UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN CIENCIAS (CINC)

Nuevos diagramas multidimensionales para la clasificación de agua

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

OSCAR ALEJANDRO USCANGA JUNCO

DIRECTORA DE TESIS
Dra. Lorena Díaz González

CO-DIRECTOR DE TESIS
Dr. Surendra Pal Verma Jaiswal †

CUERNAVACA, MORELOS FEBRERO, 2021

DR. JEAN MICHEL GRÉVY MACQUART
COORDINADOR DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada **"Nuevos diagramas multidimensionales para la clasificación de agua"** que presenta el alumno **Oscar Alejandro Uscanga Junco (10033422)** para obtener el título de **Maestro en Ciencias.**

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Outmane Oubram FCQeI-UAEM	APROBADO	
Dr. Jorge Hermosillo Valadez CINC-UAEM	APROBADO	
Dr. Pandarinath Kailasa IER-UNAM	APROBADO	
Dr. Mauricio Rosales Rivera CINC-UAEM	APROBADO	
Dra. Lorena Díaz González CINC-UAEM	APROBADO	

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LORENA DIAZ GONZALEZ | Fecha:2021-05-24 16:12:23 | Firmante

gF2qABjGBTX4cfso/W6G96tZ6qygftDY5MPOu6FFDAFfCvRQqUoTDvqYV7fBSXz2kyxae/Jiz03gUJAzuNW+ZOBJEoWTCG1IjaRG/T3px+vEFITU1Uaje8fInPCfQ4y6ki637G9BPvmS8FqoaK7gGy3/b7mvgbMUprBGAiijUioBH4W351upnNkyulJsCaRTS84MYS8THMb765MUFtYrIdCUpIDm3ImBCZaUdiM9hANYn/7F5JnCIrOEKfGS9TViaphC1GW3PvofzcSBxEV/CcBTI/LgPWxll7RtZYHt1iC8ZRWlf9FGAhWiQC6oopzi8KWeiW51zfs9HM7Hbg3wgQ==

MAURICIO ROSALES RIVERA | Fecha:2021-05-24 20:19:21 | Firmante

dQ6AkAW4tVxO5WM1W34kunkybqNIKpjii3JAHfDCYoZz6O9ijh5JtCdTgsMuFe8Qd//X7CtyktBiz2JoQgupqCyljozp0eEC2ucrq9NWRwXnwpma3oGG8bmhoo1Dk5FF/KgFnR7XDqpaoOsf+psZTro93BhLZi/IAEZSch0PlMjXJJSlyJ7aWbPCM8oaG30ja1kbWahZCfnQTe5kgZkee/+G1WQGSFkfxI2B2wbQUABdlsMdm1Tae4u9JgQ0Fa5J8BG8Ic3yX1F9kOK+YqN6bWbpXS350VvVphD4rWba+rX/A7CQsoEcaDbA4GdUrwsf+nT4cgudk20t6k1j8z0gHg==

PANDARINATH KAILASA - | Fecha:2021-05-24 20:59:04 | Firmante

H/EMQ5MtODYRuDBDgIUlsVKaREvo6IitwxS/Ik7zYfk8ot7pYzVUQh50shnFDsTrP6QTok3sPUj82bCA4OAe5gBRF4uxr0quf3/mPDK1OoEs8X1paEeveq4tNIrZZIXqpeO5ML2F1EVz6UX69Ywpk3BjrH+o/RhxlkDZo2yNKH5T/L2P7Sqsy3+vG/TI8V7NzE8X0BCTE3m4yDc7h8WIycqOLWFXa0aPJiLEX3qeABf1aMATWA7/pjaSyTn21CdWxwKU2VlfruOfEYEalicmDAOTwh5g9gYW1faS1hf42BLbZ9B5RADyFYvwl0P/W7rLLi8GXXE++JniMO/tSfcUrw==

JORGE HERMOSILLO VALADEZ | Fecha:2021-05-25 09:02:59 | Firmante

i0FXPDo6OespXS7XQUeBsIOmRq47PawgHP5jUsJTbh3uEKATvs8PRGjTPWUZCKIVk5h7PqU4ZqLegG9NwaTi2QqHdY5nLhUxec2vVeV1+67RfDV8iCgVqsuv5gEPPOkGoO70Eotj/YIXCZCoP7+TLMYH8LNS8sdsEeIzIg4Phv44/ERNrnjTSRpWOVMdvaeAtbVSLgg8VuzE4PUUIn1/wpMNq1rcgWcQii0GVGim/3q1+y8Upg59uWkjz+om7Oj7QYAlPTyJ7J4rUPwMSlNFfiPbLbB7defLz3oZSzJy2Rnvrqxp9q+TmxDjzCAUTcSpd1bBpM29jWQGoV03NOgS6A==

OUTMANE OUBRAM | Fecha:2021-05-25 10:35:13 | Firmante

UbC9LrUmQY4gcFZ27KmTvGiYxf5FRaDvy3CqM61xPLJ6Z3wObUlvBWkqdwez3U/JV+ke0Db+MC687cpzZpMWbeHtCveHhGKjH84CjSjIJ+lQ7W+PiVlf1MeVrzBj02ztFx7JD+nduOKErPJroShwl0jFrGTDAnZGjYpF7+Ae4bpM6PyIM9+1cNYitJeeMPDX+14R/irBxoEGSEK4cmixSbuQwS8oHw5eLz21i2023lG6jV/8e9xm2BA9qtZYHSYSyYixBa15S/16tfS1FjtQpci9R07rAOjTknQFhQiBfYhdhTuAd25LDP5zzRpwssjfdHJPfLBLRDYN4e0Df6aMRw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

WtxlUe

https://efirma.uaem.mx/noRepudio/Wrc6pkuiqRI5HAJ8bciNEjPboC0HvRCl

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por otórgame una beca que me permitió desempeñarme como estudiante de tiempo completo en este proyecto de maestría.

A la Dirección General de Posgrado del Instituto de Investigación en Ciencias Básicas y Aplicadas (IICBA) de la Universidad Autónoma del Estado de Morelos.

A mi madre Obdulia Junco, mi padre Ismael Uscanga y mi hermana Michelle Uscanga, por ser mi motor fundamental para seguir estudiando y alcanzar nuevas metas en mi vida, por su apoyo, comprensión y cariño. Muchas gracias, gracias infinitas.

A la Dra. Lorena Díaz González, por darme la oportunidad de trabajar con usted, por presentarme a buenas personas, por su mentoría, su apoyo y sus enseñanzas. Gracias por creer en mí.

Al Dr. Surendra Pal Verma Jaiswal, por transmitir su sabiduría a las personas que han tenido el placer de trabajar con usted. Estoy agradecido de haberle conocido.

Al Dr. Mauricio Rosales Rivera, por haberme enseñado machine learning y python, ayudarme a entender cosas fundamentales de la simulación y por su contribución en los manuscritos.

A mis sinodales el Dr. Outmane Oubram y el Dr. José Alberto Hernández Aguilar por su participación en el seguimiento de este trabajo.

A José de Jesús Rodríguez Martínez, por haber compartido parte de tu tiempo para asesorarme en matemáticas.

A Lilia, por escucharme, por tu amor y por acompañarme en este camino. Gracias por compartir tu tiempo conmigo.

Publicaciones relacionadas con esta tesis

I. Verma, S. P., Uscanga-Junco, O. A., & Diáz-González, L. (2021). A statistically coherent robust multidimensional classification scheme for water. Science of the Total Enviroment, 750, 141704. doi:https://doi.org/10.1016/j.scitotenv.2020.141704 **{Apéndice A; Artículo publicado en Science of the Total Environment}**

II. Díaz-González, L., Uscanga-Junco, O. A., Rosales-Rivera, M. (2021) Comparison of machine learning models for water multidimensional classification **{Apéndice B; Artículo publicado en Journal of Hydrology}**

Integrantes del jurado revisor de tesis

Dr. Outmane Oubram FCQeI-UAEM

Dr. Jorge Hermosillo Valadez CINC-UAEM

Dr. Pandarinath Kailasa IER-UNAM

Dr. Mauricio Rosales Rivera CINC-UAEM

Dra. Lorena Díaz González CINC-UAEM

Resumen

El diagrama de Hill-Piper (1944) es la técnica más utilizada para la clasificación de agua, funciona graficando las concentraciones de los iones mayoritarios disueltos, estos son los cationes ($Ca^{2+}, Mg^{2+}, Na^{+} y\ K^{+}$) y los aniones (${SO_4}^{-2}, Cl^{-}, {HCO_3}^{-}\ y\ {CO_3}^{2-}$). A pesar de su extenso uso en muchos estudios hidrológicos se ignoran los problemas fundamentales del uso de diagramas ternarios, los cuales se resumen a continuación: (i) violan la suposición de aleatoriedad de las variables, (ii) violan la suposición básica de que las variables están distribuidas normalmente, (iii) provocan distorsión en los datos graficados, (iv) poseen ambigüedades debido a su geometría.

En esta tesis se proponen dos nuevos modelos para la clasificación de muestras de agua (*7-hlr* y *7-molar-conc*) basados en las transformaciones *hlr* o "hybrid log-ratio transformations" de los 8 iones mayoritarios disueltos en el agua. Estos modelos son un ensamble de clasificadores que implementan análisis discriminante lineal (LDA; *Linear discriminant analysi*s) y análisis canónico, estos permiten la discriminación de las categorías catiónicas y aniónicas.

La base de datos para entrenar los nuevos modelos de clasificación consiste en 50,000 muestras generadas con simulación Monte Carlo para construir datos de distribución uniforme, estas muestras pasaron por un procedimiento de balance de cargas iónico para imitar el equilibrio químico en aguas reales y fueron procesadas a través del software DOMuDaF (Discordant Outlier from Multivariate Data through F-test of w) para encontrar y descartar datos discordantes dejando un total de 46,292 muestras como base de datos de entrenamiento.

Cada muestra de entrenamiento contiene las concentraciones en mMol/L de los 8 iones mayoritarios ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, {SO_4}^{-2}, Cl^{-}, {HCO_3}^{-}\ y\ {CO_3}^{2-}$). El tipo de agua de las muestras se obtiene con el catión y el anión mayoritarios, teniendo un total de 16 tipos de agua básicos. Los modelos logran clasificar estos 16 tipos básicos con una precisión promedio de 92.67% (*7-hlr*) y 97.08% (*7-molar-conc*).

Gracias a la capacidad de los nuevos modelos de desplegar probabilidades de clasificación es posible determinar tipos de agua híbridos, aumentando la cantidad de grupos definibles a 256 tipos de agua hibrida.

Los modelos se pusieron a prueba contra propagación de errores al 40% en cada elemento y simulación de cambios composicionales en el agua, estas pruebas de robustez rebelaron que el modelo *7-hlr* es el modelo más robusto y por lo tanto el recomendado para uso rutinario.

Tabla de contenido

Lista de figuras

Lista de tablas

Estructura de tesis

Capítulo 1: Se expone un panorama general del protocolo de investigación. En este capítulo se define la justificación y los objetivos para proponer los nuevos modelos de clasificación de agua.

Capítulo 2: Se exponen detalladamente los métodos para la clasificación de agua más utilizados, su funcionamiento y los tipos de agua que se pueden clasificar.

Capítulo 3: Este capítulo presenta el funcionamiento de la técnica de discriminación de análisis discriminante lineal y el proceso de simulación Monte Carlo para la creación de las bases de datos.

Capítulo 4: Se describe la metodología general para construir los nuevos modelos de clasificación, se explican la estrategia multiclase de los modelos propuestos y los procesos de la herramienta computacional.

Capítulo 5: Se reportan los resultados obtenidos en esta tesis sobre el rendimiento y pruebas de robustez de los nuevos modelos de clasificación.

Capítulo 6: Este capítulo está dedicado a las conclusiones, trabajos adicionales y futuros.

Capítulo 1: Introducción

La clasificación de la química del agua es una herramienta importante porque aporta información sobre los procesos hidroquímicos en la zona debido a que es un reflejo del medio de contacto y de las interacciones del agua (Kumar, 2013). El agua tiene componentes disueltos; cationes ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$) y aniones ($SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}, CO_3{}^{2-}$). Estos componentes son conocidos como "iones mayoritarios" y son considerados para la clasificación de agua (Piper, 1944). Actualmente, existen distintas metodologías para determinar la clasificación del agua, sin embargo, las más utilizadas están basadas en el diagrama de Hill-Piper (Piper, 1944). Los diagramas más utilizados en orden de popularidad son; (i) Diagrama de Piper con 4472 citas (Piper, 1944); (ii) Diagrama de Chadha con 473 citas (Chadha, 1999); **(**iii) Diagrama de Durov con 346 citas (Durov, 1948). Sin embargo, se ha demostrado por diversos autores (p. ej., Aitchison, 1986; Verma, 2015; Egozcue *et al.*, 2003) que los diagramas ternarios no son una buena opción para graficar las composiciones debido a que presentan problemas graves, los cuales se detallan en la siguiente sección.

1.1 Planteamiento del problema y justificación

El diagrama de Hill-Piper (Piper, 1944) es un procedimiento en el que se grafican las concentraciones porcentuales de los iones mayoritarios ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$ y $CO_3{}^{2-}$; Piper, 1944). Los cationes y aniones se grafican en diagramas ternarios diferentes, estos puntos después se proyectan en un diamante, que es un área dividida en segmentos, dependiendo en donde se proyecte el punto con respecto a sus concentraciones es como se le da una clasificación a esa muestra de agua (Ravikumar *et al.*, 2015). En la figura 1.1 se puede observar un diagrama de Hill-Piper (Pérez-Epinosa *et al.*, 2019).

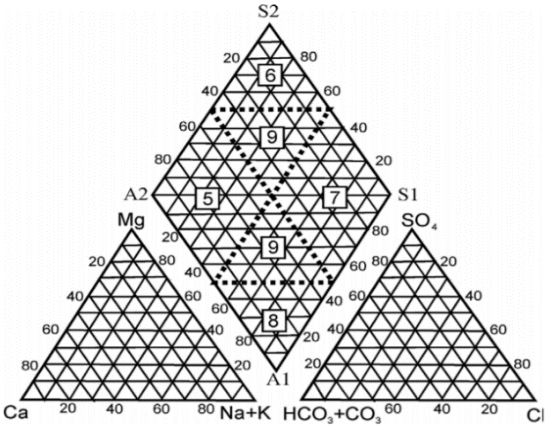


Figura 1.1 Diagrama de Hill-Piper (Pérez-Epinosa et al., 2019)

El diagrama de Hill-Piper (Piper, 1944) es actualmente, la herramienta más utilizada para identificar tipos de agua y cuenta con más 4472 citas de acuerdo con Google Schollar, sin embargo, el uso de diagramas ternarios implica los siguientes problemas:

1.- Violan la suposición de aleatoriedad. El problema de que los datos se presenten en una suma porcentual es que violan la suposición de aleatoriedad al cerrar las variables a una suma de 100%, por lo que, al tener dos variables composicionales, la tercera se determina automáticamente, por lo que esa variable es dependiente de las otras dos (Verma, 2015).

2.- Viola la suposición básica de que las variables deberían estar distribuidas normalmente, ya que, al ser datos porcentuales dentro de un contexto de variable cerrada, los valores negativos están prohibidos. En una distribución normal los datos existen de $-\infty$ a $+\infty$, por lo que, teóricamente, los valores negativos son posibles en una distribución normal (Verma, 2015).

3.- Provocan distorsión en los puntos graficados debido a la forma triangular de los diagramas ternarios (Verma, 2015; Egozcue *et al.*, 2003).

4.- Debido a su geometría, los diagramas ternarios tienen ambigüedades en la diferenciación de las concentraciones de algunos elementos, por ejemplo, $SO_4{}^{-2} y\, Cl^-$ o $Ca^{2+}\ y\ Mg^{2+}$ (Shelton *et al*., 2018).

El uso de transformaciones logarítmicas en los datos composicionales permite que las variables tomen valores tanto negativos como positivos, solucionando el problema de la distribución normal en los datos (Aitchison, 1986). El uso de las transformaciones es favorable debido a que esto permite que métodos estadísticos convencionales puedan aplicarse a los datos (Verma, 2015).

Actualmente, para la clasificación de rocas se han desarrollado otros diagramas multidimensionales que utilizan las transformaciones logarítmicas (Aitchison, 1986). y análisis discriminante lineal (LDA) y está demostrado (p. ej., Verma *et al.*, 2017) que estos diagramas funcionan mejor que los diagramas ternarios para representar estos datos composicionales (Verma, 2015). Esta metodología aún no se ha implementado para la clasificación de agua.

1.2 Objetivos del trabajo de investigación

1.2.1 Objetivo general

- Desarrollar una herramienta computacional para la clasificación de agua usando las transformaciones logarítmicas de *los* cationes $Ca^{2+}, Mg^{2+}, Na^+, K^+$ y los aniones $SO_4{}^{-2}, Cl^-, HCO_3{}^-, CO_3{}^{2-}$ mayoritarios utilizando análisis discriminante lineal (LDA) y análisis canónico.

1.2.2 Objetivos específicos

- Generación de una base de datos de entrenamiento y una base de datos de validación externa mediante simulación Monte Carlo (Verma y Quiroz-Ruiz, 2006).
- Aplicar procedimiento de balance de cargas iónico (BCI) a las muestras simuladas (Nicholson, 1933).
- Calcular las transformaciones logarítmicas *hlr* (Hybrid log-ratio transformations) de las bases de datos.
- Eliminar datos discordantes con la herramienta computacional DOMuDaF (Discordant Outlier from Multivariate Data through F-test of w; Verma *et al.*, 2016).
- Obtener funciones discriminantes mediante análisis discriminante lineal y análisis canónico (Verma *et al.*, 2017).
- Desarrollar una herramienta computacional para la clasificación de aguas en ZK framework utilizando las funciones discriminantes obtenidas del análisis discriminante lineal y análisis canónico.
- Desarrollar un módulo de robustez en la herramienta computacional de ZK framework.

Capítulo 2: Métodos para la clasificación de agua

2.1 Diagrama de Hill-Piper

El diagrama de Hill-Piper propuesto por Piper (1944) utiliza dos diagramas ternarios en donde se expresan las relaciones porcentuales de las concentraciones en unidades de miliequivalentes (mEq/L). En el primer diagrama se expresan los cationes mayores (Ca^{2+}, Mg^{2+} y $Na^{+} + K^{+}$) y en el segundo los aniones mayores (Ca^{2+}, Mg^{2+} y $Na^{+} + K^{+}$) (Kumar, 2013).

Ambos triángulos se proyectan sobre un diamante (figura 2.1), En la versión estándar del diagrama Hill-Piper existen 6 subcampos, aunque pueden existir variantes para definir los subcampos como la utilizada en Ravikumar *et al*. (2015). Al proyectar los datos de ambos diagramas ternarios sobre el romboide, la muestra caerá en alguna zona del diamante según el cation y anion dominante, las distintas combinaciones de elementos dominantes para cada subcampo son las siguientes:

- Subcampo 1: Ca^{2+} y HCO_3^{-} son dominantes; Dureza de carbonos > 50%
- Subcampo 2: Na^{+} y Cl^{-} son dominantes.
- Subcampo 3: Ca^{2+}, Mg^{2+} y Cl^{-} son dominantes.
- Subcampo 4: Ca^{2+}, Na^{+} y HCO_3^{-} son dominantes.
- Subcampo 5: Ca^{2+}y Cl^{-} son dominantes.
- Subcampo 6: Na^{+}y HCO_3^{-} son dominantes.

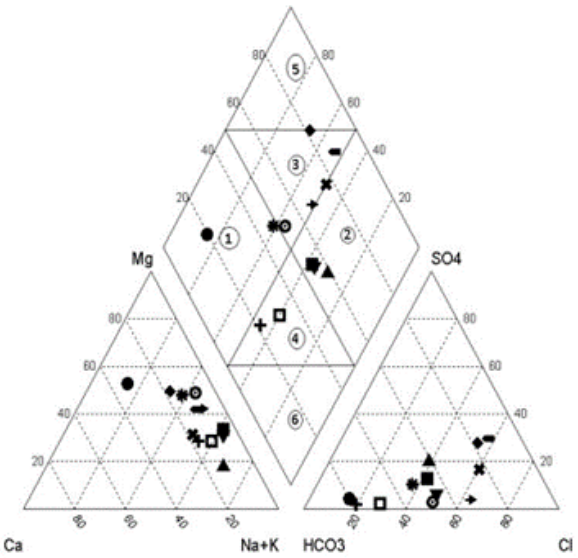


Figura 2.1 Subcampos en el diagrama de Hill-Piper (Kumar 2013)

2.1.1 Diagrama modificado de Hill-Piper

Propuesto por Handa (1965), este diagrama extiende la usabilidad del diagrama de Hill-Piper agregando un segmento para analizar la salinidad, la concentración porcentual de las tierras alcalinas ($Ca + Mg$) se grafica contra la conductividad (sales totales disueltas en 1 meq/L) abriendo la posibilidad de analizar no solo la composición del agua y su salinidad, sino también su viabilidad para riego de las aguas con baja conductividad. (Pérez-Espinosa *et. al* 2019)

Los tipos de agua que se pueden obtener, así como su clasificación en el diagrama de salinidad se pueden observar en la figura 2.2 y son los siguientes (Pérez-Espinosa *et. al* 2019):

- A1: $(Ca \ \& \ Mg) < \ HCO_3$; $(Ca \ \& \ Mg) > (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $< \ HCO_3$; *Residual* $NaHCO_3$; *nill; Hardness: non-carbonate waters*.
- A2: $(Ca \ \& \ Mg) > \ HCO_3$; $(Ca \ \& \ Mg) > (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $> \ HCO_3$; *Residual* $NaHCO_3$*; nill; non-carbonate waters*.
- A3: $(Ca \ \& \ Mg) < \ HCO_3$; $(Ca \ \& \ Mg) < (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $> \ HCO_3$; *Residual* $NaHCO_3$*; nill; Hardness: non-carbonate waters.*
- B1: $(Ca \ \& \ Mg) < \ HCO_3$; $(Ca \ \& \ Mg) > (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $< \ HCO_3$; *Residual* $NaHCO_3$*; present; non-carbonate waters.*
- B1: $(Ca \ \& \ Mg) < \ HCO_3$; $(Ca \ \& \ Mg) < (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $< \ HCO_3$; *Residual* $NaHCO_3$*; present; non-carbonate waters.*
- B1: $(Ca \ \& \ Mg) < \ HCO_3$; $(Ca \ \& \ Mg) < (Na \ \& \ K)$; $(Cl \ \& \ SO_4)$: $> \ HCO_3$; *Residual* $NaHCO_3$*; present; non-carbonate waters.*
- Salinidad: C1, Poca salinidad; C2, Salinidad media; C3, Salinidad media-alta; C4, Salinidad muy alta.
- Niveles de sodio: S1, Poco sodio; S2, Niveles medios de sodio; S3, Niveles altos de sodio; S4, Niveles muy altos de sodio.

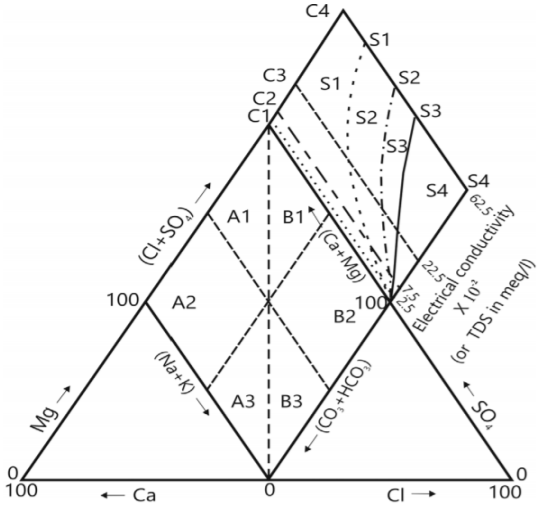


Figura 2.2 Diagrama modificado de Hill-Piper (Pérez-Espinosa et. al 2019).

2.2 Diagrama de Durov

El diagrama de Durov (Durov, 1948) es una variante del diagrama de Hill-Piper (Piper, 1944) que también utiliza dos diagramas ternarios, uno para los cationes mayores (Ca^{2+}, Mg^{2+} y $Na^{+} + K^{+}$) y otro para los aniones ($SO_4{}^{-2}, Cl^{-} y\ HCO_3{}^{-} + CO_3{}^{2-}$), estos iones se expresan en relaciones porcentuales de las concentraciones en mEq/L. La proyección para determinar el tipo de agua se hace sobre los lados de un cuadrado, la base del triángulo de cationes se proyecta en el lado derecho, mientras que la base del triángulo de aniones se proyecta en el lado superior. El cuadrado se divide en los siguientes 9 subcampos o tipos de agua (figura 2.3).

