAS DE PIPER

Dado que el proceso de clustering en los Diagramas de Piper es visual y subjetivo, la matriz de prototipos $[M]_{32\times10}$ construida a partir de la matriz $[\tilde{R}]_{40\times10}$ (Tabla 1), fue utilizada como matriz de entrada para la construcción de un nuevo mapa, obteniéndose la matriz $[M_2]_{15\times10}$. De esta manera, se disminuyó la cardinalidad de los vectores representativos de los datos originales. Las matrices M, M_2, M_3 fueron de-normalizados y representados en Diagramas de Piper para visualizar los vectores representantes de clusters (Fig. 5). Como se observa, para las muestras etiquetadas S y M, las neuronas que definen este cluster están agrupadas en la parte central del diamante en el Diagrama de Piper, observándose además una clara similitud en los contenidos iónicos de ambas aguas. Por otra parte, las neuronas correspondientes al cluster etiquetado como D graficadas en el Diagrama de Piper, muestran un enriquecimiento en CO_3^{-2} y HCO_3^- y una disminución más marcada en los contenidos de Na^+, SO_4^-, Cl^- y Mg^+.

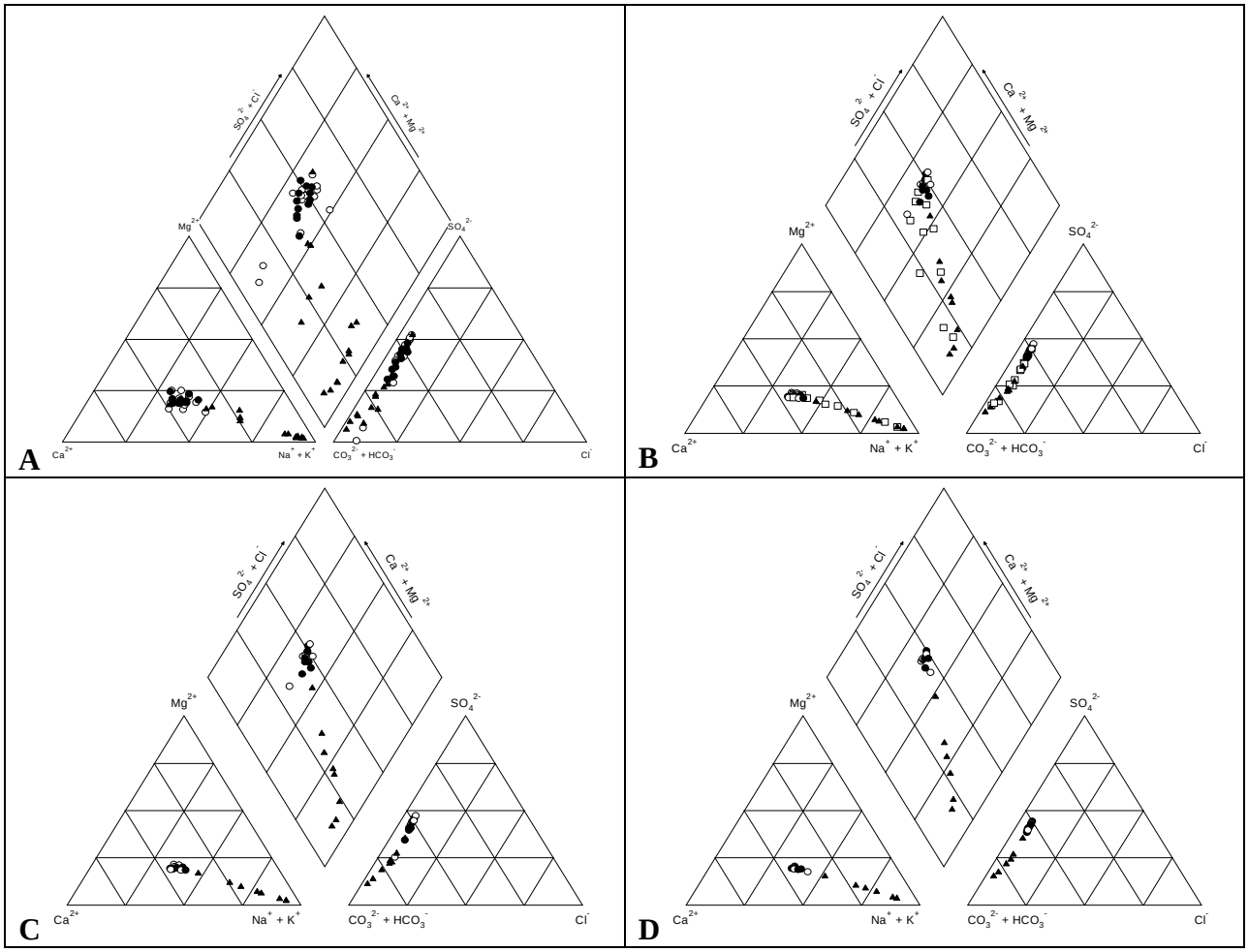

Figura 5 Diagramas de Piper construidos usando: (A) Datos de la tabla 2 (B) Matriz M (c) los vectores prototipos marcados como BMU's, considerando los datos de la tabla 5 como vectores de entrenamiento (D) Matriz M_2. ? Medium wells, ? Shallow wells, ? Deep/Regional wells ? Unlabeled simples.

CONCLUSIONES

El uso conjunto de diagrama de Piper y MAO's genera ventajas en la sistematización y actualización en bases de datos y, principalmente, el aumento en la capacidad de discriminación en muestras semejantes. El uso de los MAO`s cooperativamente con los Diagramas de Piper, reduce la subjetividad en la definición de los grupos

Los algoritmos de clustering como los MAO`s son una herramienta que permite, al igual que los Diagramas de Piper, agrupar muestras de aguas bajo condiciones definidas de similitud. Sin embargo, los Diagramas de Piper no permiten analizar grandes volúmenes de datos, por lo cual es necesario integrar las ventajas y sencillez de este método gráfico con capacidades cuantitativas que complementen el criterio científico. El uso conjunto de los MAO's con los Diagramas de Piper aumentan las capacidades de análisis de muestras de calidad de aguas subterráneas, ya que es posible incluir otras variables distintas a las consideraras en otros métodos gráficos.

La normalización de los datos de entrada para los MAO's, determina la calidad de los mapas resultantes. La elección de la función dependerá de las características de los datos de entrada. El etiquetado mediante adición permite extraer mayor información de las matrices de distancia, ya que además permite observar tendencias en la mezcla o similitud entre muestras .

Las principales ventajas en el uso de los MAO's para el análisis de muestras de aguas subterráneas radica en la capacidad de manejar grandes volúmenes de información, visualizar los resultados y, principalmente, trabajar con los datos en la dimensión original de los datos analizados, bajo las condiciones de la conjetura de naturaleza vectorial de estos datos. Esta última característica disminuye el sesgo asociado al análisis mediante reducción dimensional, como es el caso de o Diagramas de Piper, pero debe verificarse que los resultados de clustering deben ser consistentes con valores de terreno, es decir, tener significancia física (Güler *et al.*, 2002). Los datos medidos en terreno, los fenómenos estudiados y los resultados simulados deben ser comparables y ponderados en su validez, de tal manera que la información contenida sea extraída de manera eficaz y sea posible ajustar hipótesis.

REFERENCIAS

Aguilera, P.; Garrido Frenich, A., Torres, J., Castro, H., Martinez Vidal, J. & Canton, M. (2001) Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Wat. Res. 35(17): 4053–4062.

Chadha, D (1999) A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. Hydrogeology Journal 7:431-439.