- Subcampo 1: $Ca^{2+}\ y\ HCO_3{}^{-}$ son dominantes.
- Subcampo 2: $Ca^{2+}\ y\ HCO_3{}^{-}$ son dominantes
- Subcampo 3: Na^{+} y $HCO_3{}^{-}$ son cominantes
- Subcampo 4: Ca^{2+} es dominante junto a cualquier anion o Ca^{2+} y $SO_4{}^{-2}$ son dominantes.
- Subcampo 5: No hay ningún cation o anión dominante.
- Subcampo 6: : Na^{+} es dominante junto a cualquier anion o $Na^{+}\ y\ SO_4{}^{-2}$ son cominantes.
- Subcampo 7: Cl^{-} y Ca^{2+} son dominantes.
- Subcampo 8: Cl^{-} junto a cualquier cation o Cl^{-} y Mg^{2+} son dominantes.
- Subcampo 9: Cl^{-} y Na^{+} son dominantes.

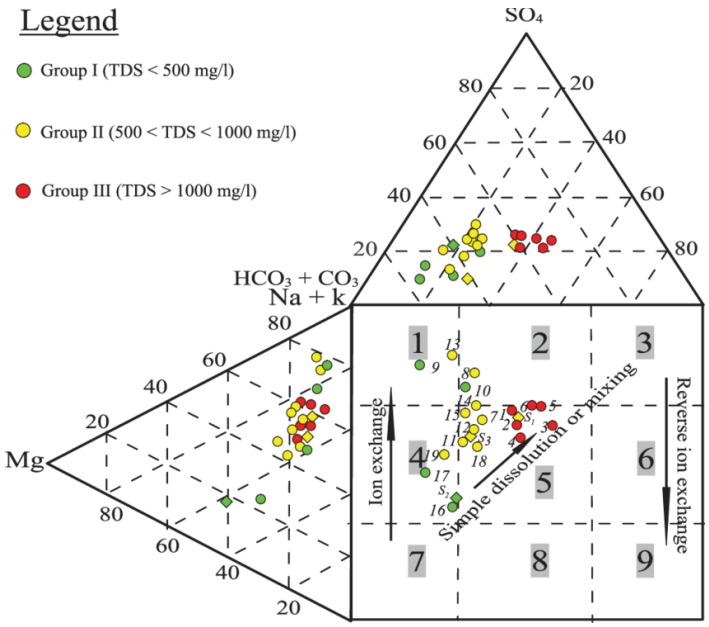


Figura 2.3 Diagrama de Durov (Mustafa et al., 2019)

2.3 Diagrama de Chadha

En el diagrama de Chadha (Chadha, 1999) se omiten los triángulos que se utilizan en los diagramas de Durov y Hill-Piper, en cambio, este diagrama utiliza los ejes X y Y en donde se grafican las diferencias de los porcentajes de los iones mayores en mEq/L. El eje X corresponde a las diferencias entre las tierras alcalinas (Ca^{2+} , Mg^{2+}) y los metales alcalinos (Na^{+}, K^{+}), en tanto que el eje Y a las diferencias entre los aniones débiles ($HCO_3{}^{-}$, $CO_3{}^{2-}$) y aniones fuertes ($SO_4{}^{-2}$, Cl^{-}). Las coordenadas obtenidas se proyectan en subcampos del diagrama para obtener el tipo de agua, dichas coordenadas se calculan con las ecuaciones 2.1 y 2.2.

$$X = (Ca^{2+} + Mg^{2+}) - (Na^{+} + K^{+}) \qquad (2.1)$$

$$Y = (HCO_3{}^{-} + CO_3{}^{2-}) - (SO_4{}^{-2} + Cl^{-}) \qquad (2.2)$$

El área de proyección es rectangular como se muestra en la figura 2.4, este se divide en los siguientes 8 subcampos y cada uno de ellos representa un tipo de agua.

- Subcampo 1: Tierras alcalinas ($Ca^{2+} + Mg^{2+}$) con excedente de metales alcalinos ($Na^{+} + K^{+}$).
- Subcampo 2: Metales alcalinos ($Na^{+} + K^{+}$) con excedente en tierras alcalinas ($Ca^{2+} + Mg^{2+}$).
- Subcampo 3: Aniones débiles ($HCO_3{}^{-} + CO_3{}^{2-}$) con excedente de aniones fuertes ($SO_4{}^{-2} + Cl^{-}$).
- Subcampo 4: Aniones fuertes ($SO_4{}^{-2} + Cl^{-}$) con excedente de aniones débiles ($HCO_3{}^{-} + CO_3{}^{2-}$).
- Subcampo 5: Tierras alcalinas y aniones débiles ($Ca^{2+} + Mg^{2+}$ y $HCO_3{}^{-} + CO_3{}^{2-}$) con excedentes de metales alcalinos y aniones fuertes ($Na^{+} + K^{+}$ y $SO_4{}^{-2} + Cl^{-}$).
- Subcampo 6: Tierras alcalinas y aniones fuertes ($Ca^{2+} + Mg^{2+}$ y $SO_4{}^{-2} + Cl^{-}$) con excedente de metales alcalinos y aniones débiles ($Na^{+} + K^{+}$ y $HCO_3{}^{-} + CO_3{}^{2-}$).
- Subcampo 7: Metales alcalinos y aniones fuertes ($Na^{+} + K^{+}$y $SO_4{}^{-2} + Cl^{-}$) con excedente de tierras alcalinas y aniones débiles ($Ca^{2+} + Mg^{2+}$y $HCO_3{}^{-} + CO_3{}^{2-}$).
- Subcampo 8: Metales alcalinos y aniones débiles ($Na^{+} + K^{+}$ y $HCO_3{}^{-} + CO_3{}^{2-}$) con excedente de tierras alcalinas y aniones fuertes ($Ca^{2+} + Mg^{2+}$ y $SO_4{}^{-2} + Cl^{-}$).

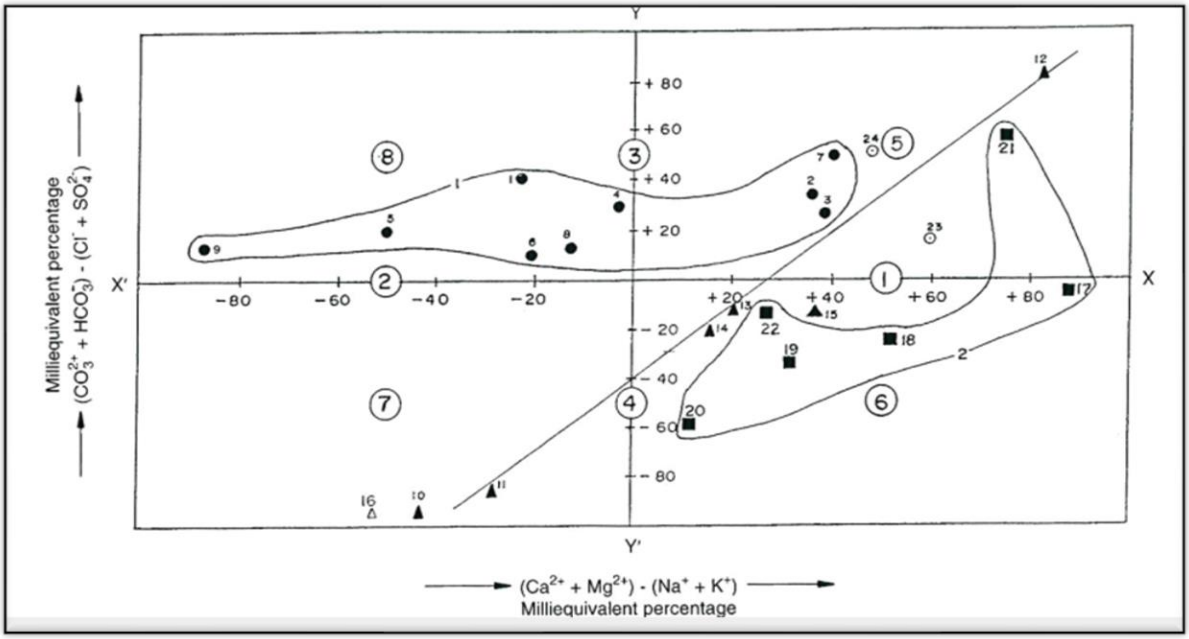


Figura 2.4 Diagrama de Chadha (Chadha, 1999)

2.4 Diagrama de Banaga

El Diagrama de Banaga propuesto por Elhag (2017) se construye con los porcentajes de las concentraciones en mEq/L de los cationes Ca^{2+}, Mg^{2+}, Na^{+} y K^{+} y los aniones ${SO_4}^{-2}, Cl^{-}, {HCO_3}^{-}$ y ${CO_3}^{2-}$. Este diagrama se construye proyectando las concentraciones porcentuales en un diamante que contrapone las tierras alcalinas (Ca^{2+}, Mg^{2+}) con los metales alcalinos (Na^{+}, K^{+}) en lados opuestos del diamante. Así también los aniones débiles (${HCO_3}^{-}, {CO_3}^{2-}$) y los aniones fuertes (${SO_4}^{-2}, Cl^{-}$) se contraponen en los lados opuestos. Este diagrama permite graficar las 8 variables en un mismo diagrama, sin embargo, se pierde la visibilidad independiente de las variables que se encuentran sumadas, por ejemplo, no se expresa si Ca^{2+} es mayor a Mg^{2+}. El diamante se divide en los siguientes subcampos (figura 2.5):

- Subcampo 1: Tierras alcalinas con aniones débiles (*Alkaline bicarbonate carbonate*).
- Subcampo 2: Tierras alcalinas con aniones fuertes (*Alkaline sulphate chloride*).
- Subcampo 3: Metales alcalinos con aniones fuertes (*Alkali sulphate chloride*).
- Subcampo 4: Metales alcalinos con aniones débiles (*Alkali bicarbonate carbonate*).
- Subcampo 5: No hay un par que se exceda ni en cationes ni en aniones.

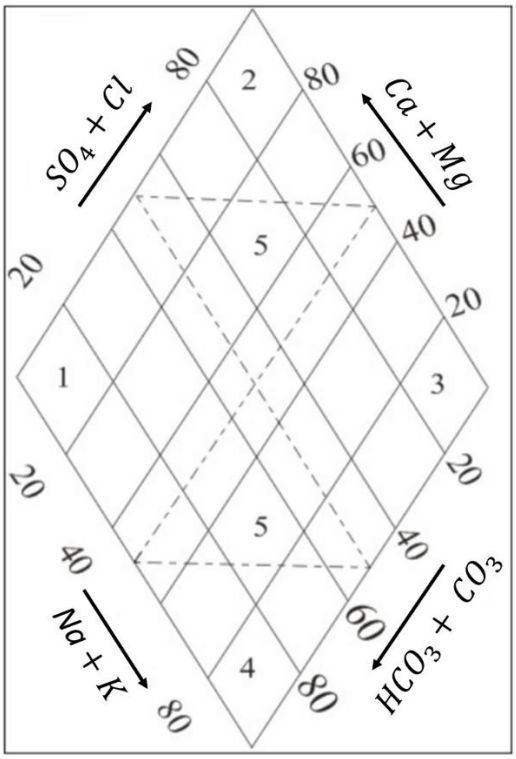


Figura 2.5 Diagrama de Banaga (Elhag, 2017)

2.5 Graficado de transformaciones logarítmicas

En el trabajo propuesto por Shelton *et al.* (2018) se explora una alternativa basada en transformaciones logarítmicas de los datos composicionales del agua, esta propuesta está basada en el diagrama propuesto por Piper (1944). Las ecuaciones 2.3 a 2.6 permiten calcular las 4 coordenadas *ilr* (*isometric log-ratio transformation*) que maximizan la interpretación geoquímica de las concentraciones (en mEq/L) de los cationes y aniones mayoritarios,

$$z_1 = \sqrt{\frac{2}{3}} \ln\left(\frac{\sqrt{Ca^{2+} * Mg^{2+}}}{Na^{+} + K^{+}}\right) \quad (2.3)$$

$$z_2 = \sqrt{\frac{1}{2}} \ln\left(\frac{Ca^{2+}}{Mg^{2+}}\right) \quad (2.4)$$

$$z_3 = \sqrt{\frac{2}{3}} \ln\left(\frac{\sqrt{{SO_4}^{-2} * Cl^{-}}}{{HCO_3}^{-} + {CO_3}^{2-}}\right) \quad (2.5)$$

$$z_4 = \sqrt{\frac{1}{2}} \ln\left(\frac{Cl^{-}}{{SO_4}^{-2}}\right) \quad (2.6)$$

En la figura 2.6 se puede observar un gráfico de transformaciones *ilr*, el gráfico consta de 4 paneles que sustituyen los componentes geométricos del diagrama de Hill-Piper:

- Inferior izquierda, que sustituye el diagrama ternario de cationes, proyecta las coordenadas z_1 y z_2; en donde z_1 expresa la proporción de $Ca^{2+} y\, Mg^{2+}$ contra $Na^{+} y\ K^{+}$, y z_2 expresa la proporción de Ca^{2+} contra Mg^{2+} , es decir, la proporción entre las concentraciones de los cationes.
- Superior derecha, que sustituye el diagrama ternario de aniones, proyecta las coordenadas z_3 y z_4; en donde z_3 expresa la proporción de ${SO_4}^{-2} y\, Cl^{-}$ contra ${HCO_3}^{-} y\, {CO_3}^{2-}$, y z_4 expresa la proporción de Cl^{-} contra ${SO_4}^{-2}$. Por lo que este panel expresa la proporción entre las concentraciones de los aniones.
- Inferior derecha, que sustituye el diamante, proyecta las coordenadas z_3 (${SO_4}^{-2} y\, Cl^{-}$ contra ${HCO_3}^{-} y\, {CO_3}^{2-}$) y z_1 ($Ca^{2+} y\, Mg^{2+}$ contra $Na^{+} y\ K^{+}$) como ejes X y Y respectivamente, en donde los valores positivos en el eje Y indican una mayor concentración en $Ca^{2+} y\, Mg^{2+}$ y los negativos una mayor concentración en $Na^{+} y\ K^{+}$. En el eje X los valores positivos significan una mayor concentración en ${SO_4}^{-2}\ y\ Cl^{-}$ y los negativos en ${HCO_3}^{-} y\, {CO_3}^{2-}$.
- Superior izquierda, es un panel que añade información y que no tiene un equivalente en el diagrama de Hill-Piper, proyecta las coordenadas z_2 (Ca^{2+} contra Mg^{2+}) y z_4 (Cl^{-} contra ${SO_4}^{-2}$) siendo los ejes X y Y respectivamente. Este panel sirve para diferenciar entre las concentraciones de esos 4 iones, pero no puede dar clasificación a una muestra de agua por si solo ya que no tiene la información de los demás componentes.

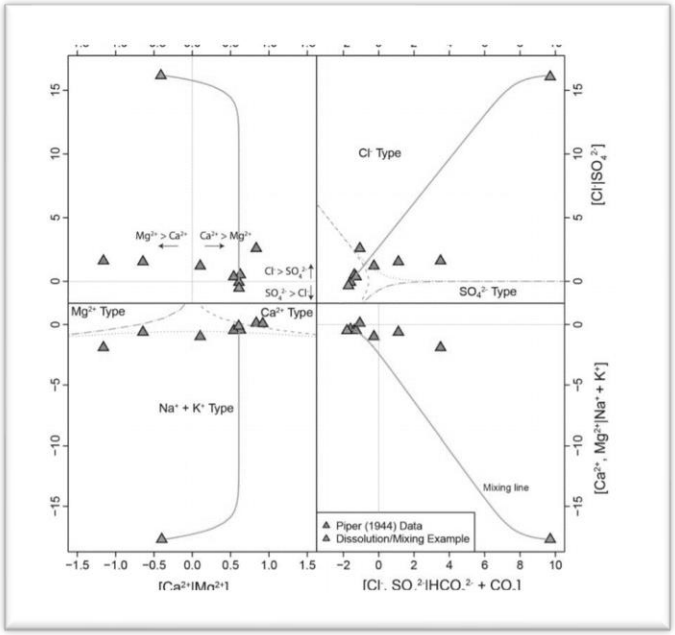


Figura 2.6 Ejemplo de diagrama con transformaciones log. (Shelton et al., 2018)

2.6 Diagrama ternario de Giggenbach

El diagrama de Giggenbach (Giggenbach, 1988) es, actualmente la herramienta más utilizada para el estudio de sistemas hidrotermales. Es un diagrama ternario que está construido con las concentraciones de *Na*, *K* y *Mg* expresadas en unidades de mg/kg, el triángulo es la combinación de pares de variables que son viables para su uso como geotermómetros (*Na-K* y *K-Mg*). Las variables del diagrama ternario se expresan de la siguiente forma: Na/1000, K/100 y $\sqrt{Mg}$. (Romano y Liotta, 2020)

Giggenbach (1988) propone dos ecuaciones empíricas para describir la dependencia de las concentraciones de los elementos con la temperatura (ecuaciones 2.7 y 2.8).

$$T(°C) = \{1390/[1.75 + \log(C_{Na}/C_K)]\} - 273.15 \qquad (2.7)$$

$$T(°C) = \{4410/[14.0 - + \log(C^2{}_K/C_{Mg})]\} - 273.15 \qquad (2.8)$$

En este diagrama ternario (figura 2.7) las condiciones de equilibrio se definen con fronteras representativas relacionadas a los geotermómetros de *Na-K* y *K-Mg*, dichas líneas se definen por el índice de madurez (MI; *Maturity Index*), que se calcula combinando las ecuaciones de los geotermómetros para obtener:

$$\log(C_k/C_{Na}) = 0.315\ \log(C^2{}_K/C_{Mg}) - 2.66 = 0.315 \log(C^2{}_K/C_{Mg}) - MI \qquad (2.9)$$

En donde un valor de *MI* = 2.66 corresponde a aguas en condiciones de equilibrio, Giggenbach (1988) propone un *MI* de 2.0 para definir la frontera entre aguas parcialmente equilibradas y aguas inmaduras (no equilibradas). (Romano y Liotta, 2020)

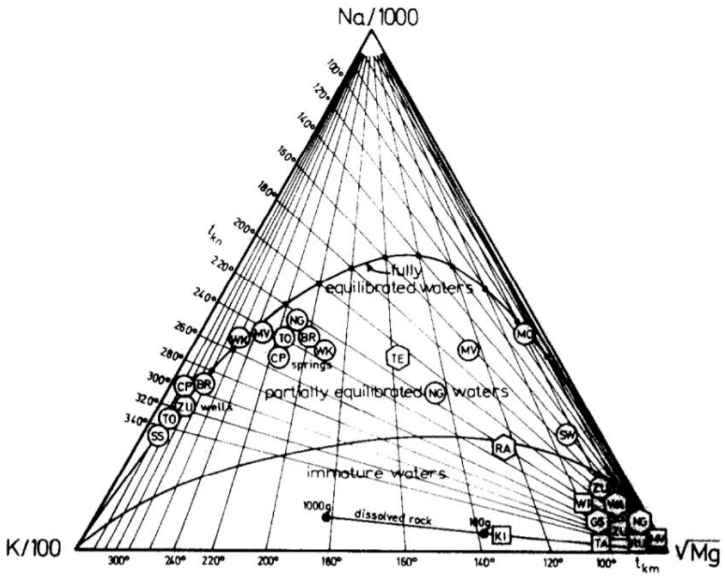


Figura 2.7 Diagrama de Giggenbach

2.7 Otras metodologías de clasificación de agua

Perceptrón multicapa

En el trabajo de Bayram y Gultekin (2010) se propuso una clasificación de aguas geotermales implementando redes neuronales artificiales, específicamente un modelo de perceptrón multicapa. Este modelo de clasificación consiste en la diferenciación de muestras de agua en 4 categorías (cold, eynal, citgol y nasa). Sin embargo, estas categorías no son una nomenclatura general, ya que atienden a la necesidad particular de clasificación en una zona de estudio.

Diagrama de Stiff

El diagrama de Stiff fue propuesto por Stiff (1951) y consiste en un polígono que se crea por una línea vertical que corta por en medio varias líneas horizontales, en las líneas horizontales de la izquierda se pone la concentración de los cationes y a la derecha los aniones, ambos en mEq/L (Güler *et al.*, 2002). En la figura 2.8 se pueden ver ejemplos de diagramas de Stiff. En este diagrama los grupos de muestras de agua que pertenecen al mismo tipo se asocian con respecto a la forma del polígono.

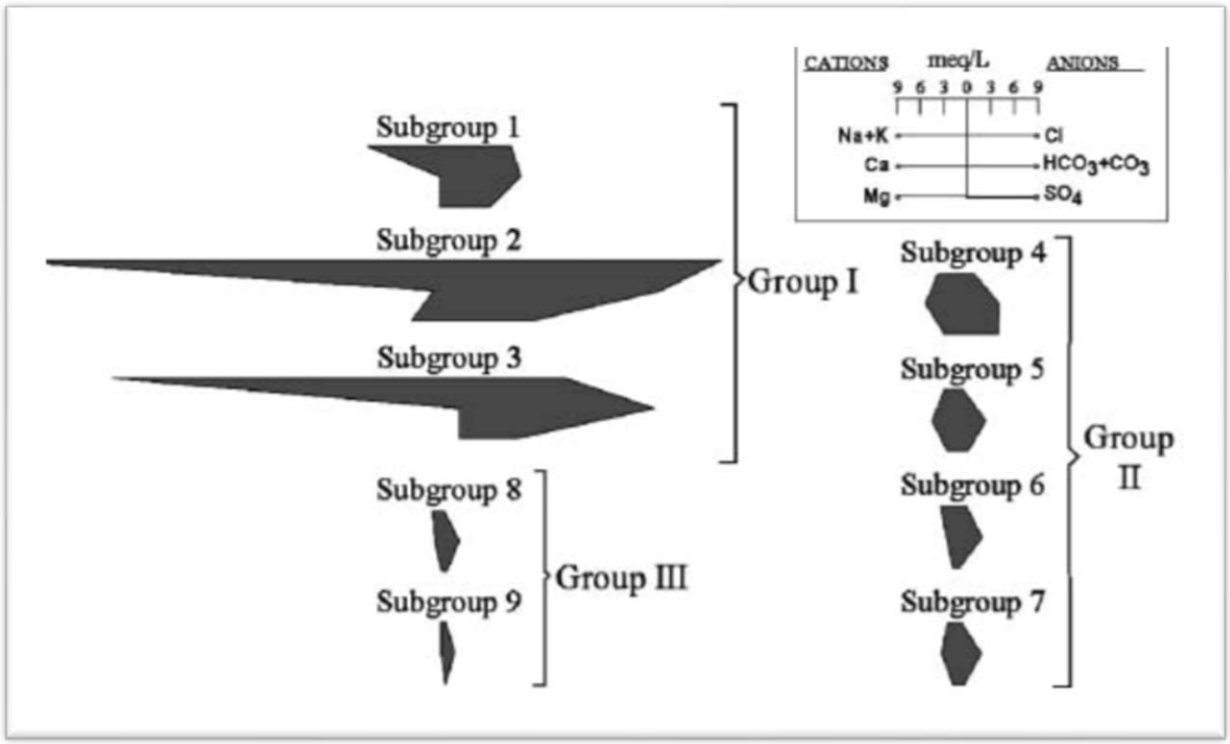


Figura 2.8 Diagramas de Stiff (Güler et al., 2002)

Capítulo 3: Fundamentos teóricos de simulación y análisis discriminante

3.1 Análisis discriminante lineal

El análisis discriminante lineal (LDA) es una técnica de clasificación y de reducción de dimensiones. El objetivo de esta técnica es transformar las características de "muestras" o "vectores" en un espacio de menos dimensiones, esta transformación maximiza la varianza entre clases y minimiza la variabilidad interna de los vectores que componen cada clase (varianza de clase) (Tharwat *et al.* 2017).

La matriz de muestras se denota como: $X = \{X_1, X_2, X_3, X_4, \dots, X_N\}$ en donde X_i es la iésima muestra y N el número total de muestras. Cada muestra individual está representada por M dimensiones ($X_i \in R^M$). Se asume que la matriz X esta particionada en c clases $X = [w_1, w_2, w_3, w_4, \dots, w_c]$, siento w cada subconjunto con las muestras que corresponden a cada clase.