Demirel, Z & Güler C. (2006) Hydrogeochemical evolution of groundwater in a mediterranean coastal aquifer, Mersin-Erdemli basin (Turkey). Environmental Geology 49:447-487.

Domenico, P. & Schwartz, F. (2001) Physical and Chemical Hydrogeology. John Wiley & Sons, New York, USA.

Fang, Y; Schwartz, F. & Schincariol, R. (2005) data mining application: self-organizing maps (SOMS) application to assisting ground-water modeling. Geological Society of America *Abstracts with Programs*, 37(7): 28.

Güler, C.; Thyne, G.; Mccray, J. & Turner, A. (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology Journal 10:455-474.

Harvey, F. & Sibray, S. (2001) Delineating ground water recharge from leaking irrigation canals using water chemistry and isotopes. Ground Water 39(3):408-421.

Kaski, S. (1997) Data exploration using Self-Organizing Maps. Ph.D. Thesis, Helsinki University of Technology, Finland.

Leischeid, G. (2003) Investigating the sustability of self-organizing maps to analyze high-dimensional water quality data sets. Geophysical Research Abstracts, 5, #01877.

Leischeid, G. (2005) Non-linear visualization and trend analysis of multivariate data sets using Self-Organizing Maps. Geophysical Research Abstracts,7, #04654, 2005.

Liem T. Tran, L.; Knight, C.; O'Neill, R.; Smith, E. & O'Connell, M. (2003) Self-Organizing Maps for Integrated Environmental Assessment of the Mid-Atlantic Region. Environmental Management 31(6):822–835.

Ochsenkühn, K; Kontoyannakos, J. & Ochsenkühn-Petropulu, M. (1997) A new approach to a hydrochemical study of groundwater flow. Journal of Hydrology 194:64-75.

Peeters, L.; Bação, F; Lobo, V. & Dassargues, A. (2006) Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map. Hydrol. Earth Syst. Sci. Discuss. 3, 1487–1516.

Rivera, D. (2006) Influencia de la interacción agua superficial, subterránea y de riego en el transporte de contaminantes de origen agrícola. Tesis Doctoral, Facultad de Ingeniería Agrícola, Universidad de Concepción.

Sánchez-Martos, F.; Aguilera, P.; Garrido-Frenich, A.; Torres, J. & Pulido-Bosch, A. (2002)Assessment of Groundwater Quality by Means of Self-Organizing Maps: Application in a Semiarid Area. Environmental Management 30(5):716–726.

Thyne, G.; Güler, C. & Poeter, E. (2004) Sequential analysis of hydrochemical data for watershed characterization. Ground Water 42(5):711-723.

Ultsch, A. (2003) Maps for the Visualization of high-dimensional Data Spaces. In *Proceedings Workshop on Self-Organizing Maps (WSOM 2003)*, Kyushu, Japan, 225-230.

Ultsch, A. & Herrmann, L. (2005) The architecture of emergent self-organizing maps to reduce projection errors. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005)* (Ed. by Verleysen, M.), Bruges, Belgium, 1-6.

Vesanto, J. (1997) Data mining techniques based on the Self-Organizing Maps. M.Sc. Thesis, Helsinki University of Technology,Finland.

Vesanto, J. (1999) SOM-based data visualization methods. Intell. Data Anal. 3(2):111-126.

Vesanto, J. (2000) Using SOM in data mining. Licentiate's Thesis, Helsinki University of technology, Finland.

Vesanto, J. (2002) Data exploration process based on the Self-Organizing Maps. Ph.D. Thesis, Helsinki University of Technology, Finland.

Vesanto, J. & Alhoniemi, E. (2000) Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 1(3):586-600.

Zhang, J-Y.; Lu, G-H. Xu, X-M. (2005) Hydrologic regionalization by using self-organizing feature maps neural network. Shuili Xuebao (Journal of Hydraulic Engineering, Chinese Hydraulic Engineering Society) 36(2):163-166, 173. (In Chinese).