El algoritmo para realizar LDA se compone de los siguientes pasos (Tharwat *et al*. 2017):

i) Calcular la media de cada clase $\bar{X}_i$ con la ecuación 3.1.

$$\bar{X}_i = \frac{1}{n_i} \; X_i \in w_j \qquad (3.1)$$

En donde n_j es el número de elementos en la clase j y w_j las muestras de la clase j.

ii) Calcular la media total $\bar{X}$ con la ecuación 3.2.

$$\bar{X} = \frac{1}{N} \sum_1^N X_i \qquad (3.2)$$

En donde N es el número total de muestras.

iii) Calcular la matriz de varianza entre clases ("*between-class matrix*"), denotada por S_B

$$S_B = \sum_{\mathbf{1}}^{\mathbf{c}} \boldsymbol{n_i} \, (\bar{X}_i - \bar{X} \,)(\bar{X}_i - \bar{X})^{\boldsymbol{T}} \qquad (3.3)$$

En donde c es el número de clases, n_i es el número de muestras en cada clase.

iv) Calcular la matriz de varianza interna de clase ("*within-class matrix*") denotada por S_w para cada una de las clases, esta matriz se calcula con la ecuación 3.4.

$$S_{wj} = \sum_{X_i \in w_j} (X_i - \bar{X}_j)(X_i - \bar{X}_j)^T \qquad (3.4)$$

En donde X_i es cada muestra de la clase j y $\bar{X}_j$ es la media de dicha clase. Al final se tienen tantas S_w como clases.

v) Calcular la matriz de transformación denotada por $\boldsymbol{W_i}$ con la ecuación 3.5

$$\boldsymbol{W_i} = {S_{wi}}^{-1} S_B \qquad (3.5)$$

En donde S_{wi} es la matriz de varianza interna de la clase i y S_B la matriz de varianza entre clases.

vi) Calcular los valores propios λ_i y los vectores propios V_i de cada matriz de transformación W_i, y ordenarlos con respecto a los valores propios λ_i. Cada eigen-vector representa un eje en el espacio transformado de LDA y su valor propio corresponde a la capacidad del eigen-vector para clasificar entre las clases (Tharwat *et al. 2017).*

Los primeros k vectores propios con el valor λ_i más alto son seleccionados para formar el nuevo espacio reducido a k dimensiones V_k^i, estos vectores propios representan las direcciones del nuevo espacio.

vii) Proyectar las muestras de cada clase en su nuevo espacio k-dimensional con la ecuación 3.6.

$$\phi_j = X_i V_k^j, \ \ x_i \epsilon\, w_j \qquad (3.6)$$

En donde ϕ_j se refiere al espacio transformado de las muestras X_i que pertenecen a la clase w_j. Este espacio transformado ϕ_j contiene las muestras transformadas de tal forma que se maximiza la distancia entre las clases y la distancia interna entre la media de cada clase con sus muestras. Esta técnica de discriminación asume que las clases tienen una distribución multinormal, en la sección 4.4 se muestra el procedimiento para poder cumplir este requisito. La implementación de LDA utilizada en este trabajo es del software Statistica ®.

3.2 Procedimiento para la generación de la base de datos de entrenamiento

El primer objetivo de este proyecto, es generar una base de datos simulada en la que se encuentren representadas las 16 categorías de tipos de agua determinadas por los iones mayoritarios, las clases son las siguientes; (i) *calcium-sulfate*, (ii) *calcium-chloride*, (iii) *calcium-bicarbonate*, (iv) *calcium-carbonate*, (v) *magnesium-sulfate*, (vi) *magnesium-chloride*, (vii) *magnesium-bicarbonate*, (viii) *magnesium-carbonate*, (ix) *sodium-sulfate*, (x) *sodium-chloride*, (xi) *sodium-bicarbonate*, (xii) *sodium-carbonate*, (xiii) *potassium-sulfate*, (xiv) *potassium-chloride*, (xv) *potassium-bicarbonate*, (xvi) *potassium-carbonate*.

Los 16 grupos se forman de la generación de 8 variables que corresponden a las concentraciones en mMol/L de los 8 iones mayoritarios ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$, $SO_4{}^{-2}, Cl^{-}$ $HCO_3{}^{-}$ y $CO_3{}^{2-}$). Se implementaron distintas metodologías para lograr este objetivo y en esta sección se abordan las que contribuyeron al entendimiento y resolución final de este problema.

3.2.1 Generación de muestras aleatorias con distribución normal

La primera aproximación a la generación de los datos recurriendo a la simulación fue en base a la metodología de Verma y Quiroz-Ruiz (2006), los pasos de dicha metodología son los siguientes:

i) Generar números pseudoaleatorios uniformemente distribuidos en el espacio (0, 1) para esto se utiliza el algoritmo Marsenne Twister de Matsumoto y Nishimura (1998) para generar los números U (0,1).

ii) Comprobar que los números pseudoaleatorios de la distribución uniforme U (0,1) corresponden a una distribución IID U(0,1); es decir que están independientemente e idénticamente distribuidos. Esto se demostró en el trabajo de Verma y Quiroz-Ruiz (2006) utilizando el método de Marsaglia (1968) graficando los números generados en dos y tres dimensiones para observar la correcta distribución uniforme de los valores generados por el algoritmo Marsenne Twister de Matsumoto y Nishimura (1998).

iii) Convertir los números pseudoaleatorios en variables continuas de una distribución normal N(0,1). En el procedimiento de Verma y Quiroz-Ruiz (2006) se usó el método polar de Marsaglia y Bray (1964). pero en una modificación posterior del procedimiento de Monte Carlo en el trabajo de Verma *et al*. (2017) el método polar se cambió por el método Ziggurat propuesto por Doornik (2005). El método Ziggurat demostró ser más eficiente para la generación de una distribución N (0,1) en el trabajo de Thomas *et al.* (2007) en donde se hizo una comparación con el método polar (Marsaglia y Bray, 1964).

Con dicho conocimiento se desarrolló una metodología de simulación que buscaría la representatividad de las 16 clases que ocurren por combinación cruzada entre cada elemento mayoritario de los cationes (Ca^{2+}, Mg^{2+},Na^{+} y K^{+}) y los aniones ($SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$ y $CO_3{}^{2-}$). La metodología para la generación de las muestras de las 16 clases representadas fue la siguiente:

i) Determinar parámetros de la simulación.

Determinar una media y una desviación estándar para una variable predominante y una variable no predominante, es decir que, en el grupo de variables catiónicas (Ca^{2+}, Mg^{2+} , Na^{+}, K^{+}), para garantizar que después de una simulación todos los datos correspondieran por ejemplo, a la clase *calcium* (Ca^{2+}) dicha variable debería tener un valor más alto con respecto a las otras 3 variables, para que al final, todos los elementos resultantes de la simulación fuesen clasificados como *calcium* en la parte catiónica. La misma lógica se aplica para la parte aniónica.

Los parámetros seleccionados para garantizar que todas las muestras pertenecieran a una clase son: N (135.7460, 20.3116) para la variable predominante y N (14.7513, 7.7295) para las variables no predominantes. La simulación se hizo considerando que los valores de las variables representarían la concentración en mM/L porque es en esa unidad en donde se busca la representatividad de las clases.

ii) Determinar el número de simulaciones

Para obtener valores representativos de todas las clases en combinación cruzada de las variables catiónicas y aniónicas fue necesario hacer 16 simulaciones, en cada simulación se consideraron las 8 variables y se dejó una variable catiónica predominante y una variable aniónica predominante para garantizar que el resultado de la simulación perteneciera, por ejemplo, a *calcium-sulfate*.

En la tabla 3.1 se pueden observar los parámetros de la simulación para generar los valores de las concentraciones en mM/L de todas las variables, dichos parámetros fueron utilizados para la generación de las muestras de la clase *calcium-sulfate*.

Tabla 3.1 Media y desviación estándar de la clase calcium-sulfate

Clase		**Ca^{2+}**	**Mg^{2+}**	**Na^{+}**	**K^{+}**	**$SO_4{}^{-2}$**	**Cl^{-}**	**$HCO_3{}^{-}$**	**$CO_3{}^{2-}$**
calcium-sulfate	$\bar{x}$	135.74	14.75	14.75	14.75	135.74	14.75	14.75	14.75
	s	20.31	7.72	7.72	7.72	20.31	7.72	7.72	7.72

De manera que, utilizando la lógica anterior, se harían 16 simulaciones en donde cada variable catiónica y cada variable aniónica en combinación fueran el par de variables predominantes, el número de muestras en cada clase se presenta en la tabla 3.2.

Tabla 3.2 Componentes predominante para cada clase

Clase	**Componentes predominantes**	
	Cationes	**Aniones**
calcium-sulfate	Ca^{2+}	$SO_4{}^{-2}$
calcium-chloride	Ca^{2+}	Cl^{-}
calcium-bicarbonate	Ca^{2+}	$HCO_3{}^{-}$
calcium-carbonate	Ca^{2+}	$CO_3{}^{2-}$
magnesium-sulfate	Mg^{2+}	$SO_4{}^{-2}$
magnesium-chloride	Mg^{2+}	Cl^{-}
magnesium-bicarbonate	Mg^{2+}	$HCO_3{}^{-}$
magnesium-carbonate	Mg^{2+}	$CO_3{}^{2-}$
sodium-sulfate	Na^{+}	$SO_4{}^{-2}$
sodium-chloride	Na^{+}	Cl^{-}
sodium-bicarbonate	Na^{+}	$HCO_3{}^{-}$
sodium-carbonate	Na^{+}	$CO_3{}^{2-}$
potassium-sulfate	K^{+}	$SO_4{}^{-2}$
potassium-chloride	K^{+}	Cl^{-}
potassium-bicarbonate	K^{+}	$HCO_3{}^{-}$
potassium-carbonate	K^{+}	$CO_3{}^{2-}$

En esta metodología se simularon un total de 3015 muestras por cada una de las 16 clases. Sin embargo, resultó no ser el método óptimo después de hacer un análisis discriminante lineal. La técnica de discriminación se tiene que aplicar para clasificar la parte catiónica y un análisis independiente para identificar la parte aniónica. Considerando que se hace el análisis discriminante lineal con 3 clases utilizado en Agrawal *et al*. (2004) se hizo una prueba para clasificar tres grupos aniónicos, en este caso, *sulfate*, *bicarbonate y carbonate*.

Se graficaron las muestras utilizando las funciones discriminantes DF1 y DF2. En la figura 3.1 se puede apreciar que las 3 nubes de datos de *sulfate* (verde), *bicarbonate* (azul) y *carbonate* (rojo) están muy alejadas unas de otras, es decir que, no hay suficientes datos en las zonas en donde ocurre un cambio de clase. Esto indica que las clases no se encuentran correctamente representadas.

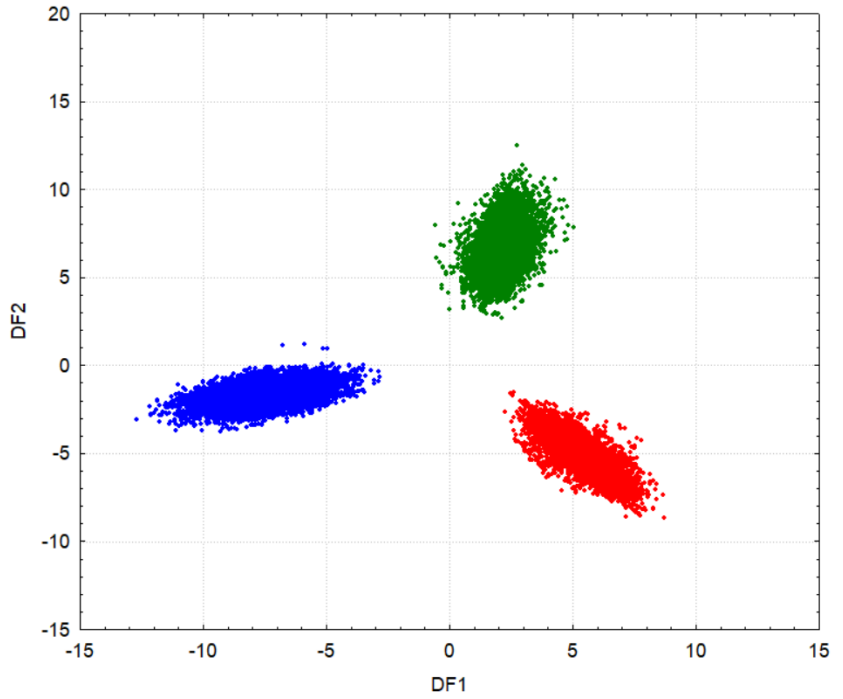


Figura 3.1 Grafico con funciones discriminantes DF1 y DF2 de simulación normal, Verde: sulfate, Azul: bicarbonate, Rojo: carbonate.

3.2.2 Generación de muestras aleatorias con distribución uniforme

De acuerdo con el primer paso del procedimiento de Verma y Quiroz-Ruiz (2006), es necesario generar números pseudoaleatorios uniformemente distribuidos U (0,1) antes de poder generar los números con distribución normal. Solo con este paso se decidió abordar otra perspectiva para encontrar la representatividad de las 16 clases, generando los valores de 8 variables en donde cada variable corresponde a una distribución uniforme U (0,1).

En total se generaron 50,000 muestras (cada muestra se constituye de las 8 variables). Los números pseudoaleatorios de U (0,1) se multiplicaron por un escalar c=100 para aumentar el rango máximo de los valores, este número fue seleccionado arbitrariamente para cubrir la representatividad de las variables en unidades de mMol/L.

Siguiendo el procedimiento para generar los diagramas discriminantes propuesto por Agrawal *et al*. (2004) se realizó un análisis discriminante lineal de 3 clases; *calcium-sulfate*, *calcium-chloride* y *calcium-carbonate*. En la figura 3.2 se aprecia que la representatividad es mucho más clara y las fronteras entre las clases están mejor definidas cuando las variables siguen una distribución uniforme y cuando se hizo una sola simulación de un número definido de muestras.

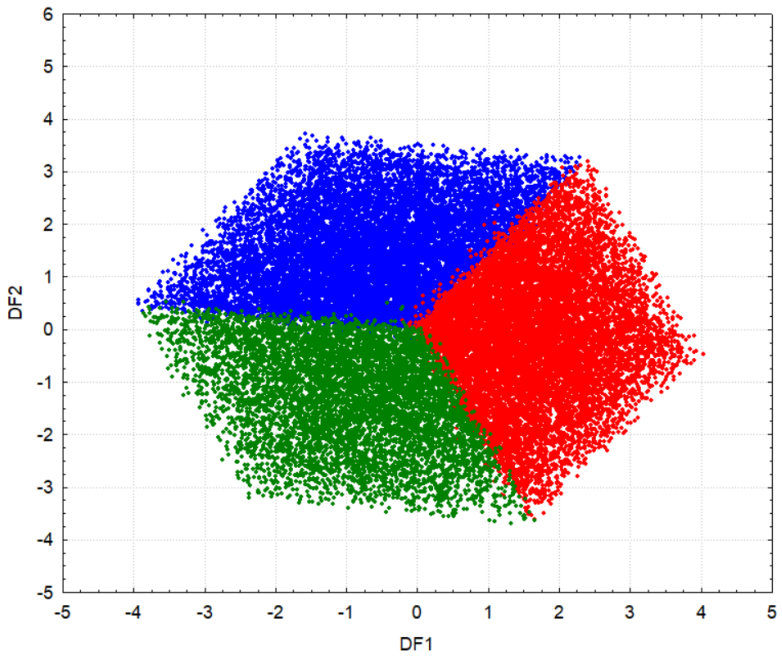


Figura 3.2 Grafico con funciones discriminantes DF1 y DF2 de simulación uniforme, Verde: sulfate, Azul: bicarbonate, Rojo: carbonate.

La tabla 3.3 muestra cuantas muestras de las 50,000 corresponden a cada una de las 16 clases, esto es por el criterio del catión y anión mayoritarios. Al aplicar la metodología de Verma y Quiroz-Ruiz (2006) utilizando el algoritmo propuesto por Matsumoto y Nishimura (1998), se logró obtener una buena representatividad para generar los diagramas con análisis discriminante lineal, además, de mostrar de forma explícita en la figura 3.2 la definición de las fronteras. Esta simulación fue la que se seleccionó para continuar con los distintos pasos de la metodología.

Tabla 3.3 Numero de muestras por clase

Clase	**Número de muestras**
calcium-sulfate	3102
calcium-chloride	3033
calcium-bicarbonate	3131
calcium-carbonate	3031
magnesium-sulfate	3120
magnesium-chloride	3143
magnesium-bicarbonate	3161
magnesium-carbonate	3160
sodium-sulfate	3082
sodium-chloride	3120

sodium-bicarbonate	3148
sodium-carbonate	3134
potassium-sulfate	3147
potassium-chloride	3128
potassium-bicarbonate	3113
potassium-carbonate	3247
Total	50000

3.2.3 Generación de números pseudoaleatorios de suma constante con distribución uniforme

Esta sección explica uno de los procedimientos que se consideró para generar la base de datos inicial, el procedimiento no fue el seleccionado debido a la versatilidad y facilidad de uso del método expuesto en la sección 3.3.2. La primera aproximación a al desarrollo de un nuevo diagrama para la clasificación de agua fue recurriendo a una traducción literal del diagrama de Hill-Piper (Piper, 1944) en donde se intentaba llenar las posibilidades de clasificación mediante simulación, pero intentando encontrar la representatividad en el diagrama ternario.

Usaré como ejemplo 3 variables de la simulación expuesta en la sección 3.3.2 para exponer este punto. La variable *calcium*, que tiene una distribución U (0, 100), corresponde al histograma de la figura de la figura 3.3. La frecuencia con la que aparece cada rango de valores es similar (distribución uniforme).

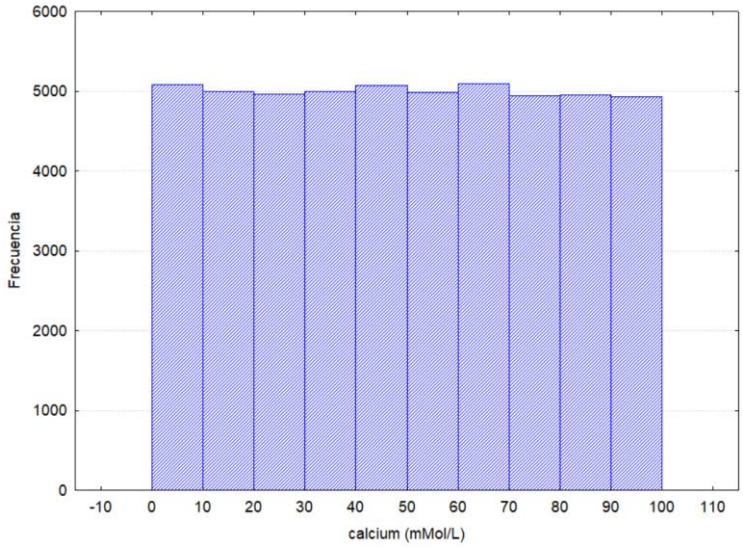


Figura 3.3 Histograma calcium

Sin embargo, cuando las 3 variables (*calcium*, *magnesium*, *sodium*), se grafican en un diagrama ternario para intentar ver la representatividad en dicho plano, cambian su distribución. Esto ocurre debido a que en el triángulo se grafican las relaciones porcentuales de las concentraciones, resultando en un diagrama como el de la figura 3.4. Hay más puntos mientras los valores se aproximan más a las fronteras o al centro. Esto indica que, en esta representación, los valores que tenían de una distribución uniforme se aproximan a una distribución normal al ser presentados como relaciones porcentuales, esto ocurre porque en efecto, los datos están normalizados. En la figura 3.5, se puede observar que la distribución de la variable *calcium*, una vez está en relaciones porcentuales, tiende a seguir una distribución normal.

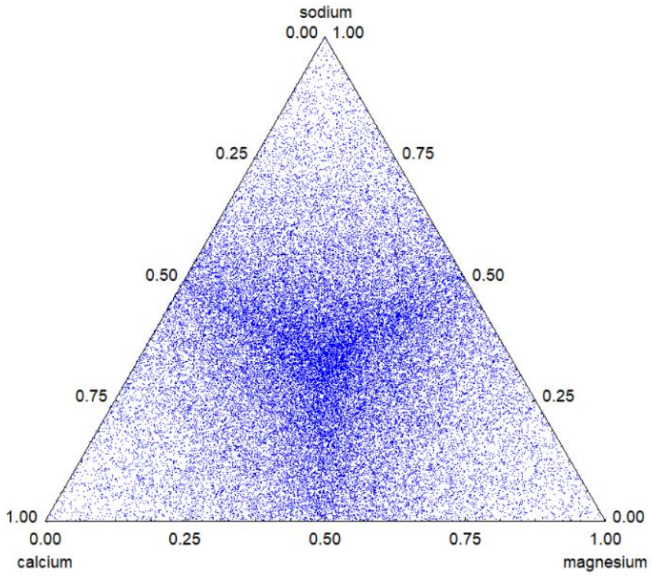


Figura 3.4 Diagrama ternario calcium, magnesium y sodium

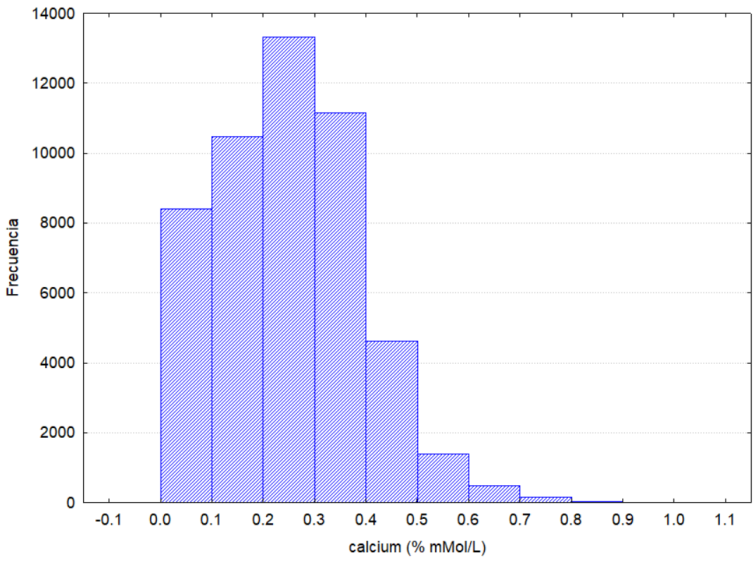


Figura 3.5 Histograma calcium (% mMol/L)

Las observaciones anteriores llevaron la siguiente pregunta, ¿Es posible generar 3 números pseudoaleatorios de suma constante tal que dichos números una vez normalizados sigan una distribución uniforme? O Incluso, *¿n* números? Se diseñó el siguiente algoritmo para generar muestras que cumplieran con las características anteriores, siendo 3, el número de variables consideradas para la generación de una sola muestra.

1.- Inicio

2.- Generar *a*, *b* y *c* de una distribución U (0,1)

3.- *suma = a + b + c*

4.- *anorm = a / suma*

5.- *bnorm= b / suma*

6.- *cnorm = c / suma*

7.- si *anorm y bnorm y cnorm* no se encuentran en la lista de muestras, añadir a la lista de muestras, Si no, volver al paso 2.

8.- Fin

Con el algoritmo anterior, se pudieron generar 628 muestras, las probabilidades de que ocurran varios eventos en los que al generar 3 variables una sola variable una vez normalizada corresponda a más del 95% o menos del 5% de la concentración, son muy bajas, aproximadamente 0.0015%, estos son los espacios que se ven poco nutridos en la figura 3.4.

Se busca que, por cada variable, aumente la frecuencia de sus datos normalizados por debajo y por encima de la medía (aprox. ~ 0.33333), para que la distribución mostrada en la figura 3.5**,** se asemeje cada vez más a una distribución uniforme. De cumplirse estas condiciones se podría apreciar una distribución uniforme en la proyección de los datos en el diagrama ternario, que expresa las relaciones porcentuales de las variables.

Por lo que no es viable resolver este problema mediante la generación de números pseudoaleatorios que primero siguen una distribución uniforme intentando verificar las condiciones en las que una muestra de n variables no se repita, ya que esta metodología solo sería factible con cuando n es infinito.

Este problema requiere un algoritmo más sofisticado para demostrar si es posible o no, generar *n* números de suma constante que sigan una distribución uniforme. Por lo que, al no conocer una metodología que pudiese resolver este problema, se ha decidido tomar metodología de generación bases de datos al procedimiento expuesto en la sección 3.3.2.

Capítulo 4: Metodología general

La metodología para el desarrollo de los diagramas multidimensionales para la clasificación de agua consta de 6 fases: (i) Generación de la base de datos inicial de las concentraciones en mMol/L de los cationes $Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$ y los aniones $SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}, CO_3{}^{2-}$ mediante simulación Monte Carlo (Verma y Quiroz-Ruiz, 2006) ; (ii) Balanceo iónico de cargas; (iii) Cálculo de las transformaciones logarítmicas para abrir el campo (Verma *et al.* , 2016); (iv) Eliminación de datos discordantes con el programa DOMuDaF (Discordant Outlier from Multivariate Data through F-test of w) (Verma *et al.*, 2016); (v) Obtención de funciones discriminantes mediante el análisis discriminante lineal y el análisis canónico (Agrawal *et al.* ,2004); (vi) Desarrollo de una herramienta computacional para clasificar nuevas muestras con la capacidad de llevar a cabo pruebas de robustez (Verma *et al.*, 2016). En la figura 4.1 se aprecia de forma esquemática la metodología general.

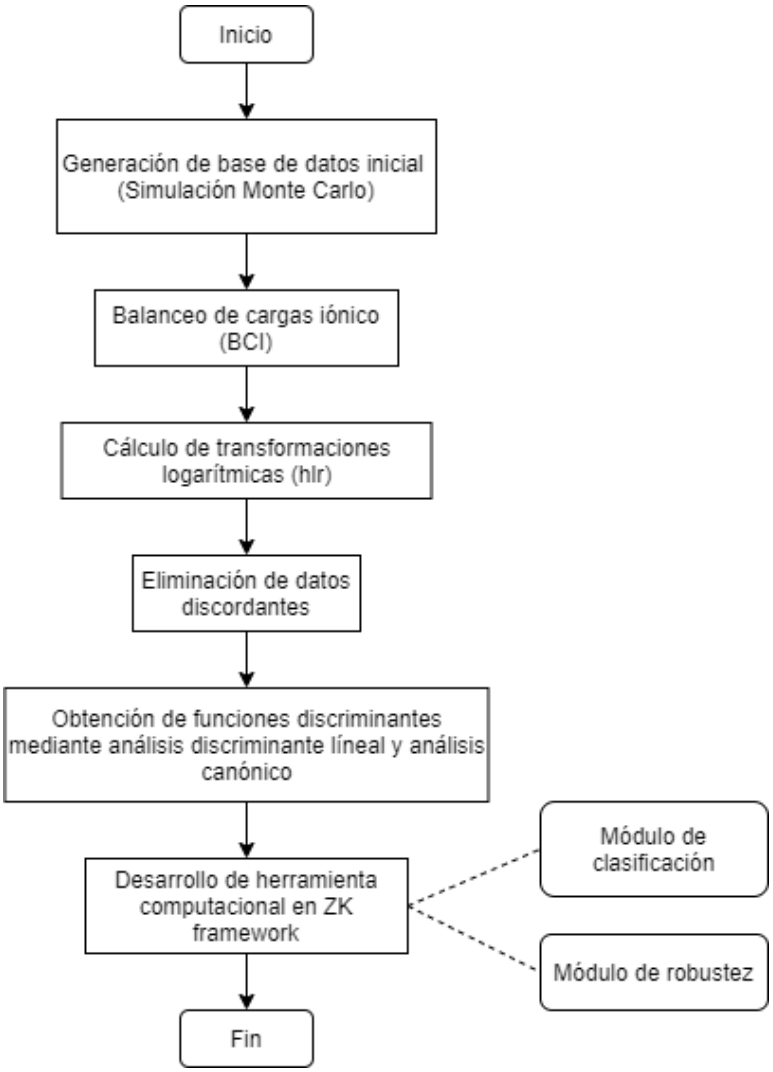


Figura 4.1 Metodología general

4.1 Generación de bases de datos mediante simulación Monte Carlo

El primer paso de la metodología es generar una base de datos de entrenamiento inicial para poder utilizar el análisis discriminante lineal (LDA) y análisis canónico. Para crear la base de datos se utilizó el procedimiento expuesto en la sección 3.3.2 "Generación de muestras aleatorias con distribución uniforme". Se utiliza el procedimiento propuesto por Verma y Quiroz-Ruiz (2006) para generar 8 cadenas de valores que pertenezcan a la distribución U (0,1), cada uno de estos valores es multiplicado por un escalar $c = 100$ para tener 8 distribuciones U (0,100).

Cada muestra simulada se compone de 8 valores que corresponden a los cationes y aniones mayoritarios ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, {SO_4}^{-2}, Cl^{-}, {HCO_3}^{-} y\ {CO_3}^{2-}$). Para la base de datos de entrenamiento se generaron 50,000 muestras y para la base de datos de validación externa se generaron 8,000 muestras, esta base de datos sirve para visualizar el rendimiento de los modelos de clasificación en un conjunto de datos que no haya sido utilizado en el entrenamiento (Géron, 2019).

En la tabla 4.1 se visualizan cuantas muestras de la generación pertenecen a cada una de las 16 clases que fueron determinadas con combinación cruzada del catión y el anión mayoritarios en las bases de datos de entrenamiento y de validación externa.

Tabla 4.1 Número de muestras

Clase	**Catión mayoritario**	**Anión mayoritario**	**Número de muestras**	
			Entrenamiento	**Validación e.**
calcium-sulfate	Ca^{2+}	${SO_4}^{-2}$	3102	510
calcium-chloride	Ca^{2+}	Cl^{-}	3033	479
calcium-bicarbonate	Ca^{2+}	${HCO_3}^{-}$	3131	517
calcium-carbonate	Ca^{2+}	${CO_3}^{2-}$	3031	511
magnesium-sulfate	Mg^{2+}	${SO_4}^{-2}$	3120	480
magnesium-chloride	Mg^{2+}	Cl^{-}	3143	526
magnesium-bicarbonate	Mg^{2+}	${HCO_3}^{-}$	3161	466
magnesium-carbonate	Mg^{2+}	${CO_3}^{2-}$	3160	508
sodium-sulfate	Na^{+}	${SO_4}^{-2}$	3082	468
sodium-chloride	Na^{+}	Cl^{-}	3120	523
sodium-bicarbonate	Na^{+}	${HCO_3}^{-}$	3148	525
sodium-carbonate	Na^{+}	${CO_3}^{2-}$	3134	481
potassium-sulfate	K^{+}	${SO_4}^{-2}$	3147	530
potassium-chloride	K^{+}	Cl^{-}	3128	488
potassium-bicarbonate	K^{+}	${HCO_3}^{-}$	3113	512
potassium-carbonate	K^{+}	${CO_3}^{2-}$	3247	476
Total			50000	8000

4.2 Balanceo y ajuste de balance de cargas

La calidad de los datos composicionales de las muestras generadas manualmente fue evaluada mediante el cálculo del parámetro BCI (balance de cargas iónico) (ecuación 4.1) propuesto por Nicholson (1993).

$$BCI = \frac{|\sum cationes + \sum aniones|}{|\sum aniones - \sum cationes|} \qquad (4.1)$$

Donde ∑ aniones y ∑ cationes esta dado en las unidades mEq/L. Este parámetro se estableció de manera arbitraria a 0.00005 (máximo desbalance permitido). Como las muestras son generadas pseudoaleatoria mente es muy poco probable que cumplan con un balanceo mínimo del 0.00005, entonces se ajusta un balance entre los cationes y los aniones mediante un factor de desbalance **(**ecuación 4.2**)** que se aplica a un lado de la muestra (parte catiónica o parte aniónica) para conseguir un valor más apto de BCI.

$$Factor\ de\ desbalance = \frac{|\sum cationes|}{|\sum aniones|} \qquad (4.2)$$

El factor de desbalance se aplica con un incremento pseudoaleatorio entre el 1 y 10%. El factor de desbalance se puede aplicar tanto a la parte catiónica o aniónica de la muestra para mejorar el BCI, esto es determinado de forma pseudoaleatoria. El procedimiento de balance de cargas iónico se encuentra esquematizado en la figura 4.2.

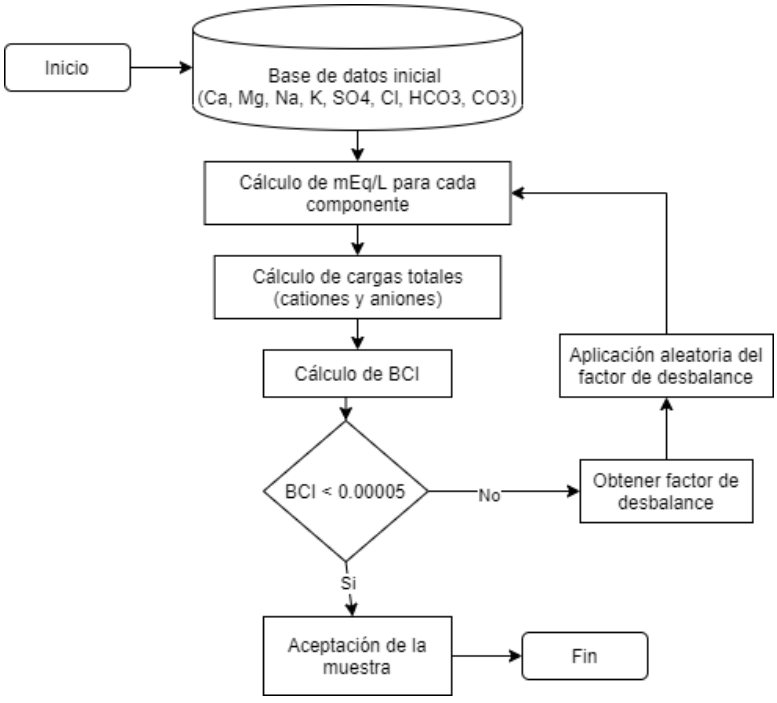


Figura 4.2 Procedimiento BCI

4.3 Transformaciones logarítmicas (hlr)

Debido a los problemas que provoca la representación de datos en los diagramas ternarios (Verma, 2015) se usan transformaciones logarítmicas para abrir las variables a valores positivos y negativos antes de realizar el análisis discriminante lineal.

En este trabajo se utiliza la transformación *hlr* o "hybrid log-ratio transformation" (ecuación 4.3) utilizada en Verma *et al.* (2016).

$$hlr_{j+1} = \ln\left[\frac{(V_1 * V_2 * \dots V_n)^{\frac{1}{n}}}{V_{j+1}}\right]; \quad j = 1, 2, \dots, (n-1) \tag{4.3}$$

En donde $V_1 \dots V_n$ corresponden a las concentraciones en mM/L de los cationes $Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$ y los aniones ${SO_4}^{-2}, Cl^{-}, {HCO_3}^{-}, {CO_3}^{2-}$. Las ecuaciones de las transformaciones utilizadas se pueden observar en la tabla 4.2.

Tabla 4.2 Ecuaciones hlr

$hlr_2 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{Mg}\right]$	(4.4)
$hlr_3 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4**Cl*HCO_3*CO_3)^{\frac{1}{8}}}{Na}\right]$	(4.5)
$hlr_4 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{K}\right]$	(4.6)
$hlr_5 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{SO_4}\right]$	(4.7)
$hlr_6 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{Cl}\right]$	(4.8)
$hlr_7 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{HCO_3}\right]$	(4.9)
$hlr_8 = \ln\left[\frac{(Ca*Mg*Na*K*SO_4*Cl*HCO_3*CO_3)^{\frac{1}{8}}}{CO_3}\right]$	(4.10)

4.4 Separación de datos discordantes

Para comprobar la multinormalidad de las transformaciones logarítmicas *hlr* (hlr_2-hlr_8) de los 8 iones mayores se utilizó el programa DOMuDaF (Discordant Outlier from Multivariate Data through F-test of w; Verma *et al*., 2016). En el reporte obtenido por dicha herramienta se pueden separar las muestras discordantes, de manera que los datos restantes son los que se utilizarán para realizar el análisis discriminante lineal, ya que esta técnica estadística asume la multinormalidad de los grupos. El procedimiento que lleva acabo DOMuDaF se puede visualizar en la figura 4.3.

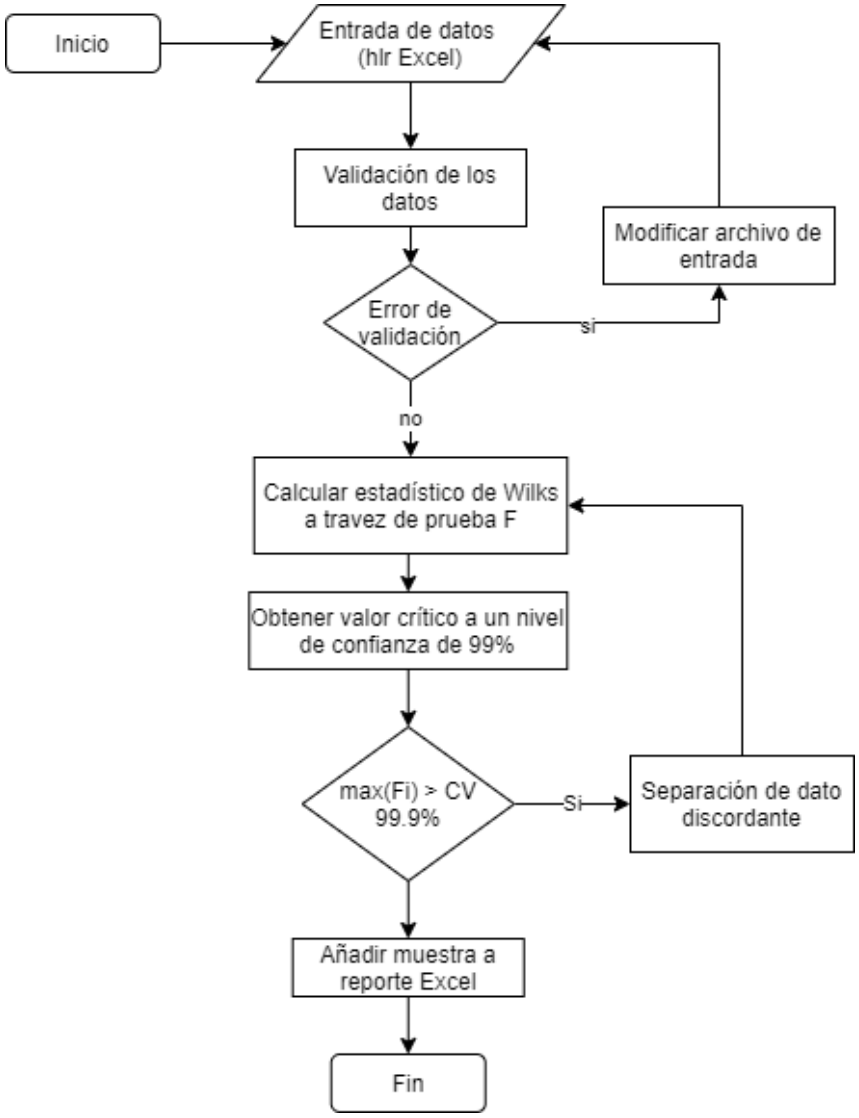


Figura 4.3 Procedimiento DOMuDaF (Verma et al., 2016)

En la tabla 4.3 se puede visualizar el número de datos que siguen una distribución multinormal para cada una de las 16 clases en las bases de datos de entrenamiento y de validación externa. En la base de datos de entrenamiento el procedimiento de DOMuDaF (Verma *et al*., 2016) se hizo con las transformaciones *hlr* (hlr_2, hlr_3, hlr_4, hlr_5, hlr_6, hlr_7 y hlr_8), mientras que en la base de datos de validación externa se hizo con las concentraciones en mMol/L de los iones mayoritarios (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, ${SO_4}^{-2}$, Cl^{-}, ${HCO_3}^{-}$ y ${CO_3}^{2-}$).

Este procedimiento es un requisito para el uso de la técnica estadística de análisis discriminante lineal (Verma *et al*., 2016). Sin embargo, la multinormalidad de los grupos no es un requisito en otras técnicas de discriminación. En el **Apéndice B**, se hace el entrenamiento y evaluación de otras técnicas de discriminación (*Categorial Boosting* y *Support Vector Machines)* en donde la separación de datos discordantes en los grupos no fue necesaria.

Tabla 4.3 Numero de muestras censuradas por clase

	Número de muestras censuradas	
Clase	**Entrenamiento**	**Validación E.**
calcium-sulfate	2853	467
calcium-chloride	2818	450
calcium-bicarbonate	2909	487
calcium-carbonate	2788	469
magnesium-sulfate	2892	428
magnesium-chloride	2870	476
magnesium-bicarbonate	2908	433
magnesium-carbonate	2915	481
sodium-sulfate	2869	427
sodium-chloride	2901	483
sodium-bicarbonate	2927	483
sodium-carbonate	2862	441
potassium-sulfate	2909	503
potassium-chloride	2927	458
potassium-bicarbonate	2923	474
potassium-carbonate	3021	443
Total	46292	7403

4.5 Construcción de modelos de clasificación

El análisis discriminante lineal (LDA) es una técnica estadística que se utiliza para diferenciar entre miembros de diferentes grupos predefinidos (Agrawal *et al.*, 2004). Él método calcula una combinación lineal de las dimensiones originales en donde las clases del nuevo espacio están divididas por líneas rectas. Es un requisito que las clases con las que se entrenará al modelo sean multinormales, esto se cumple separando los datos discordantes con el software DOMuDaF (Verma *et al.*, 2016).

En un análisis discriminante lineal de 3 clases se obtienen dos funciones discriminantes (DF1 y DF2), esto permite que el diagrama resultante sea útil para una representación gráfica, sin embargo, al involucrar, por ejemplo, 4 clases, se obtienen 3 ($n_{clases} - 1$) funciones discriminantes para diferenciar dichas clases. Tres funciones discriminantes no son factibles para una representación gráfica (Agrawal *et al.*, 2004). Siguiendo la estrategia para lidiar con más de 3 clases de Agrawal *et al.* (2004) se decidió mantener el número de clases por diagrama en 3, de manera que, en caso de ser necesario, los diagramas resultantes tendrán sentido si se les quiere interpretar gráficamente, a este ensamble de 3 LDA para un mismo diagrama se denomina como "tres a la vez". Cada clasificador de k clases resulta también en k centroides de $k - 1$ dimensiones, con la distancia de estos centroides a las funciones discriminantes aplicadas a una muestra, es posible determinar la categoría de un ejemplo desconocido (sección 4.5.3).

En este trabajo existen 4 clases catiónicas y 4 clases aniónicas, de manera que, para aplicar LDA de estas categorías siguiendo la estrategia multiclase "tres a la vez", es necesario determinar las clases que participaran en la discriminación. Cada uno de los diagramas LDA y las clases que discrimina se pueden apreciar en la tabla 4.4, en donde los diagramas del 1 al 4 son responsables de la clasificación catiónica mientras que del 5 al 8 de la clasificación aniónica.

Tabla 4.4 Clasificadores LDA "tres a la vez"

No.	Diagrama	Clases que discrimina
1	Ca-Mg-Na	*calcium, magnesium, sodium*
2	Ca-Mg-K	*calcium, magnesium, potassium*
3	Ca-Na-K	*calcium, sodium, potassium*
4	Mg-Na-K	*magnesium, sodium, potassium*
5	SO4-Cl-HCO3	*sulfate, chloride, bicarbonate*
6	SO4-Cl-CO3	*sulfate, chloride, carbonate*
7	SO4-HCO3-CO3	*sulfate, bicarbonate, carbonate*
8	Cl-HCO3-CO3	*chloride, bicarbonate, carbonate*

Se propusieron dos modelos de clasificación LDA, *7-hlr* y *7-molar-conc*. El primero, se construye utilizando las transformaciones *hlr* ($hlr_2 - hlr_8$; sección 4.3) como variables descriptivas, mientras que el segundo se construye utilizando las concentraciones en mMol/L de los iones mayores ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, {SO_4}^{-2}, Cl^{-}, {HCO_3}^{-} y\ {CO_3}^{2-}$). Ambos modelos utilizan la estrategia multiclase expuesta en la tabla 4.4.

Las funciones discriminantes para clasificación catiónica y aniónica para el modelo *7-hlr* se describen como en la ecuación 4.11, en el modelo *7-molar-conc* las funciones discriminantes para cationes y aniones corresponden a las ecuaciones 4.12 y 4.13 respectivamente.

$$DF = C_1 * hlr_2 + C_2 * hlr_3 + C_3 * hlr_4 + C_4 * hlr_5 + C_5 * hlr_6 + C_6 * hlr_7 + C_7 * hlr_8 + C_0 \quad (4.11)$$

$$DF = C_1 * Ca + C_2 * Mg + C_3 * Na + C_4 * K + C_5 * SO_4 + C_6 * Cl + C_7 * HCO_3 + C_0 \quad (4.12)$$

$$DF = C_1 * Mg + C_2 * Na + C_3 * K + C_4 * SO_4 + C_5 * Cl + C_6 * HCO_3 + C_7 * CO_3 + C_0 \quad (4.13)$$

En donde C_i corresponde a cada uno de los coeficientes de la función discriminante (en total 7), C_0 una constante, $hlr_2 - hlr_8$ cada una de las transformaciones logarítmicas y $Ca,\ Mg,\ Na,\ K,\ SO_4,\ Cl,\ HCO_3\ y\ CO_3$ las concentraciones en mM/L de la muestra.

4.5.1 Cálculo de probabilidades de clasificación en los modelos

Dos modelos ensamblados LDA "tres a la vez" componen los modelos *7-hlr* y *7-molar-conc*, se usa un ensamble para la clasificación catiónica y otro para la clasificación aniónica.

Para calcular las probabilidades de una muestra desconocida de pertenecer a cualquier de las clases catiónicas o aniónicas de los modelos, es necesario aplicar las funciones discriminantes en dicha muestra y calcular su distancia con los centroides para cada uno de los diagramas que pertenecen al ensamble. La distancia con los centroides se calcula como sigue:

$$d_i = \sqrt{(DF1_j - CX_i)^2 + (DF2_j - CY_i)^2} \quad (4.14)$$

En donde $DF1_j$ y $DF2_j$, son las funciones discriminantes del clasificador LDA j, CX_i y Cy_i son las coordenadas del centroide de la clase i y d_i es la distancia de la muestra desconocida al centroide de la clase i.

Cuando se conocen las distancias de las categorías en el clasificador LDA "tres a la vez", la probabilidad de que la muestra desconocida pertenezca a la clase i se calcula como sigue:

$$g_i = e^{-(\frac{{d_i}^2}{2})} \quad (4.15)$$

Cuando se conocen las 3 probabilidades de las categorías en el clasificador LDA, estas se normalizan. Al sumar las probabilidades de los 4 diagramas que pertenecen a la clasificación catiónica o aniónica de algún modelo, la probabilidad máxima de clasificación es de 75%,

4.5.2 Modelo greater-molar-conc

Greater-molar-conc se refiere al nombre a priori que se le da a una muestra de agua basada en su catión mayoritario ($Ca,\ Mg,\ Na,\ K$) y su anión mayoritario ($SO_4,\ Cl,\ HCO_3,\ CO_3$) en unidades de mMol/L. Existen un total de 16 tipos de agua por combinación cruzada de cada catión y cada anión tal y como se muestran en la tabla 4.3 en la columna de "Clase".

Durante este trabajo, me referiré a *greater-molar-conc* como un modelo de clasificación de agua basado únicamente en los iones mayoritarios, esto debido a que este criterio se puede comparar con los modelos más sofisticados que implementan análisis discriminante lineal y análisis canónico como lo son los modelos 7-*hlr* y 7-*molar-conc*.

4.5.3 Determinación de tipos de agua híbridos

Los dos modelos desarrollados con análisis discriminante lineal (LDA) y análisis canónico despliegan un conjunto de probabilidades de las clases que discrimina, esto se puede utilizar para determinar tipos de agua híbridos. Se define un umbral entre estas probabilidades para determinar si el tipo de agua es básico (un nombre) o híbrido (dos nombres), tanto para la clase catiónica como la aniónica.

Los diagramas dan 4 probabilidades para las clases catiónicas ($Ca,\ Mg,\ Na,\ K$) y 4 probabilidades para las clases aniónicas ($SO_4,\ Cl,\ HCO_3,\ CO_3$). De estas 4 probabilidades, se denota como P_m a la probabilidad más alta y P_n a la segunda probabilidad más alta, las condiciones para que la clase catiónica o aniónica sean básicas son: Si ($P_m \geq 0.5$) y (($P_m - P_n$) ≥ 0.25) y ($P_n \leq 0.25$). Si no se cumplen dichas condiciones entre P_m y P_n, entonces el tipo de agua será híbrido.

Un ejemplo de tipo de agua híbrido sería "*calcium-magnesium-sulfate*", en este ejemplo, solo la clase catiónica es híbrida y la clase aniónica es básica. De esta forma existen 16 clases catiónicas: 4 básicas (*calcium*, *magnesium*, *sodium*, *potassium*) y 12 híbridas (*calcium-magnesium, calcium-sodium, calcium-potassium, magnesium-calcium, magnesium-sodium, magnesium-potassium, sodium-calcium, sodium-magnesium, sodium-potassium, potassium-calcium, potassium-magnesium, potassium-sodium*). De la misma forma 16 clases aniónicas: 4 básicas (*sulfate, chloride, bicarbonate, carbonate*) y 12 híbridas (*sulfate-chloride, sulfate-bicarbonate, sulfate-carbonate, chloride-sulfate, chloride-bicarbonate, chloride-carbonate, bicarbonate-sulfate, bicarbonate-chloride, bicarbonate-carbonate, carbonate-sulfate,*

carbonate-chloride, *carbonate-bicarbonate*). Por combinación cruzada, se determinaron un total de 256 clases combinadas (clase catiónica + clase aniónica).

Estas condiciones se aplican al final de la aplicación de los modelos *7-hlr* y *7-molar-conc*, permitiendo conocer tipos híbridos, es importante mencionar que esta opción no se encuentra disponible con el modelo *greater-molar-conc*, debido a que dicho modelo no genera probabilidades.

4.5.4 Evaluación de los modelos de clasificación

Para visualizar el rendimiento de los modelos *7-hlr* y *7-molar-conc* en las bases de datos de entrenamiento y validación externa, se calculan las precisiones individuales para cada clase catiónica y aniónica (ecuación 4.16), así como la precisión total de clasificación (4.17).

$$Precisión_i = \frac{\mathrm{VP}}{\mathrm{VP+FP}} \qquad (4.16)$$

En donde se calcula la precisión individual de la categoría catiónica o aniónica i, VP corresponde al número de verdaderos positivos y FP al número de falsos positivos.

$$Precisión = \frac{\mathrm{PC}}{\mathrm{PT}} \qquad (4.17)$$

En donde PC es el número de predicciones correctas y PT es el número de predicciones totales**,** este cálculo se efectúa considerando las 4 clases catiónicas y las 4 clases aniónicas, teniendo una precisión total catiónica y aniónica.

4.6 Diseño de la herramienta computacional

Para construir una herramienta computacional para la clasificación de agua se necesitan definir 2 módulos. El primero es el módulo de clasificación y se refiere a la clasificación de muestras desconocidas, en las que, de acuerdo con un modelo seleccionado, se darán los tipos de agua. El segundo es el módulo de robustez que implementa dos procedimientos: propagación de errores y cambios composicionales. A continuación, se explican los módulos.

4.6.1 Módulo de clasificación

El módulo de clasificación es el encargado de asignar una clase a muestras de agua desconocidas, es necesario seleccionar un modelo de clasificación (*7-hlr*, *7-molar-conc o greater-molar-conc*), ya que para determinar la clasificación en cada modelo es necesario llevar a cabo operaciones diferentes. El funcionamiento de este módulo se puede visualizar en la figura 4.4.

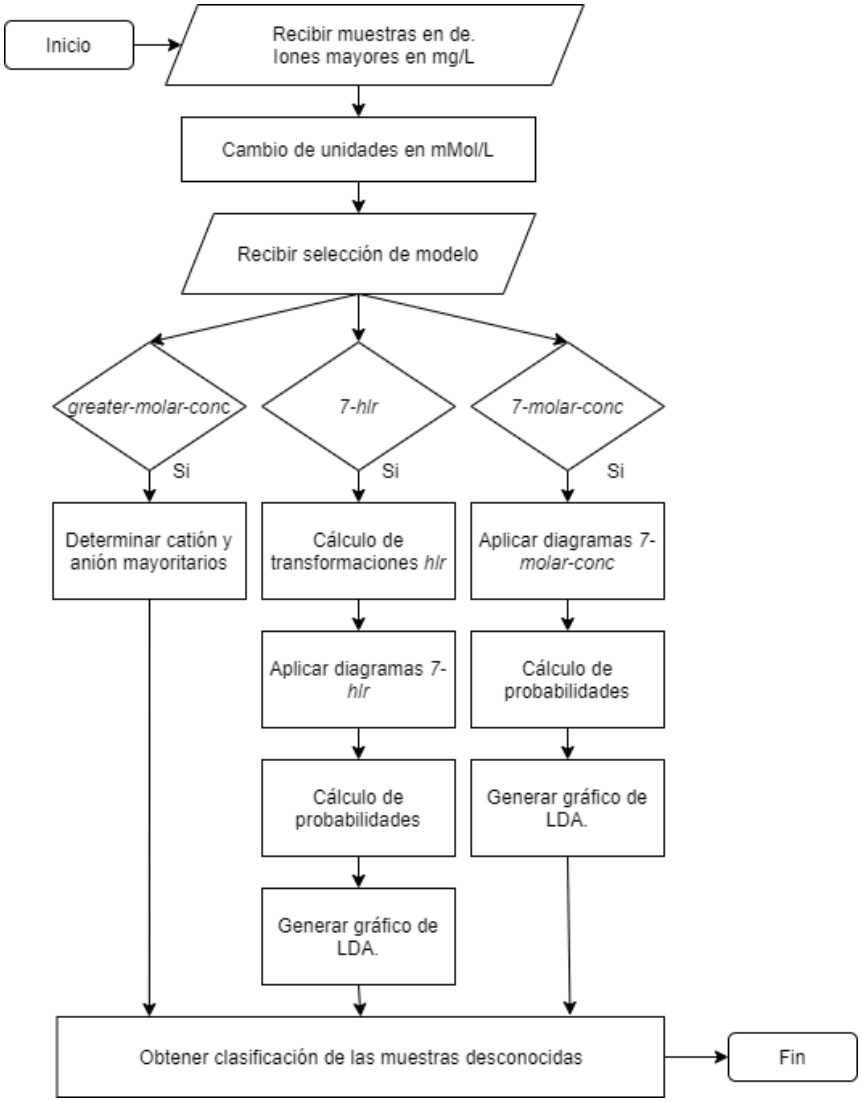


Figura 4.4 Diagrama de flujo de módulo de clasificación

4.6.2 Módulo de robustez

La robustez es la capacidad que tiene un modelo de que sus predicciones persistan bajo distintos entornos con variaciones o perturbaciones. Se definen dos procesos de robustez para los modelos *7-hlr*, *7-molar-conc* y *greater-molar-conc*; (i) propagación de errores, (ii) cambios composicionales. Ambos procesos se explican a continuación.

i) Propagación de errores

Consiste en seleccionar una muestra inicial de agua que tiene las concentraciones de los iones mayores ($Ca,\ Mg,\ Na,\ K, SO_4,\ Cl,\ HCO_3,\ CO_3$) en unidades de mg/L y sus medidas de dispersión (U^{99}ó RSD ó S) para cada uno de los elementos. Después se efectúa una simulación Monte Carlo (Verma y Quiroz, 2006) para generar 2200 de 8 variables con distribución N (0,1), cada una de esas muestras se multiplica por el valor de dispersión (U^{99}ó RSD ó S) y se suma el valor correspondiente a cada variable de la muestra inicial. Al terminar, los datos simulados pasan por el módulo de clasificación. El flujo de este procedimiento se encuentra ilustrado en la figura 4.5.

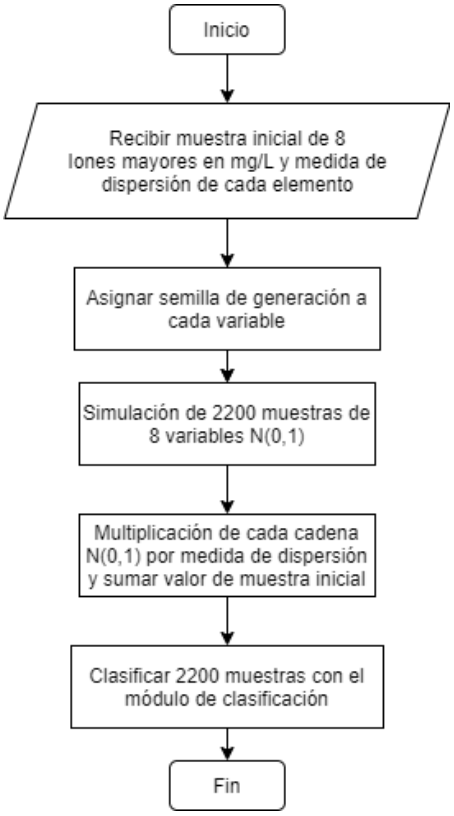


Figura 4.5 Diagrama de flujo de del procedimiento de propagación de errores

Cuando las muestras simuladas entran al módulo de clasificación, es posible conocer la proporción de muestras que permanecieron en el tipo de agua de la muestra inicial a pesar de la propagación de una medida de dispersión.

ii) Cambios composicionales

En este proceso se selecciona una muestra inicial de los iones mayores ($Ca,\ Mg,\ Na,\ K,\ SO_4,\ Cl,\ HCO_3\ y\ CO_3$) en unidades de mg/L y un porcentaje de disminución o adición para cada uno de los iones mayores (puede ser 0, que quiere decir que no cambia). Entonces se simula cambios en la composición química del agua a través de la adición o sustracción de las concentraciones de los iones mayores hasta un límite inferior de 0.002 mg/L o un límite superior de 50000 mg/L o hasta que la muestra cambie el tipo de agua inicial en el modelo seleccionado (*7-hlr* ó *7-molar-conc* ó *greater-molar-conc*), con esta simulación se puede contar el número de pasos que soporta cada modelo antes de pasar por una condición de paro. Este procedimiento se encuentra ilustrado en la figura 4.6.

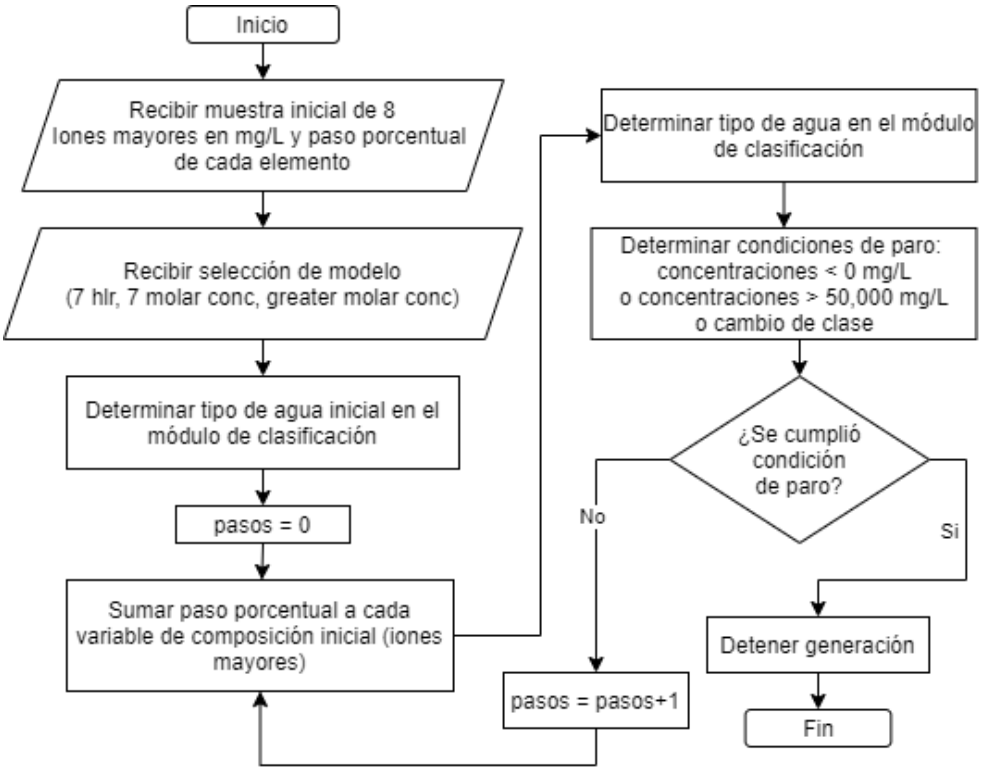


Figura 4.6 Diagrama de flujo del procedimiento de cambios composicionales

Capítulo 5: Resultados

5.1 Entrenamiento y evaluación de los modelos de clasificación

Cuando se entrenan los modelos de LDA y análisis canónico (*7-hlr y 7-molar-conc*), se obtienen las funciones discriminantes y los centroides de cada una de las clases (sección 4.5.1). En las tablas 5.1 y 5.2 se presentan las constantes de las funciones discriminantes y las coordenadas de los centroides para utilizar el modelo *7-hlr*. De la misma forma, las tablas 5.3 y 5.4 corresponden a las constantes y centroides del modelo *7-molar-conc*.

Tabla 5.1 Constantes de las funciones discriminantes (7-hlr)

Diagrama	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_0
DF1-hCa-Mg-Na	-0.4840	-1.8464	-0.7552	-0.5560	-0.5913	-0.6041	-0.5319	0.0195
DF2-hCa-Mg-Na	1.9873	0.5890	0.8572	0.9238	0.9003	0.8920	0.9229	0.0230
DF1-hCa-Mg-K	1.9524	0.8069	0.4520	0.8561	0.8235	0.8405	0.8423	0.0232
DF2-hCa-Mg-K	-0.6321	-0.7887	-1.8504	-0.5630	-0.6300	-0.6264	-0.5645	0.0170
DF1-hCa-Na-K	-0.9876	-0.9277	-1.9160	-0.7067	-0.7678	-0.7620	-0.6951	-0.0046
DF2-hCa-Na-K	0.5747	1.6787	0.0002	0.4877	0.4838	0.5062	0.4630	-0.0078
DF1-hMg-Na-K	1.0514	-0.0756	-0.8972	0.2167	0.1665	0.1742	0.2296	0.0223
DF2-hMg-Na-K	-0.4741	1.1190	-0.6415	-0.0388	-0.0417	-0.0298	-0.0621	-0.0134
DF1-hSO4-Cl-HCO3	0.0088	-0.0344	-0.0421	0.9387	-0.7352	-0.7560	-0.2193	0.0071
DF2-hSO4-Cl-HCO3	-0.0097	-0.0001	-0.0060	-0.0151	0.9605	-0.9894	0.0103	-0.0162
DF1-hSO4-Cl-CO3	-0.0022	-0.0139	-0.0157	1.0772	-0.3384	-0.0505	-0.9100	0.0013
DF2-hSO4-Cl-CO3	0.0114	-0.0439	-0.0544	0.0617	-1.2710	-0.1880	0.5893	0.0218
DF1-hSO4-HCO3-CO3	0.0056	-0.0369	-0.0417	-0.3566	-0.1549	-1.0755	0.8689	0.0007
DF2-hSO4-HCO3-CO3	-0.0030	0.0124	0.0265	-1.0265	0.1129	0.7957	0.6022	-0.0084
DF1-hCl-HCO3-CO3	0.0129	-0.0371	-0.0312	-0.1961	-0.2396	-1.1014	0.8371	0.0032
DF2-hCl-HCO3-CO3	0.0164	-0.0256	-0.0118	-0.1004	-1.1998	0.5411	0.3969	0.0094

Tabla 5.2 Centroides del modelo 7-hlr						
Diagrama	**Coordenadas de los centroides**					
	Para la clase 1		**Para la clase 2**		**Para la clase 3**	
	CX_1	CY_1	CX_2	CY_2	CX_3	CY_3
hCa-Mg-Na	-0.7638	0.7991	-0.3097	-1.0437	1.0617	0.2602
hCa-Mg-K	0.7463	-0.8150	-1.0683	-0.2378	0.3304	1.0204
hCa-Na-K	-0.9639	0.5432	-0.0045	-1.0799	0.9347	0.5354
hMg-Na-K	-1.0101	0.4375	0.1103	-1.0726	0.8852	0.6222
hSO4-Cl-HCO3	-1.0942	0.0032	0.5410	-0.9363	0.5466	0.9210
hSO4-Cl-CO3	-1.0682	-0.2791	0.2858	1.0470	0.7783	-0.7631
hSO4-HCO3-CO3	0.1314	1.0850	0.8727	-0.6496	-1.0095	-0.4250
hCl-HCO3-CO3	0.0691	1.0665	0.9010	-0.5862	-0.9761	-0.4698

Tabla 5.3 Constantes de las funciones discriminantes (7-molar-conc)								
Diagrama	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_0
DF1-cCa-Mg-Na	-0.01951	-0.01952	0.03812	-0.00057	0.00026	0.00002	0.00023	-0.02171
DF2-cCa-Mg-Na	0.03357	-0.03380	0.00013	-0.00013	0.00022	0.00025	0.00045	-0.02224
DF1-cCa-Mg-K	-0.01825	-0.01961	-0.00044	0.03802	-0.00024	0.00044	-0.00001	-0.07075
DF2-cCa-Mg-K	-0.03424	0.03294	-0.00007	0.00051	-0.00007	-0.00029	0.00003	0.04585
DF1-cCa-Na-K	0.00120	-0.00015	0.03198	-0.03321	0.00015	-0.00041	0.00014	0.02541
DF2-cCa-Na-K	0.03808	-0.00051	-0.01882	-0.01671	-0.00021	-0.00029	-0.00014	-0.00685
DF1-cMg-Na-K	0.00059	-0.00267	-0.03107	0.03393	-0.00029	0.00039	-0.00014	-0.04207
DF2-cMg-Na-K	0.00031	-0.03770	0.02031	0.01563	0.00052	0.00031	0.00031	-0.02996
DF1-cSO4-Cl-HCO3	-0.00042	-0.00009	0.00021	0.00318	-0.03369	0.03098	-0.00006	-0.00974
DF2-cSO4-Cl-HCO3	-0.00029	-0.00021	-0.00027	-0.03722	0.01529	0.02042	0.00073	0.01552
DF1-cSO4-Cl-CO3	0.00011	0.00004	0.00012	0.02137	-0.03793	0.00036	0.01717	0.00991
DF2-cSO4-Cl-CO3	0.00011	0.00004	0.00012	-0.03231	-0.00261	-0.00001	0.03482	-0.00842
DF1-cSO4-HCO3-CO3	0.00061	0.00027	0.00019	0.01544	0.00011	-0.03818	0.02231	0.04082
DF2-cSO4-HCO3-CO3	-0.00006	0.00018	0.00003	0.03524	-0.00011	-0.00365	-0.03153	-0.00214
DF1-cCl-HCO3-CO3	-0.00043	-0.00019	0.00011	0.00032	-0.03138	0.03324	-0.00174	-0.00125
DF2-cCl-HCO3-CO3	-0.00078	-0.00011	-0.00058	0.00117	0.01977	0.01658	-0.03702	-0.00956

Tabla 5.4 Centroides del modelo 7-molar-conc						
Diagrama	**Coordenadas de los centroides**					
	Para la clase 1		**Para la clase 2**		**Para la clase 3**	
	CX_1	CY_1	CX_2	CY_2	CX_3	CY_3
cCa-Mg-Na	-0.82945	1.37613	-0.82280	-1.35400	1.64039	0.00365
cCa-Mg-K	-0.79939	-1.39078	-0.85851	1.33212	1.61573	0.03207
Ca-Na-K	0.06155	1.56751	1.41701	-0.81273	-1.44982	-0.71521
cMg-Na-K	-0.13204	-1.54071	-1.37709	0.87152	1.48111	0.66004
cSO4-Cl-HCO3	0.12452	-1.53943	-1.48447	0.66359	1.34228	0.86543
cSO4-Cl-CO3	0.90632	-1.31127	-1.63764	-0.09988	0.72636	1.40341
cSO4-HCO3-CO3	0.67873	1.43151	-1.62147	-0.15296	0.95777	-1.26969
cCl-HCO3-CO3	-1.37900	0.83489	1.44371	0.69942	-0.08313	-1.53416

Para evaluar el rendimiento de los modelos *7-hlr* y *7-molar-conc* se calcularon las métricas descritas en la sección 4.5.4 y los resultados pueden observarse en la tabla 5.5 para la base de datos de entrenamiento y la base de datos de validación externa. El modelo *7-molar-conc* tiene precisiones totales (catiónicas ya aniónicas) mayores que el modelo *7-hlr* tanto en la base de datos de entrenamiento como en la base de datos de validación externa, esto hace evidente que el modelo *7-molar-conc* discrimina con mayor precisión los 16 tipos básicos de agua.

Tabla 5.5 Precisión de entrenamiento y de validación para los modelos 7-hlr y 7-molar-conc										
Base de datos de entrenamiento (46292 muestras)										
modelo	**Precisión catiónica individual**				**Precisión catiónica total**	**Precisión aniónica individual**				**Precisión aniónica total**
	Ca	**Mg**	**Na**	**K**		**SO4**	**Cl**	**HCO3**	**CO3**	
7-hlr	93.57	94.16	91.63	90.26	92.34	93.57	92.50	92.33	93.78	93.04
7-molar-conc	94.62	95.73	99.65	99.76	97.35	94.33	99.77	99.76	94.05	96.81
base de datos de validación externa (7403 muestras)										
7-hlr	93.13	93.44	91.08	90.06	91.88	93.14	92.58	93.43	93.02	93.04
7-molar-conc	94.21	95.26	99.88	99.72	97.14	93.52	99.83	99.83	93.76	96.57

5.2 Herramienta computacional

Se desarrolló la herramienta computacional llamada *WaterMClaSys_lda* en ZK framework, se implementaron los módulos de clasificación y robustez descritos en la sección 4.6. En el módulo de clasificación esta herramienta puede recibir hasta 8,000 muestras de agua de los iones mayores en unidades de mg/L y clasificar las muestras con cualquiera de los modelos de clasificación disponibles (*7-hlr*, *7-molar-conc* y *greater-molar-conc*).

El módulo de robustez realiza los procedimientos de propagación de errores y cambios composicionales desde una muestra inicial, este módulo se utilizó para obtener los resultados de la sección 5.3 "Pruebas de robustez". *WaterMClaSys_lda* se encuentra disponible para su uso junto con las plantillas para cada funcionalidad en el portal http://tlaloc.ier.unam.mx/watermclasys_lda.

5.3 Pruebas de robustez

En total existen 64 muestras de prueba tomadas de la base de datos de entrenamiento para evaluar la robustez de los modelos *7-hlr* y *7-molar-conc*: 16 muestras más cercanas al 75% (75H, muestras muy interiores) y al 50% (50H, muestras interiores) en el modelo *7-hlr*; 16 muestras más cercanas al 75% (75M, muestras muy interiores) y al 50% (50H, muestras interiores) en el modelo *7-molar-conc*. Las muestras de las pruebas 75H y 50H junto con las probabilidades de la categoría a la que pertenecen, se presentan en la tabla 5.6., de la misma forma las 75M y 50M se aprecian en la tabla 5.7. Algunas de las muestras de prueba son iguales para ambos modelos, sin embargo, esto no debería ser considerado una constante, ya que los modelos *7-hlr* y *7-molar-conc*, a pesar de estar aplicados a la misma muestra, resultan en probabilidades diferentes. No se determinaron muestras de prueba para el modelo *greater-molar-conc* ya que este no calcula probabilidades de clasificación

Tabla 5.6 Pruebas 75H y 50H

				Concentraciones (mMol/L)								Probabilidades 7-hlr	
No.	**Catión mayor**	**Anión mayor**	**Prueba**	**Ca**	**Mg**	**Na**	**K**	**SO4**	**Cl**	**HCO3**	**CO3**	**Prob. Catiónica**	**Prob. Aniónica**
1	Ca	SO_4	75H	87.81	9.97	19.53	20.50	93.80	17.74	25.62	2.31	66.43	70.34
2	Ca	Cl	75H	79.72	20.78	6.35	9.63	30.15	90.89	15.99	24.89	70.60	65.32
3	Ca	HCO_3	75H	64.64	18.40	11.93	2.83	17.75	2.34	123.31	9.84	68.38	72.98
4	Ca	CO_3	75H	88.87	17.60	9.36	16.40	8.36	16.18	11.46	97.17	70.00	72.26
5	Mg	SO_4	75H	17.27	87.98	8.92	6.18	89.30	4.59	13.97	14.23	72.15	72.06
6	Mg	Cl	75H	12.78	66.12	7.19	0.82	10.19	87.18	29.02	14.61	72.58	67.03
7	Mg	HCO_3	75H	10.70	70.92	5.28	24.21	20.36	20.44	109.81	10.88	68.07	70.54
8	Mg	CO_3	75H	6.12	44.50	5.54	4.97	13.86	4.96	4.11	37.48	72.46	66.94
9	Na	SO_4	75H	21.70	19.73	135.00	8.17	80.58	13.07	4.74	23.53	71.85	69.26
10	Na	Cl	75H	8.07	0.93	52.18	2.57	4.29	34.05	11.35	9.38	72.10	66.25
11	Na	HCO_3	75H	13.59	14.57	55.64	5.54	3.53	5.30	74.42	15.35	68.63	71.23
12	Na	CO_3	75H	24.00	11.09	108.02	12.82	6.51	16.49	10.63	75.45	70.61	71.03
13	K	SO_4	75H	6.99	12.36	9.32	100.34	63.38	13.11	3.23	2.62	73.37	72.92
14	K	Cl	75H	6.14	22.61	4.63	62.35	4.62	90.88	8.17	8.10	68.84	73.44
15	K	HCO_3	75H	2.90	12.50	8.00	107.89	5.74	24.05	83.32	13.91	73.35	68.90
16	K	CO_3	75H	4.06	11.64	15.13	100.06	15.12	1.08	8.66	53.30	72.41	69.56

1	Ca	SO_4	50H	56.95	37.52	28.78	13.21	60.81	34.53	29.75	22.52	49.51	49.67
2	Ca	Cl	50H	89.80	43.47	73.08	6.37	30.16	99.01	24.14	81.26	49.94	50.36
3	Ca	HCO_3	50H	49.91	22.46	14.43	29.34	31.54	46.58	66.68	6.08	49.79	50.36
4	Ca	CO_3	50H	63.46	35.59	12.63	39.54	32.36	31.36	25.99	64.10	50.07	50.34
5	Mg	SO_4	50H	52.72	89.59	57.45	12.19	93.73	68.63	23.71	37.24	50.10	49.79
6	Mg	Cl	50H	4.10	78.44	40.45	64.04	61.11	81.19	35.98	15.09	50.10	50.08
7	Mg	HCO_3	50H	50.94	90.34	67.73	7.96	79.74	34.14	103.40	30.61	49.06	50.03
8	Mg	CO_3	50H	28.00	88.64	50.12	45.43	48.34	34.37	33.95	81.92	50.42	50.03
9	Na	SO_4	50H	46.42	38.18	76.40	28.45	69.61	43.58	12.09	39.58	49.45	50.39
10	Na	Cl	50H	9.32	47.50	84.36	49.03	39.46	74.76	41.17	26.10	50.69	50.52
11	Na	HCO_3	50H	52.53	11.04	90.28	52.61	12.05	72.73	96.40	38.40	50.05	50.08
12	Na	CO_3	50H	40.84	43.48	74.56	12.77	43.36	17.88	25.60	62.88	50.35	50.10
13	K	SO_4	50H	19.27	55.00	63.89	99.52	80.15	1.67	64.63	42.68	50.36	50.33
14	K	Cl	50H	23.58	45.14	45.69	64.72	1.36	99.83	63.91	40.69	49.42	50.32
15	K	HCO_3	50H	28.92	54.87	15.90	74.41	48.84	33.08	76.14	25.49	49.90	50.37
16	K	CO_3	50H	4.71	49.91	45.35	87.53	34.47	23.41	26.98	61.38	50.11	49.06

Tabla 5.7 Pruebas 75M y 50M

				Concentraciones (mMol/L)								**Probabilidades 7-molar-conc**	
No.	**Catión mayor**	**Anión mayor**	**Prueba**	**Ca**	**Mg**	**Na**	**K**	**SO4**	**Cl**	**HCO3**	**CO3**	**Prob. Catiónica**	**Prob. Aniónica**
1	Ca	SO_4	75M	87.81	9.97	19.53	20.50	93.80	17.74	25.62	2.31	74.78	74.84
2	Ca	Cl	75M	88.76	35.02	24.60	7.47	37.90	110.89	31.34	30.80	74.49	74.84
3	Ca	HCO_3	75M	92.62	9.22	19.54	41.90	38.66	37.00	105.71	22.54	74.50	74.74
4	Ca	CO_3	75M	88.87	17.60	9.36	16.40	8.36	16.18	11.46	97.17	74.83	74.93
5	Mg	SO_4	75M	17.27	87.98	8.92	6.18	89.30	4.59	13.97	14.23	74.86	74.87
6	Mg	Cl	75M	20.24	78.85	24.80	11.60	23.87	112.42	20.15	27.13	74.34	74.94
7	Mg	HCO_3	75M	5.88	86.64	10.22	9.63	14.85	42.80	110.07	11.16	74.88	74.86
8	Mg	CO_3	75M	11.26	79.54	11.81	18.34	4.21	27.43	12.46	81.71	74.65	74.56
9	Na	SO_4	75M	21.70	19.73	135.00	8.17	80.58	13.07	4.74	23.53	75.00	74.58
10	Na	Cl	75M	29.91	24.03	101.61	30.10	27.51	94.45	11.51	39.30	74.79	74.45
11	Na	HCO_3	75M	10.39	31.59	94.21	16.97	18.04	34.52	101.52	11.52	74.75	74.83
12	Na	CO_3	75M	24.00	11.09	108.02	12.82	6.51	16.49	10.63	75.45	74.96	74.56
13	K	SO_4	75M	8.21	27.99	32.89	106.38	86.16	26.10	10.27	1.50	74.88	74.74
14	K	Cl	75M	19.86	23.37	10.62	81.76	8.21	110.86	19.04	16.26	74.48	74.97
15	K	HCO_3	75M	6.59	12.24	33.32	96.35	19.43	5.27	86.07	18.55	74.80	74.69
16	K	CO_3	75M	16.18	12.47	39.20	102.90	11.96	4.28	17.70	76.75	74.82	74.60
1	Ca	SO_4	50M	79.52	10.30	67.80	74.86	66.95	15.87	51.16	60.69	51.37	50.14
2	Ca	Cl	50M	33.38	12.78	26.42	13.51	22.20	39.17	30.97	8.86	50.07	49.19
3	Ca	HCO_3	50M	76.58	70.82	60.74	21.70	58.43	55.75	76.66	64.00	50.36	48.96
4	Ca	CO_3	50M	72.72	57.92	67.68	22.31	56.07	58.10	45.33	67.85	49.85	49.52
5	Mg	SO_4	50M	57.80	68.90	46.52	53.98	78.09	73.49	3.07	60.59	51.70	50.03
6	Mg	Cl	50M	45.31	62.25	48.30	52.95	33.34	71.42	57.78	60.25	49.11	49.59
7	Mg	HCO_3	50M	25.57	48.79	30.55	44.26	36.04	13.36	52.69	42.71	49.89	50.17
8	Mg	CO_3	50M	65.41	75.83	11.15	66.46	70.02	14.77	55.85	74.72	49.87	50.46
9	Na	SO_4	50M	50.37	14.72	69.62	63.30	55.09	39.60	16.98	48.17	50.47	50.40
10	Na	Cl	50M	33.73	57.12	63.79	41.85	27.85	67.23	54.23	55.08	48.65	50.32
11	Na	HCO_3	50M	18.33	75.63	85.59	72.42	54.26	62.90	73.68	50.43	49.83	49.25
12	Na	CO_3	50M	47.36	44.32	69.77	61.29	55.81	31.04	47.23	62.27	50.69	49.50
13	K	SO_4	50M	52.34	66.29	26.06	73.04	69.84	65.03	59.03	36.31	50.89	49.03
14	K	Cl	50M	32.13	36.63	5.08	47.39	38.08	56.24	48.07	4.76	51.72	51.10
15	K	HCO_3	50M	35.36	40.84	19.15	51.92	42.18	27.17	51.25	30.34	50.74	49.14
16	K	CO_3	50M	22.95	83.29	93.45	100.77	63.35	54.90	69.58	77.77	50.52	51.43

Las pruebas de robustez consisten en la simulación de datos a partir de las muestras de prueba 75H, 75M, 50H y 50M. Las pruebas se nombran según el proceso de robustez (sección 4.6.2), con un "40" para indicar propagación de errores al 40% y "C" para indicar un proceso de cambios composicionales, en este caso de un aumento en $Na\ y\ Ca$ al 1% en cada paso. De manera que las pruebas son las siguientes; 75H40, 75M40, 50H40, 50M40, 75HC y 75MC.

La propagación de errores se determinó basándose en los errores analíticos porcentuales reportados en Verma (2013), en donde elementos como SO_4 y Mg presentan errores analíticos de 54% y 92% respectivamente, por lo que errores del 40% pueden ocurrir en mediciones reales de agua. Las pruebas de cambios composicionales (75HC y 75MC) simulan la interacción del agua con minerales como la plagioclasa ($(Na, Ca)(Si, Al)_3O_8$). Estas pruebas de robustez también se documentaron en el **Ápendice A**.

Los resultados de los datos generados en las simulaciones de las pruebas de robustez, se evaluarion en todos los modelos (*7-hlr, 7-molar-conc, greater-molar-conc).* el porcentaje de muestras que permanecieron en las clases iniciales en las pruebas relacionadas con muestras muy interiores (75H40 y 75M40) para cada modelo se presenta en la tabla 5.8 y en la figura 5.1. De la misma manera, para las pruebas de muestras interiores (50H40 y 50M40) se presenta la tabla 5.9 y figura 5.2.

Tabla 5.8 Resultados de pruebas 75H40 y 75M40

		Porcentaje de muestras que permanecieron en la categoría inicial					
No.	**Clase**	**75H40**			**75M40**		
		7-hlr	**7-molar-c.**	**greater-m.-c.**	**7-hlr**	**7-molar-c.**	**greater-m.-c.**
1	*calcium-sulfate*	93.73	93.46	93.05	93.73	93.46	93.05
2	*calcium-chloride*	92.91	91.14	92.00	88.01	86.51	87.19
3	*calcium-bicarbonate*	95.50	95.64	95.68	89.28	86.55	85.92
4	*calcium-carbonate*	96.683	96.32	96.32	96.683	96.32	96.32
5	*magnesium-sulfate*	96.32	96.229	96.047	96.32	96.229	96.047
6	*magnesium-chloride*	92.867	93.321	93.23	91.731	91.186	91.413
7	*magnesium-bicarbonate*	92.549	91.958	91.913	93.23	92.413	92.549
8	*magnesium-carbonate*	92.821	93.003	92.731	94.275	94.275	93.503
9	*sodium-sulfate*	94.821	94.775	94.73	94.821	94.775	94.73
10	*sodium-chloride*	95.366	93.912	94.502	90.186	90.55	90.959
11	*sodium-bicarbonate*	94.775	93.639	94.73	90.595	89.868	90.368
12	*sodium-carbonate*	96.274	95.866	95.638	96.274	95.866	95.638
13	*potassium-sulfate*	97.91	97.274	97.319	93.049	91.322	91.277
14	*potassium-chloride*	94.184	92.095	92.776	94.321	93.458	93.912
15	*potassium-bicarbonate*	96.274	96.184	96.32	92.503	92.14	92.821
16	*potassium-carbonate*	94.366	94.457	94.548	91.186	91.05	90.595
Las celdas grises indican el modelo con el porcentaje de permanencia más alto.							

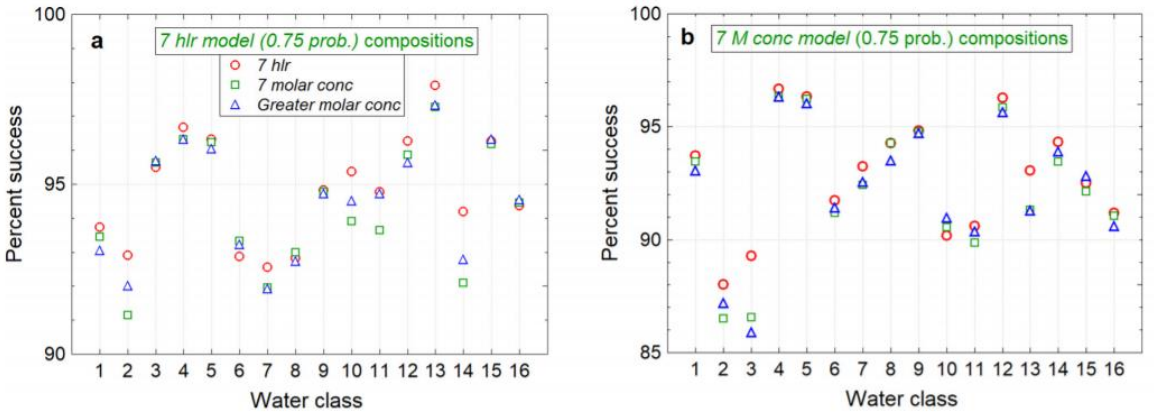


*Figura 5.1 Grafica de resultados de pruebas 75H40 y 75M40 (tomada del **Apéndice A**)*

Tabla 5.9 Resultados de pruebas propagación de errores al 40% en muestras interiores

		Pruebas de robustez					
No.	**Clase**	**50H40**			**50M40**		
		7-hlr	**7-molar-c.**	**greater-m.-c.**	**7-hlr**	**7-molar-c.**	**greater-m.-c.**
1	*calcium-sulfate*	59.428	59.791	58.155	23.444	21.081	19.355
2	*calcium-chloride*	43.526	35.302	35.075	30.986	34.166	32.803
3	*calcium-bicarbonate*	62.108	64.153	63.38	37.165	35.12	35.348
4	*calcium-carbonate*	61.381	62.29	61.199	20.809	21.49	20.173
5	*magnesium-sulfate*	56.156	54.975	54.521	19.718	20.036	19.627
6	*magnesium-chloride*	50.75	41.845	41.89	16.674	18.31	17.765
7	*magnesium-bicarbonate*	48.251	42.163	43.117	19.219	23.58	22.899
8	*magnesium-carbonate*	70.059	71.558	70.014	31.395	31.122	30.622
9	*sodium-sulfate*	59.064	55.157	55.929	22.99	23.807	23.716
10	*sodium-chloride*	72.331	71.331	72.467	44.843	43.344	44.571
11	*sodium-bicarbonate*	44.889	45.07	46.07	17.356	16.538	17.628
12	*sodium-carbonate*	53.885	53.885	55.429	26.715	25.806	26.306
13	*potassium-sulfate*	52.612	45.025	45.116	19.219	21.717	20.672
14	*potassium-chloride*	53.158	45.706	45.525	33.894	31.622	33.394
15	*potassium-bicarbonate*	59.155	54.43	56.929	41.209	35.12	38.664
16	*potassium-carbonate*	56.474	63.017	62.562	13.994	16.129	15.357
Las celdas grises indican el modelo con el porcentaje de permanencia más alto.							

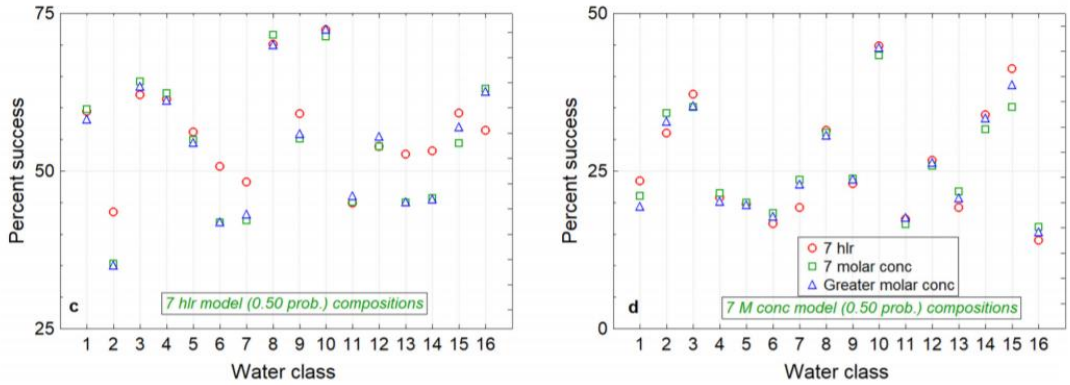


*Figura 5.2 Resultados de pruebas 50H40 y 50M40 (tomada del **Apéndice A**)*

En las pruebas 75HC y 75MC se omiten las categorías que tengan como iones mayoritarios a Ca y Na, pues el incremento de estos elementos en dichas clases solo hace que el proceso de cambio composicional llegue al máximo permitido de 50,000 mg/L. El número de pasos de cada prueba se puede visualizar en la tabla 5.10 y en la figura 5.3.

Tabla 5.10 Resultados en pruebas de cambios composicionales

		Pruebas de robustez					
No.	**Clase**	**75HC**			**75MC**		
		7-hlr	**7-molar-conc**	**greater-m.-c.**	**7-hlr**	**7-molar-conc**	**greater-m.-c.**
5	*magnesium-sulfate*	172	163	163	172	163	163
6	*magnesium-chloride*	169	165	165	106	118	116
7	*magnesium-bicarbonate*	194	190	190	197	216	214
8	*magnesium-carbonate*	190	201	199	163	194	191
13	*potassium-sulfate*	248	238	238	124	117	117
14	*potassium-chloride*	275	229	232	192	140	142
15	*potassium-bicarbonate*	267	261	261	106	105	106
16	*potassium-carbonate*	195	189	189	101	96	96

Las celdas grises indican el modelo que dio más pasos antes de llegar a la condición de paro.

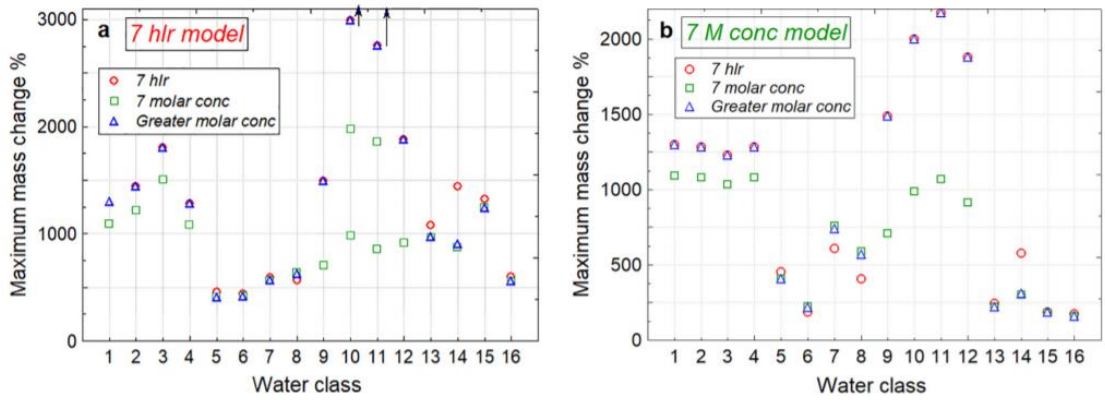


*Figura 5.3 Resultados de pruebas 75HC y 75MC (tomada del **Apéndice A**)*

En la tabla 5.11 se muestran un resumen de las pruebas de robustez para todos los modelos, corresponde a los promedios en las 16 muestras iniciales de cada prueba en cada modelo. El modelo *7-hlr* es el modelo más robusto pues los promedios de porcentaje en las pruebas 75H40, 75M40 y 50H40 y los pasos promedio de las pruebas 75HC y 75MC son mayores en este modelo. La única prueba que no fue favorable para *7-hlr* fue 50M40, que corresponde a la propagación de errores en muestras interiores (cercanas al 50%) en del modelo *7-molar-conc*, esta prueba dio el mismo porcentaje promedio en los 3 modelos.

Tabla 5.11 Resumen de pruebas de robustez			
	Resultados promedio de las pruebas.		
Prueba	***7-hlr***	***7-molar-conc***	***greater-molar-conc***
75H40	94.8%	94.3%	94.5%
75M40	92.9%	92.2%	92.3%
50H40	56.5%	54.1%	54.2%
50M40	26.2%	26.2%	26.2%
75HC	258.0	232.4	253.4
75MC	213.3	191.8	212.3
Las celdas grises indican el modelo con el porcentaje promedio o el número de pasos más alto.			

5.4 Aplicación en muestras de agua subterráneas

En el **Apéndice A** se utilizó el modelo *7-hlr* (el modelo más robusto) en un caso de aplicación real, estas muestras fueron originalmente clasificadas utilizando el diagrama de Hill-Piper (figura 5.4) y corresponden a muestras agua subterránea en el estado de Tamil Nadu al sur de la India (Tabla 5.12; Kumar, 2013).

Tabla 5.12 Muestras de agua subterránea (Kumar, 2013)								
	Concentraciones en mg/L							
Id	Ca^{2+}	Mg^{2+}	Na^{+}	K^{+}	$SO_4{}^{-2}$	Cl^{-}	$HCO_3{}^{-}$	$CO_3{}^{2-}$
GW1	40	36	258	4	132	188	336	1
GW2	38	53	161	86	12	138	573	1
GW3	60	100	115	18	194	284	165	1
GW4	40	86	182	1	187	273	110	1
GW5	34	102	161	113	29	372	329	24
GW6	22	80	133	3	14	184	311	24
GW7	50	52	145	26	78	213	140	1
GW8	32	32	113	6	6	53	361	14
GW9	36	74	107	10	52	142	348	1
GW10	28	87	322	34	58	369	580	1
GW11	30	29	14	2	7	21	183	2
GW12	20	77	271	4	91	230	427	42

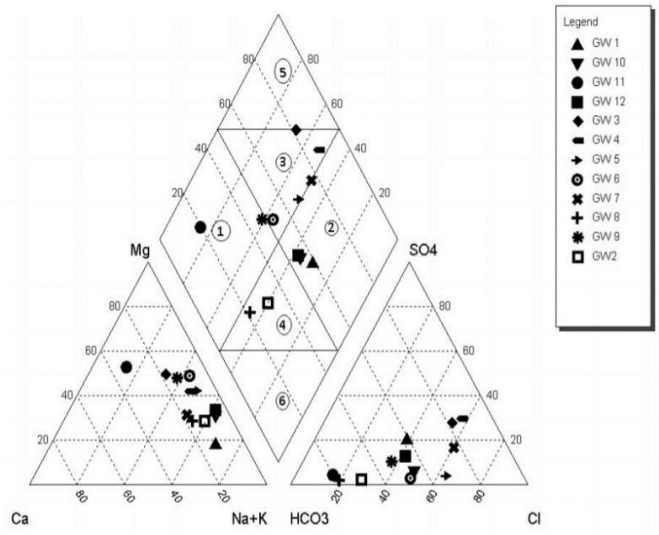


Figura 5.4 Diagrama de Hill-Piper de muestras de muestras de agua subterránea (Kumar, 2013)

En la tabla 5.13 se presentan los tipos de agua obtenidos para cada muestra en el diagrama de Hill-Piper y los obtenidos con el modelo *7-hlr* en su tipo básico e híbrido. Las muestras GW1, GW4, GW9 y GW11 tienen problemas de ambigüedad en el tipo de agua en el diagrama de Hill-Piper (Piper, 1944), por ejemplo, la muestra GW1, en su representación aniónica (diagrama ternario derecho), la muestra se encuentra claramente más cerca de la esquina HCO_3, incluso, en unidades de mEq/L (unidades del diagrama Hill-Piper), el anión mayoritario es HCO_3, sin embargo, el tipo de agua reportado en Kumar (2013) es la zona 2, que de acuerdo al autor corresponde al tipo $Na\text{-}Cl$.

Las predicciones, tanto básicas como híbridas del modelo *7-hlr*, son mucho más coherentes, en especial la híbrida, que algunas veces propone un tipo aniónico $HCO_3\text{-}Cl$, sugiriendo a que las concentraciones de estos elementos son parecidas. En la figura 5.5 se presentan los gráficos generados por el programa *WaterMClaSys_lda* con las funciones discriminantes del modelo ensamblado *7-hlr* para las muestras de agua subterránea en Kumar (2013).

Tabla 5.13 Comparación de tipos de agua en muestras reales

	Herramienta de clasificación		
id	**Hill-Piper**	***7-hlr* (básico)**	***7-hlr* (híbrido)**
GW1	$\boldsymbol{Na\text{-}Cl}$	*sodium-bicarbonate*	*sodium bicarbonate-chloride*
GW2	$Ca\text{-}Na\text{-}HCO_3$	*sodium-bicarbonate*	*sodium-potassium bicarbonate-chloride*
GW3	$Ca - Mg - Cl$	*sodium-chloride*	*sodium-magnesium chloride*
GW4	$\boldsymbol{Ca\text{-}Mg\text{-}Cl}$	*sodium-chloride*	*sodium-magnesium chloride*
GW5	$Ca\text{-}Mg\text{-}Cl$	*sodium-chloride*	*sodium chloride-bicarbonate*
GW6	$Ca\text{-}Mg\text{-}Cl$	*sodium-bicarbonate*	*sodium-magnesium bicarbonate-chloride*
GW7	$Na\text{-}Cl$	*sodium-chloride*	*sodium chloride-bicarbonate*
GW8	$Ca\text{-}Na\text{-}HCO_3$	*sodium-bicarbonate*	*sodium bicarbonate-chloride*
GW9	$\mathbf{Ca\text{-}HCO_3}$	*sodium-bicarbonate*	*sodium-magnesium bicarbonate-chloride*
GW10	$Na\text{-}Cl$	*sodium-bicarbonate*	*sodium bicarbonate-chloride*
GW11	$\boldsymbol{Ca\text{-}HCO_3}$	*magnesium-bicarbonate*	*magnesium-sodium bicarbonate-chloride*
GW12	$Na\text{-}Cl$	*sodium-bicarbonate*	*sodium-magnesium bicarbonate-chloride*

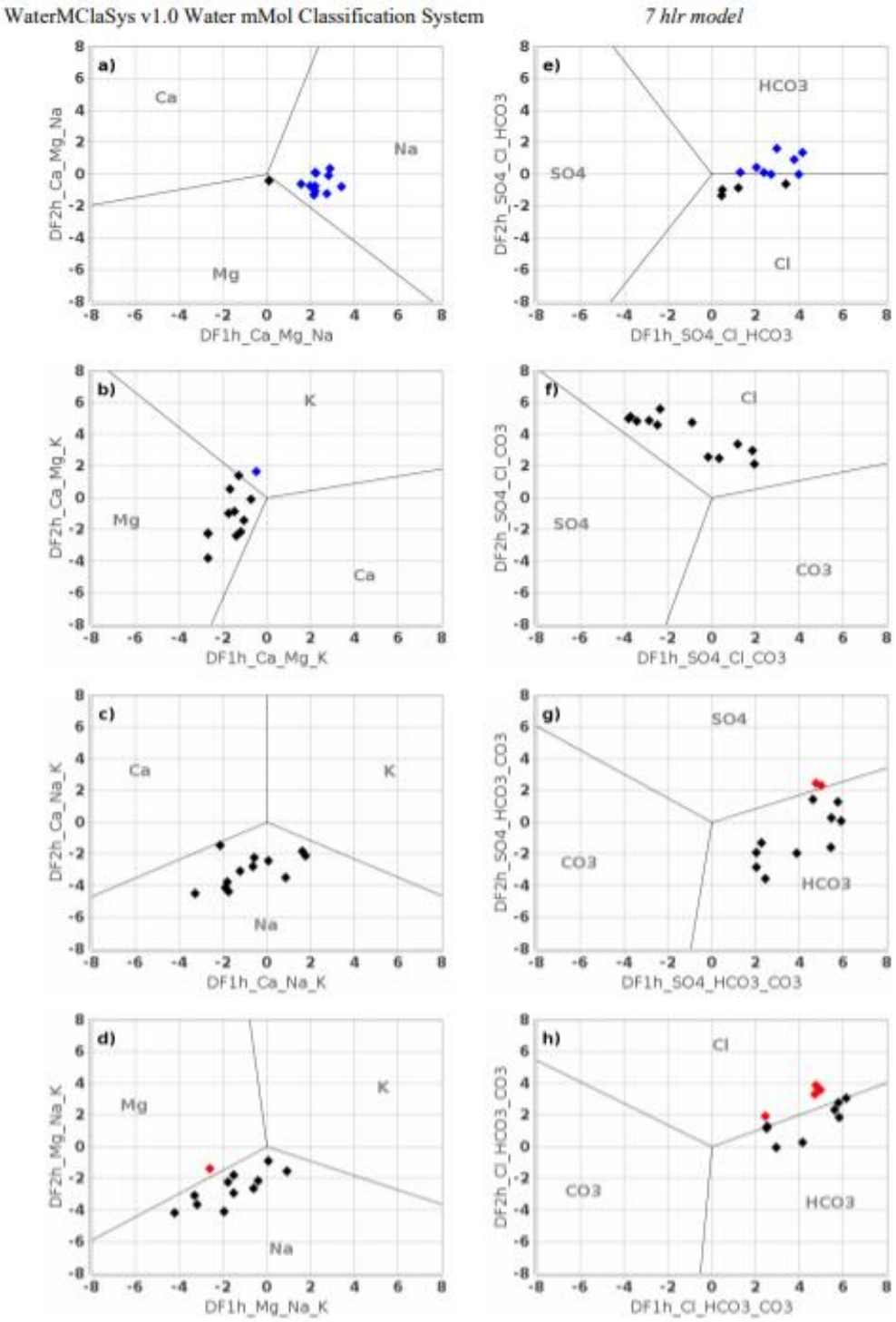


Figura 5.5 Diagramas WaterMClaSys_lda para muestras de agua subterráneas (Kumar, 2013)

Capítulo 6: Conclusiones, trabajos adicionales y futuros

6.1 Conclusiones

Se simularon dos bases de datos para entrenamiento y validación externa con la metodología propuesta por Verma y Quiroz-Ruiz (2006), este procedimiento permitió entrenar y evaluar los modelos de clasificación con bases de datos con representatividad en las 16 clases básicas determinadas por los iones mayoritarios. Estas bases de datos pasaron por el software DOMuDaF (Verma *et al.*, 2016) para separar datos discordantes con el fin de utilizar correctamente la técnica de análisis discriminante lineal y análisis canónico (que supone la multinormalidad de las clases).

Se entrenaron dos modelos de clasificación de agua (*7-hlr* y *7-molar-conc*), ambos modelos tienen buena precisión para poder discriminar muestras de agua. Se visualizó con un caso de aplicación (Kumar, 2013; **Apéndice A**) que existen inconsistencias en la forma en la que el diagrama de Hill-Piper (Piper, 1944) clasifica muestras de agua como en los ejemplos GW1, GW4, GW9 y GW11 en las que directamente, la clasificación propuesta por el diagrama de Hill-Piper es confusa y poco coherente debido a la división en el diamante para nombrar los tipos de agua. Este problema se resuelve en la propuesta de esta tesis gracias a que los tipos de agua están directamente relacionados con los iones mayoritarios. El traslape de muestras en el diagrama de Hill-Piper evita que sea posible diferenciar, por ejemplo, concentraciones de Ca y Mg en la parte catiónica o SO_4 y Cl en la parte aniónica, estas diferencias son explícitas en los modelos *7-hlr* y *7-molar-conc*.

Una cualidad importante de los nuevos modelos es que no dependen del uso de diagramas ternarios ni de las relaciones porcentuales, evitando así la amplificación-reducción de errores y la violación de la aleatoriedad en las variables composicionales (Verma, 2015). Estos errores existen en otras propuestas de clasificación ya que usan diagramas ternarios (Piper, 1944; Durov, 1948; Handa, 1965; Giggenbach, 1988). Sin embargo, aún hay funcionalidades que los modelos *7-hlr* y *7-molar-conc* no contemplan, como el análisis de salinidad del diagrama modificado de Hill-Piper y el análisis de condiciones de equilibrio del diagrama de Giggenbach, por lo que el uso de los nuevos modelos es primordialmente la de asignar tipos de agua a muestras desconocidas.

La capacidad de los modelos *7-hlr* y *7-molar-conc* de poder identificar tipos híbridos gracias al cálculo interno de probabilidades añade resolución a la cantidad de tipos de agua que se pueden identificar hasta un número de 256 tipos de agua, esto permite que muestras que los modelos pudieran identificar incorrectamente (porque las concentraciones de dos elementos son muy parecidas), se puedan interpretar como un tipo de agua híbrido.

Las pruebas de robustez rebelaron que el modelo *7-hlr* es el modelo más robusto debido a que puede soportar más perturbaciones como propagación de errores y cambios

composicionales antes de que el modelo colapse y de un resultado incorrecto. *7-hlr* entonces, es el modelo recomendado. Estas pruebas de robustez y clasificación se hicieron con el programa de libre uso *WaterMClaSys_lda,* en donde se pueden ocupar los modelos de clasificación propuestos.

6.2 Trabajos adicionales

Adicionalmente se trabajó en el artículo del **Apéndice B**, publicado la revista *Journal of Hydrology* en donde el modelo *7-hlr* se compara con otras técnicas de *Machine Learning* como *Categorical Boosting* y *Support Vector Machines*, a continuación, se presenta el *Abstract* de dicho trabajo.

Abstract

Para encontrar un modelo mejorado para la clasificación de agua, propusimos 4 modelos utilizando *Categorial Boosting* (CatBoost) y *Support Vector Machines (SVM)* con 3 *kernels*: linear, polinomial y de función radial. Estos modelos se compararon con el modelo *7-hlr* basado en análisis discriminante lineal y análisis canónico. Una base de datos de entrenamiento (50,000 muestras) y otra base de datos de validación independiente (8,000 muestras) con un balance de cargas iónico adecuado de 4 cationes ($Ca,\ Mg,\ Na,\ K$) y 4 aniones ($SO_4,\ Cl,\ HCO_3,\ CO_3$) se generaron a través de simulación Monte Carlo. Las 16 clases iniciales fueron asignadas por con el catión y anión mayoritario (criterio *GMC* o *greater-molar-conc*). Siete datos transformados *hlr* se usaron como variables de entrenamiento y de validación externa para los nuevos modelos de clasificación. Estos modelos generan probabilidades de clasificación para cada categoría, por lo que se pueden determinar hasta 256 tipos híbridos. El modelo WClassCB mostró los mejores valores de precisión en el conjunto de datos de entrenamiento. Sin embargo, el modelo WClassVL es el procedimiento recomendado porque generaliza mejor que los otros modelos en el conjunto de validación externa. Los nuevos modelos se desempeñan mejor que el modelo WClassHLR con hasta 7% de mejora. El uso de todos los modelos (WClassHLR, WClassCB, WClassVL, WClassVP, WClassVR) se ilustra con 4 aplicaciones en muestras de agua subterránea de la India y Nigeria. Todos los modelos tienen dificultades clasificando muestras reales cuando hay más de un catión o anión mayoritario, pero pueden recuperar la clasificación sugiriendo un tipo de agua híbrido. Se desarrolló el programa WaterClaSys_ML para la aplicación de estos modelos.

6.3 Trabajos futuros

En el **Apéndice A** se desarrolló una propuesta para añadir más dimensiones a la clasificación de agua añadiendo variables adicionales como salinidad, pH, As, Cr, Li, Fe, Al, SiO_2, Pb, Se, Hg, P, S, NH_3, NH_4, NO_x, PO_x, H_xC_x. Parámetros como *TDS* (total de sólidos disueltos) se presentarán automáticamente junto a otros parámetros que tuviesen importancia hidroquímica. Para esta propuesta es necesario ampliar la metodología de simulación Monte Carlo para involucrar las nuevas variables y entrenar un nuevo modelo de clasificación para poder aplicar esta clasificación de agua a más problemas de interés ambiental.

Referencias

Bayram, A. F., & Gultekin, S. S. (2010). Classifying of the Simav geothermal waters with artificial neural network method. *In Proceedings World Geothermal Congress, Bali Indonesia*, 25-29.

Agrawal, S., Guevara, M., & Verma, S. (2004). Discriminant Analysis Applied to Establish Major-Element. *International Geology Review, 46*, 575–594.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London, UK: Chapman and Hall.

Chadha, D. K. (1999). A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. *Hydrogeology Journal, 7(5)*, 431-439.

Doornik, J. (2005). *An Improved Ziggurat Method to Generate Normal Random Samples.* Oxford, UK: University of Oxford.

Durov, S. A. (1948). Natural waters and graphic representation of their composition. *In Dokl Akad Nauk SSSR, 59(3)*, 87-90.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35(3)*, 279-300. doi:https://doi.org/10.1023/A:1023818214614

Elhag, A. B. (2017). New diagram useful for classification of groundwater quality. *Journal of Geology & Geophysics, 6*, 279. doi: 10.4172/2381-8719.1000279

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems* (Second ed.). Canada: O'Reilly Media.

Giggenbach, W. F. (1988). Geothermal solute equilibria. Derivation of Na-K-Mg-Ca geoindicators. *Geochimica et Cosmochimica Acta, 52*, 2749-2765. doi:https://doi.org/10.1016/0016-7037(88)90143-3

Güler, C., Thyne, G. D., McCray, J. E., & Turner, A. K. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal, 10*, 455–474. doi:https://doi.org/10.1007/s10040-002-0196-6

Handa, B. K. (1965). Modified Hill-Piper diagram for classification of groundwater in arid and semi-arid regions. *Geochemical Society of India Bulletin, 1*, 20-24.

Iwalewa, T., Makkawi , M., Elamin , A., & Al-Shaibani , A. (2013). Groundwater Management Case Study, Eastern Saudi Arabia: Part II – Solute Transport Simulation and Hydrochemistry. *European Journal of Scientific Research, 109(4)*, 650-667.

Kumar, P. J. (2013). *Interpretation of groundwater chemistry using piper and chadha´s diagrams: a comparative study from perambalur taluk.* Elixir Geoscience 54, 12208- 12211.

Marsaglia, G. (1968). Random numbers fall mainly in the planes. *National Academy of Science Proccesdings, 61(1)*, 25-28.

Marsaglia, G., & Bray, T. A. (1964). A convenient method for generating normal variables. *Society for Industrial and Applied Mathematics, SIAM Review, 6(3)*, 260–264.

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister; A 623-dimensionally equidistributed uniform pseudorandom number generator:. *Association for Computing Machinery, ACM Transactions of Modeling and Computer Simulations, 8(1)*, 3-30.

Mustafa, E., Mohamed, A., Ehab, Z., & Shouakar, S. (2019). Hydrochemical and stable isotopes indicators for detecting sources of groundwater contamination close to Bahr El-Baqar drain, eastern Nile Delta, Egypt. *Water Science, 33(1)*, 54-64.

Nicholson, K. (1933). *Geothermal Fluids: Chemistry and Exploration Techniques.* Berlin Heidelberg, Germany: Springer-Verlag.

Pérez-Espinosa, R., Kailasa, P., & Hernández-Campos, F. J. (2019). CCWater - A computer program for chemical classification of geothermal waters. *Geosciences Journal, 23*, 621-635. doi:https://doi.org/10.1007/s12303-018-0064-6

Piper, A. M. (1944). A graphic procedure in the geochemical interpretation of water analyses. *Transactions American Geophysical Union, 25*, 914-923.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogoush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Neural Information Processing Systems*, 6637-6647.

Ravikumar, P., Prakash, K., & Somashekar, R. (2015). A comparative study on usage of Durov and Piper diagrams to interpret hydrohemical processes in groundwater from SRLIS. *Earth Science - Elixir International Journal, 80*, 31073 - 31077.

Romano, P., & Liotta, M. (2020). Using and abusing Giggenbach ternary Na-K-Mg diagram. *Chemical Geology, 541*, 119577. doi:https://doi.org/10.1016/j.chemgeo.2020.119577

Schoeller, H. (1955). Géochemie des eaux souterraines. *Revue de l'Institut Francais du Petrole, 10*, 230-244.

Shelton, J. L., Englea, M. A., Buccianti, A., & Blondes, M. S. (2018). The isometric log-ratio (ilr)-ion plot: a proposed alternative to the Piper diagram. *Journal of Geochemical Exploration, 190*, 130-141. doi:https://doi.org/10.1016/j.gexplo.2018.03.003

Stiff Jr., H. A. (1951). The interpretation of chemical water analysis by means of patterns. *Journal of petroleum technology, 3*, 15-17.

Tharwat, A., Gaber, T., Ibrahin, A., & Ella Hassanien, A. (2017). *Linear Discriminant Analysis: A detailed tutorial.* Alemania: Department of Computer Science and Engineering, Frankfurt University of Applied Sciences.

Thomas, D., Luk, W., Leong, P., & Villasenor, J. (2007). Gaussian random number generators. *ACM Computing Surveys, 39(4)*.

Verma, M. P. (2013). IAEA inter-laboratory comparisons of geothermal water chemistry: critiques on analytical uncertainty, accuracy, and geothermal reservoir modelling of Los Azufres, Mexico. *Journal of Iberian Geology, 39* , 57-72.

Verma, S. P. (2015). Monte Carlo comparison of conventional ternary diagrams with new log-ratio bivariate diagrams and an example of tectonic discrimination. *Geochemical Journal, 49*, 393-412.

Verma, S. P., & Agrawal, S. (2011). New tectonic discrimination diagrams for basic and ultrabasic volcanic. *Revista Mexicana de Ciencias Geológicas, 28(1)* , 24-44.

Verma, S. P., & Quiroz-Ruiz, A. (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista Mexicana de Ciencias Geológicas, 23*, 133-161.

Verma, S. P., Rivera-Gomez, M. A., Diáz-González, L., & Quiroz-Ruiz, A. (2016). Log-ratio transformed major-element based multidimensional classification for altered high-Mg igneous rocks. *Geochemistry, Geophysics, Geosystems, 17*, 4955-4972. doi:https://doi.org/10.1002/2016GC006652

Verma, S., Rivera Gómez, M., Diáz González, L., Pandarinath, K., Amezcua Valdez, A., Rosales RIvera, M., . . . Armstrong Altrin, J. (2017). Multidimensional classification of magma types for altered igneous rocks and application to their tectonomagmatic discrimination and igneous provenance of siliciclastic sediments. *Lithos, 278- 281*, 321- 330.

Apéndice A. A statistically coherent robust multidimensional classification scheme for water

Science of the Total Environment 750 (2021) 141704

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A statistically coherent robust multidimensional classification scheme for water

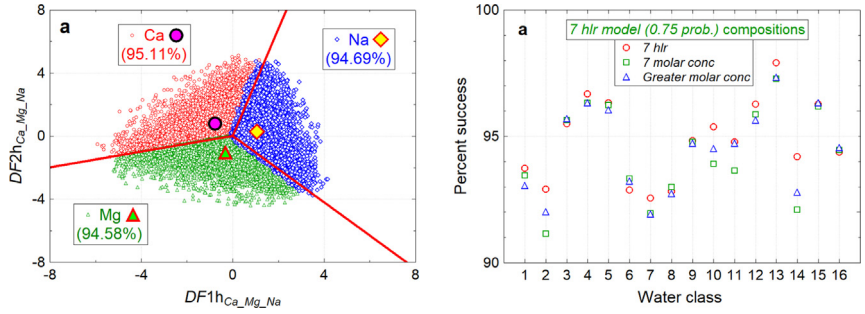
Surendra P. Verma [a,*], Oscar Alejandro Uscanga-Junco [b], Lorena Díaz-González [c]

[a] Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Priv. Xochicalco s/no., Col. Centro, Apartado Postal 34, Temixco, Mor. 62580, Mexico
[b] Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma de Estado de Morelos, Cuernavaca, Morelos 62209, Mexico
[c] Centro de Investigación en Ciencias, Universidad Autónoma de Estado de Morelos, Cuernavaca, Morelos 62209, Mexico

HIGHLIGHTS

- The first multidimensional scheme trained from the nomenclature of 16 water types
- Probability-based concept used for the first time for water classification
- Comparison of the multidimensional molar and log-ratio classifications
- The most robust *7 hlr model* capable of the nomenclature of 256 water types

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 May 2020
Received in revised form 13 August 2020
Accepted 13 August 2020
Available online 15 August 2020

Editor: José Virgílio Cruz

Keywords:
Ternary diagrams
Molar concentrations
Log-ratio molar concentrations
New water nomenclature

ABSTRACT

Serious limitations of the existing water classification schemes prompted us to propose a new statistically coherent water nomenclature system. An extensive database of ionic charge-balanced concentrations of 8 elements (4 cations Ca, Mg, Na, and K; and 4 anions SO_4, Cl, HCO_3, and CO_3), in 46,292 multivariate outlier-free simulated samples, was used for training the multidimensional classification system. The initial assignment for 16 classes was achieved from the greater molar concentration concept of each cation and anion, called the *Greater molar conc* model. Seven hybrid log-ratios (hlr) from 8 elemental concentrations were used for linear discriminant analysis (LDA) and canonical analysis to propose 16 multidimensional discriminant functions from the *7 hlr model*. The LDA and canonical analysis were also performed on the initial molar concentrations of 7 elements, without any log-transformation, which was designated as the *7 M conc model*. The robustness of these three classification systems (*7 hlr*, *7 M conc*, and *Greater molar conc*) was tested against analytical uncertainty propagation and mineral-water interaction effects. The *7 hlr model*, due to its higher robustness, was considered as the best option for the nomenclature of the 16 types of water. From the probability concept, it was possible to identify hybrid water types, along with the basic or primary types of water. Our water classification scheme (*7 hlr model* under the "basic+hybrid" option) can classify as many as 256 different classes of water. Due to the clearly high complexity of the proposed classification scheme, we developed a new online computer program *WaterMClaSys_LDA* (Water Molar Classification System from Linear Discriminant Analysis) available at our web portal http://tlaloc.ier.unam.mx, for use by anyone after registration and log-in. The usefulness of the new classification scheme is illustrated by applications to groundwater, lake water, and geothermal water samples from South India, Mongolia, and western Turkey, respectively.

© 2020 Elsevier B.V. All rights reserved.

* Corresponding author.
E-mail addresses: spv@ier.unam.mx (S.P. Verma), usju5296@gmail.com (O.A. Uscanga-Junco), ldg@uaem.mx (L. Díaz-González).

https://doi.org/10.1016/j.scitotenv.2020.141704
0048-9697/© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Sheth et al. (2002), in their work on the "calc-alkaline series" nomenclature of igneous rocks, reproduced, from Anderson (1999), this quote of Abbé de Condillac "*… we cannot improve the language of any science without at the same time improving the science itself; neither can we, on the other hand, improve a science, without improving the language or nomenclature which belongs to it.* – Abbé de Condillac".

Similar statements can be traced throughout the human history. Thus, the nomenclature in any science is of fundamental importance not only for efficient human communication but also for identifying or quantifying unusual time- or space-related changes, such as in environmental science or mining research. The water research is, of course, no exception.

There have been many attempts to propose a nomenclature scheme for water and to present relevant computer programs (e.g., Chadha, 1999; Romani, 1981; Lee, 1998; Rao, 1998; Güler et al., 2002; Ahmad et al., 2003; Ray and Mukherjee, 2008; Giménez-Forcada, 2010; Elhag, 2016; Teng et al., 2016; Pérez-Espinosa et al., 2019), although the Hill-Piper diagram (Hill, 1940; Piper, 1944) still constitutes the most frequently used nomenclature with over 3800 citations according to Google Scholar (accessed on 13th February 2020). At several occasions (e.g., Xue et al., 2019), the "famous" diagram is simply cited as the Piper diagram, without even providing the relevant reference.

The Hill-Piper classification is based mainly on two ternary diagrams constructed from cations and anions normalized concentration data in mg/L (Ca-Mg-(Na + K) for cations; and SO_4-Cl-(HCO_3 + CO_3) for anions) and a related diamond field, which could provide the nomenclature of only 5 water types. Durov (1948) used a rectangular field, instead of a diamond field. A modified Hill-Piper diagram, capable of providing irrigation water classification, was proposed by Handa (1965) by incorporating salinity in the chemical variables. Romani (1981) proposed a modification of the Hill-Piper ternary diagrams (equilateral triangles) in terms of right-angled isosceles triangles, stating that this change represented an advantage in that the resulting central field becomes a square instead of a diamond. Another modification of the Hill-Piper diagrams consisted of the Banaga diagrams, which were compared with the traditional diagrams (Elhag, 2016). The visualization of concentration patterns was documented by Stiff (1951). Similarly, modifications of the Durov diagrams (Durov, 1948) have also been proposed (e.g., Lloyd, 1965; Chadha, 1999). Multi-rectangular diagrams based on milliequivalent/l (meq/L) normalized concentrations are also available (Ahmad et al., 2003; Ray and Mukherjee, 2008).

Kemp (1971) proposed acid-base molar composition diagrams and argued in favor of the total dissolved solids as a parameter of great importance for the classification of waters. Shterev (1973) tentatively proposed water classification based on water-rock interactions in space and time, which involved more chemical components than traditionally used (4 cations and 4 anions) for the water nomenclature. D'Amore et al. (1983) defined 6 new parameters from 7 chemical components of water samples from central Italy, which were used to complement the Hill-Piper diagram. Stuyfzand (1989) introduced a classification scheme based on the Cl-content, alkalinity, the most important cation and anion and a parameter [Na + K + Mg]-corrected for sea salt and proposed two ternary diagrams, one each for cations and anions.

Several researchers (e.g., Lee, 1998; Rao, 1998; Al-Bassam and Khalil, 2012; Teng et al., 2016; Pérez-Espinosa et al., 2019) presented computer programs to facilitate the use of these graphical techniques and establish the water nomenclature. A combination of the graphical procedure and multivariate statistical techniques such as principal component analysis has also been attempted (Güler et al., 2002).

However, false trends (Butler, 1979) and serious consequences of distortion and amplification-reduction of analytical errors in ternary diagrams, caused by closure and constant sum problems (Chayes, 1960, 1971; Aitchison, 1986), documented by Verma (2012a, 2015), were not considered in any of these studies. As explained by Butler (1979), most workers erroneously quoted the mixing trends in ternary diagrams as straight lines. These shortcomings of ternary diagrams were ignored by all workers, including Shelton et al. (2018), for water classification and process analysis. In fact, if the analytical errors, inevitably present in all experimental work, were taken in account (Verma, 2012a, 2015), the shape of the mixing curve in a ternary diagram will be even more complex, with varying thickness along the curve.

These problems can be overcome only by abandoning the ternary diagrams and opening the closure through log-ratio transformations (Verma, 2015, 2020a). Therefore, it is time to replace the ternary diagrams for water classification and resort to the use of log-ratio transformations and linear discriminant analysis (LDA) along with canonical analysis (Verma, 2020a).

Isometric log-ratio transformation was used by Shelton et al. (2018), but only the transformation in terms of new plots was attempted without increasing the dimensions of the solution. The emphasis of these authors was in terms of the comparison of the ternary and new ilr-based plots, for which the findings of Butler (1979) and Verma (2012a, 2015) were ignored. Shelton et al. (2018) also followed the combination of alkalis ($Na^+ + K^+$) and HCO_3^- and CO_3^{2-} ($HCO_3^- + CO_3^{2-}$) as done by all other workers. It was also not clear how the class boundaries in the new plots were constructed; the percent success values for their replacement diagrams were also not reported by Shelton et al. (2018).

There is no special geochemical reason to combine the two alkali cations and two anions for their use in ternary diagrams and leave the two alkaline earth elements as separate entities. Why were the two alkaline earth elements (Ca and Mg) not combined in any classification, leaving Na and K as separate entities? Further, why should we combine any of them after all? Therefore, it would be worthwhile to explore a totally new classification scheme without combining any of these 8 ionic species.

Another point that needs attention is that we always think in terms of element concentrations (mg/L or μg/L) for water classification, which is a natural consequence of the way the water analysis is commonly expressed in terms of the mass/volume unit, and not as the number of atoms of the chemical elements. Would it not be better to name the water class after the highest number of atoms of a cation or anion, instead of concentrations? Our answer is yes; hence, we used molar concentrations to propose a new multidimensional classification scheme for water nomenclature as well as to show its greater robustness against analytical errors and chemical modifications.

At this stage, we refrained from using molal concentrations (e.g., milliMoles/kg, i.e., mM/kg) although it would be much better to do so. The reason for not shifting to mass/mass units was that it is customary, at present, to measure and report water analysis in mass/volume units. In future, it may be useful to shift from molar (milliMoles/L, i.e., mM/L) to molal (mM/kg) units, because the molal analysis would be more precise and accurate than the molar determinations (see the discussion in Verma, 2020a). For example, the weighing errors for a liquid in a calibrated precision balance could be around 0.01% or even 0.001%, whereas the errors of volume measurements in a volumetric flask would be much higher (certainly >0.01%). The dilutions of liquids inevitably required, such as those used for calibrations, would be better handled by mass/mass units. Therefore, it should be better to move from molar to molal units to reduce the final uncertainties in water analysis.

Our multidimensional proposal is based on the hybrid log-ratios (hlr; Verma, 2020a; Eqs. S1 to S4 in the supplementary file) of the molar concentrations (mM/L) of 4 cations and 4 anions and the use of all 8 ions (*7 hlr model* in terms of 7 hlr transformations) for a robust multidimensional solution through the LDA and canonical analysis. Similarly, the molar concentrations, without log-ratio transformations, were also used for the multidimensional solution in terms of the *7 M conc model*. However, in the LDA and canonical analysis, it was not possible to use all 8 concentration variables. The original classification scheme based on the highest molar concentration (*Greater molar conc*

model) was also evaluated. Robustness schemes and probability criteria were presented to decide on the best scheme(s) for the water nomenclature into 16 classes. Further probability-based criteria were put forth to decipher up to 256 water types from the combination of the basic and hybrid water types, being an innovative approach for the water nomenclature.

2. Database and statistical methodology

Our innovative procedure commenced with the establishment of a representative database of 8 compositional variables in 16 main types of water compositions from Monte Carlo simulations (Law and Kelton, 2000), following the precise and accurate method of Verma and Quiroz-Ruiz (2006) for generating U(0,1) independently for each element. The primary water class nomenclature was achieved from the highest anion and cation molar concentrations (the *Greater molar conc model*).

The 16 assigned classes represent the cross-combinations of 4 cation and 4 anion classes based on molar concentrations. The mM/L concentration data were converted to mEq/L and adjusted for charge balance of better than ±0.00004% (i.e., a perfect balance). Because we wanted to apply the LDA and canonical analysis to all 8 transformed compositions, i.e., 7 hlr variables and the LDA requires that the individual classes be multi-normally distributed (Morrison, 1990) in terms of these 7 unitless variables, the multi-normality of 16 classes was achieved individually from the DOMuDaF computer program (Verma et al., 2016). The statistical characteristics of the multivariate outlier-free data (number of samples n, centroid or mean $\bar{\boldsymbol{x}}$ and 99% uncertainty U_{99} values) for the 16 classes are summarized in Table 1. The concentration values were simply truncated to one decimal place before listing them in Table 1. A total of 46,292 multivariate discordant outlier free, ionic charge-balanced analyses were available for training the classification system. Note that, prior to undertaking the LDA and canonical analysis, the training classes were all set to similar sizes, free from multivariate discordant outliers, and balanced in their sizes; therefore, our results should be considered as with little bias and insignificant overfitting.

The LDA and canonical analysis were also applied to the 7 hybrid log-ratios (hlr_2 to hlr_8; hlr_1 was absent because the first classifying element Ca can only occupy the numerator; see eqs. *S1* and *S2*) as well as to 7 (out of 8; LDA was not applicable to all of them) original molar concentrations (Ca to HCO_3, without CO_3 for cations classification; and Mg to CO_3, without Ca for anions classification).

The main idea for proposing the multidimensional classification of 16 types of water was to achieve independently the classification of 4 cation and 4 anion groups. However, for 4 classes, the LDA would provide 3 (one less than the number of classes) discriminant functions, requiring a three-dimensional diagram to visualize them. It would not be appropriate to use only 2 of these 3 functions, because the totality of variance can only be explained by all 3 discriminant functions. Therefore, it is advisable to evaluate 3 classes at a time, which would require making 4 sets of 3 classes each. For example, for the 4 cation classes (Ca, Mg, Na, and K), we constructed 4 sets as follows: Ca, Mg, and Na; Ca, Mg, and K; Ca, Na, and K; and Mg, Na, and K. The anion classification (SO_4, Cl, and HCO_3, and CO_3) was similarly achieved in 4 sets. The 16 water types could thus be obtained by cross combinations.

When 3 groups or classes are evaluated from the LDA, 2 discriminant functions are obtained. The discriminant function equations can be expressed in its general form as follows (Eq. (1)):

$$DF\alpha h_\tau = (hc_2 \times hlr_2) + (hc_3 \times hlr_3) + (hc_4 \times hlr_4) + (hc_5 \times hlr_5) + (hc_6 \times hlr_6) + (hc_7 \times hlr_7) + (hc_8 \times hlr_8) + hc_0 \quad (1)$$

where α takes values of 1 and 2; τ represents the diagram type; h stands for hybrid log-ratios (hlr); and coefficients hc_2 to hc_8 and hc_0 are listed in the first upper half of Table S1.

As an example, the equations for the classification of the Ca-Mg-Na plot would be as follows (Eqs. (2) and (3)):

$$DF1h_{Ca\_Mg\_Na} = (-0.48404 \times hlr_2) + (-1.84647 \times hlr_3) + (-0.75526 \times hlr_4) + (-0.55609 \times hlr_5) + (-0.59135 \times hlr_6) + (-0.60414 \times hlr_7) + (-0.53193 \times hlr_8) + 0.01956 \quad (2)$$

$$DF2h_{Ca\_Mg\_Na} = (1.98732 \times hlr_2) + (0.58901 \times hlr_3) + (0.85724 \times hlr_4) + (0.92387 \times hlr_5) + (0.90036 \times hlr_6) + (0.89207 \times hlr_7) + (0.92290 \times hlr_8) + 0.02301 \quad (3)$$

The LDA also provided the respective centroids (Table S2) for each set of discriminant function equations and the respective plots, although the latter are not really required for classification. The classification is based on probability calculations as originally suggested by Agrawal (1999) and later by Verma (2012b, 2020a) who provided detailed explanation on the probability concept. Nevertheless, because the scientists are used to visualize the data in diagrams, we will illustrate our methodology also with plots. The discriminant function required for the diagrams were calculated from DF1 and DF2 type equations. For example, the $DF1h_{Ca\_Mg\_Na} - DF2h_{Ca\_Mg\_Na}$ functions required for the classification of Ca, Mg and Na (Ca-Mg-Na diagram) can be calculated from Eqs. (2) and (3). For the other diagrams of three at a time (Ca-Mg-K, Ca-Na-K, and Mg-Na-K), the discriminant functions can be similarly calculated (see coefficients in the first upper half of Table S1).

One may argue against the use of log-ratio transformations (measurement unit-free variables), especially because the cation and anion concentration data in water are not strictly a closed constant sum system, contrary to the rock compositional data. Therefore, instead of the above-mentioned scheme, we could devise a scheme that is based directly on all molar concentrations. Therefore, an alternative scheme was designed from first making the dataset free from multivariate outliers in terms of molar concentrations (DOMuDaF; Verma et al., 2016) and then applying the LDA and canonical analysis to the censored (outlier-free) database.

Finally, we may also decide to use the initial class-assignment nomenclature scheme based on the highest molar concentrations of a cation and an anion. Thus, in practice, we have three schemes (see Table 1 for the discrimination variables): LDA and Canonical analysis of 7 hlr variables (the *7 hlr model*); LDA and canonical analysis of 7 M concentrations (the *7 M conc model*); and the simple identification of the greater molar concentration variables (*Greater molar conc model*).

As for the *7 hlr model*, only 7 M concentrations could be used in the multidimensional solution as the *7 M conc model*. For cations, CO_3 was excluded from the concentration data, i.e., Ca, Mg, Na, K, SO_4, Cl, and HCO_3 molar concentrations were used. For anions, Ca was excluded the discriminant function equations, with the remaining elements Mg, Na, K, SO_4, Cl, HCO_3, and CO_3 being used for the classification. The discriminant function equations like Eq. (4) were constructed from the coefficients of the lower second half of Table S1.

$$DF\alpha m_\tau = (mc_1 \times mC_{elem1}) + (mc_2 \times mC_{elem2}) + (mc_3 \times mC_{elem3}) + (mc_4 \times mC_{elem4}) + (mc_5 \times mC_{elem5}) + (mc_6 \times mC_{elem6}) + (mc_7 \times mC_{elem7}) + mc_0 \quad (4)$$

where α is 1 or 2; τ represents the diagram type; m stands for molar concentrations (mC_{elem1} to mC_{elem7}); and coefficients mc_1 to mc_7 and mc_0 are listed in the second half of Table S1.

The third classification scheme, the *Greater molar conc model*, was the same as that used for assigning the initial classes to the water compositions, i.e., the greatest of the 4 cation and anion molar concentrations for the nomenclature of cations and anions, respectively.

With these 3 proposals, it is necessary to decide on the best option. As experimental data are always characterized by the central tendency

Table 1
Statistical summary of simulated concentration data for multi-normally distributed simulated samples for the LDA and canonical analysis.

No.	Water class	n	Ca		Mg		Na		K		SO_4		Cl		HCO_3		CO_3	
			$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}	$\bar{x}$	U_{99}
Concentrations (mg/L)																		
1	Calcium sulfate	2853	3249.2	30.0	964.3	25.3	933.9	26.1	1562.1	44.1	7835.7	75.3	1406.5	39.5	2442.7	68.3	2376.6	62.1
2	Calcium chloride	2818	3073.8	31.2	908.1	24.6	861.2	25.0	1482.2	42.6	4015.6	107.9	3165.0	38.0	2634.1	75.7	2623.2	68.0
3	Calcium bicarbonate	2909	3088.0	31.5	912.6	23.9	873.3	25.1	1489.2	42.9	4111.2	106.9	1530.0	43.7	5393.6	63.0	2643.8	66.6
4	Calcium carbonate	2788	3278.0	31.1	975.5	25.5	919.9	25.8	1562.9	44.5	3814.8	101.2	1455.0	40.1	2472.4	69.7	4886.0	46.5
5	Magnesium sulfate	2892	1586.4	41.7	1990.5	19.1	917.4	26.0	1563.7	44.7	7807.8	72.2	1409.6	39.7	2477.5	68.8	2398.0	62.1
6	Magnesium chloride	2870	1496.2	40.5	1871.9	18.7	887.0	24.7	1500.1	42.0	4126.5	104.0	3142.1	36.5	2740.3	74.7	2585.9	66.7
7	Magnesium bicarbonate	2908	1462.7	39.5	1839.5	19.0	869.3	24.7	1448.6	41.1	4106.7	104.7	1525.0	41.8	5362.6	63.4	2484.5	65.3
8	Magnesium carbonate	2915	1620.2	41.0	1982.5	19.2	915.7	25.9	1543.2	44.2	3855.1	99.6	1426.6	39.8	2443.3	66.9	4882.6	47.4
9	Sodium sulfate	2869	1693.6	44.3	1046.2	26.6	2041.3	23.2	1697.0	48.1	7349.2	76.3	1359.7	38.2	2333.6	64.9	2195.7	59.4
10	Sodium chloride	2901	1581.9	42.7	957.4	25.7	1907.5	22.7	1574.0	44.8	3832.2	102.2	2912.9	35.4	2467.5	70.8	2357.0	63.5
11	Sodium bicarbonate	2927	1566.3	41.8	988.8	25.2	1896.7	22.5	1570.2	44.8	3809.4	98.7	1430.0	40.4	5061.7	61.0	2387.8	62.7
12	Sodium carbonate	2862	1727.5	44.6	1038.5	27.2	2043.8	23.6	1699.2	47.0	3583.3	95.0	1351.8	38.6	2369.4	64.7	4574.0	46.5
13	Potassium sulfate	2909	1699.8	43.7	1058.0	26.7	991.8	27.4	3461.5	39.0	7345.0	75.0	1350.1	37.4	2316.4	64.7	2237.5	58.8
14	Potassium chloride	2927	1599.7	41.6	976.6	25.1	930.0	26.5	3217.9	39.4	3766.4	102.0	2983.9	37.0	2492.2	71.6	2385.7	63.3
15	Potassium bicarbonate	2923	1593.6	41.9	968.5	25.8	931.6	26.9	3259.6	38.9	3827.4	103.3	1479.6	41.6	5102.9	61.0	2342.0	64.3
16	Potassium carbonate	3021	1700.3	43.7	1043.7	26.5	993.2	27.6	3477.6	40.0	3565.7	93.2	1341.6	37.6	2295.3	64.3	4595.8	44.9
Concentrations (milliEquivalent/L, mEq/L)																		
1	Calcium sulfate	2853	162.15	1.50	79.35	2.08	40.62	1.14	39.95	1.13	163.15	1.57	39.68	1.12	40.03	1.12	79.21	2.07
2	Calcium chloride	2818	153.39	1.56	74.73	2.03	37.46	1.09	37.91	1.09	83.61	2.25	89.28	1.07	43.17	1.24	87.43	2.27
3	Calcium bicarbonate	2909	154.10	1.57	75.10	1.97	37.99	1.09	38.09	1.10	85.60	2.23	43.16	1.23	88.40	1.03	88.11	2.22
4	Calcium carbonate	2788	163.58	1.55	80.27	2.10	40.01	1.12	39.97	1.14	79.43	2.11	41.05	1.13	40.52	1.14	162.84	1.55
5	Magnesium sulfate	2892	79.17	2.08	163.79	1.57	39.91	1.13	39.99	1.14	162.57	1.50	39.76	1.12	40.60	1.13	79.92	2.07
6	Magnesium chloride	2870	74.66	2.02	154.04	1.54	38.58	1.07	38.37	1.07	85.92	2.17	88.63	1.03	44.91	1.22	86.18	2.22
7	Magnesium bicarbonate	2908	72.99	1.97	151.37	1.56	37.81	1.07	37.05	1.05	85.51	2.18	43.02	1.18	87.89	1.04	82.81	2.18
8	Magnesium carbonate	2915	80.85	2.05	163.13	1.58	39.83	1.12	39.47	1.13	80.27	2.07	40.24	1.12	40.04	1.10	162.73	1.58
9	Sodium sulfate	2869	84.52	2.21	86.09	2.19	88.79	1.01	43.40	1.23	153.02	1.59	38.36	1.08	38.25	1.06	73.18	1.98
10	Sodium chloride	2901	78.94	2.13	78.78	2.11	82.97	0.99	40.26	1.14	79.79	2.13	82.17	1.00	40.44	1.16	78.56	2.12
11	Sodium bicarbonate	2927	78.16	2.09	81.37	2.08	82.50	0.98	40.16	1.15	79.32	2.05	40.34	1.14	82.96	1.00	79.58	2.09
12	Sodium carbonate	2862	86.21	2.22	85.45	2.24	88.90	1.03	43.46	1.20	74.61	1.98	38.13	1.09	38.83	1.06	152.45	1.55
13	Potassium sulfate	2909	84.82	2.18	87.06	2.20	43.14	1.19	88.53	1.00	152.93	1.56	38.09	1.06	37.96	1.06	74.57	1.96
14	Potassium chloride	2927	79.83	2.08	80.37	2.06	40.45	1.15	82.30	1.01	78.42	2.12	84.17	1.04	40.85	1.17	79.51	2.11
15	Potassium bicarbonate	2923	79.53	2.09	79.70	2.12	40.52	1.17	83.37	1.00	79.69	2.15	41.74	1.17	83.63	1.00	78.06	2.14
16	Potassium carbonate	3021	84.85	2.18	85.88	2.18	43.20	1.20	88.94	1.02	74.24	1.94	37.85	1.06	37.62	1.05	153.17	1.50
Concentrations (milliMoles/L, mM/L)																		
1	Calcium sulfate	2853	81.07	0.75	39.67	1.04	40.62	1.14	39.95	1.13	81.57	0.78	39.68	1.12	40.03	1.12	39.60	1.03
2	Calcium chloride	2818	76.70	0.78	37.36	1.01	37.46	1.09	37.91	1.09	41.81	1.12	89.28	1.07	43.17	1.24	43.71	1.13
3	Calcium bicarbonate	2909	77.05	0.79	37.55	0.99	37.99	1.09	38.09	1.10	42.80	1.11	43.16	1.23	88.40	1.03	44.06	1.11
4	Calcium carbonate	2788	81.79	0.77	40.14	1.05	40.01	1.12	39.97	1.14	39.71	1.05	41.05	1.13	40.52	1.14	81.42	0.77
5	Magnesium sulfate	2892	39.58	1.04	81.90	0.79	39.91	1.13	39.99	1.14	81.28	0.75	39.76	1.12	40.60	1.13	39.96	1.03
6	Magnesium chloride	2870	37.33	1.01	77.02	0.77	38.58	1.07	38.37	1.07	42.96	1.08	88.63	1.03	44.91	1.22	43.09	1.11
7	Magnesium bicarbonate	2908	36.50	0.98	75.68	0.78	37.81	1.07	37.05	1.05	42.75	1.09	43.02	1.18	87.89	1.04	41.40	1.09
8	Magnesium carbonate	2915	40.43	1.02	81.57	0.79	39.83	1.12	39.47	1.13	40.13	1.04	40.24	1.12	40.04	1.10	81.37	0.79
9	Sodium sulfate	2869	42.26	1.11	43.04	1.09	88.79	1.01	43.40	1.23	76.51	0.79	38.36	1.08	38.25	1.06	36.59	0.99
10	Sodium chloride	2901	39.47	1.07	39.39	1.06	82.97	0.99	40.26	1.14	39.90	1.06	82.17	1.00	40.44	1.16	39.28	1.06
11	Sodium bicarbonate	2927	39.08	1.04	40.68	1.04	82.50	0.98	40.16	1.15	39.66	1.03	40.34	1.14	82.96	1.00	39.79	1.04
12	Sodium carbonate	2862	43.10	1.11	42.73	1.12	88.90	1.03	43.46	1.20	37.30	0.99	38.13	1.09	38.83	1.06	76.22	0.77
13	Potassium sulfate	2909	42.41	1.09	43.53	1.10	43.14	1.19	88.53	1.00	76.47	0.78	38.09	1.06	37.96	1.06	37.29	0.98
14	Potassium chloride	2927	39.91	1.04	40.18	1.03	40.45	1.15	82.30	1.01	39.21	1.06	84.17	1.04	40.85	1.17	39.76	1.05
15	Potassium bicarbonate	2923	39.76	1.05	39.85	1.06	40.52	1.17	83.37	1.00	39.85	1.08	41.74	1.17	83.63	1.00	39.03	1.07
16	Potassium carbonate	3021	42.42	1.09	42.94	1.09	43.20	1.20	88.94	1.02	37.12	0.97	37.85	1.06	37.62	1.05	76.59	0.75
Hybrid log-ratios (hlr) of the molar concentrations (mM/L) of 4 cations and 4 anions																		
			hlr_2		hlr_3		hlr_4		hlr_5		hlr_6		hlr_7		hlr_8			
1	Calcium sulfate	2853	0.216	0.034	0.231	0.037	0.255	0.039	−0.710	0.014	0.260	0.038	0.238	0.037	0.217	0.034		
2	Calcium chloride	2818	0.303	0.036	0.331	0.039	0.308	0.037	0.183	0.036	−0.786	0.016	0.184	0.039	0.117	0.034		
3	Calcium bicarbonate	2909	0.283	0.034	0.343	0.040	0.337	0.040	0.157	0.035	0.191	0.038	−0.775	0.016	0.105	0.033		
4	Calcium carbonate	2788	0.207	0.034	0.207	0.034	0.244	0.037	0.264	0.039	0.225	0.034	0.218	0.037	0.244	0.039		
5	Magnesium sulfate	2892	−0.717	0.014	0.254	0.038	0.264	0.039	−0.711	0.014	0.259	0.038	0.229	0.037	0.197	0.032		
6	Magnesium chloride	2870	−0.622	0.013	0.305	0.036	0.322	0.038	0.143	0.032	−0.758	0.015	0.136	0.036	0.157	0.033		
7	Magnesium bicarbonate	2908	−0.635	0.013	0.312	0.038	0.334	0.038	0.122	0.033	0.157	0.037	−0.780	0.016	0.175	0.034		
8	Magnesium carbonate	2915	−0.711	0.014	0.266	0.038	0.284	0.039	0.201	0.034	0.244	0.037	0.243	0.038	−0.708	0.014		
9	Sodium sulfate	2869	0.131	0.033	−0.778	0.016	0.177	0.038	−0.632	0.014	0.308	0.038	0.306	0.038	0.326	0.035		
10	Sodium chloride	2901	0.242	0.036	−0.722	0.016	0.251	0.039	0.209	0.033	−0.710	0.015	0.250	0.038	0.238	0.035		
11	Sodium bicarbonate	2927	0.188	0.033	−0.706	0.015	0.252	0.037	0.221	0.034	0.260	0.039	−0.710	0.015	0.229	0.035		
12	Sodium carbonate	2862	0.165	0.034	−0.768	0.016	0.158	0.036	0.308	0.034	0.328	0.038	0.282	0.036	−0.619	0.013		
13	Potassium sulfate	2909	0.120	0.032	0.183	0.037	−0.773	0.016	−0.629	0.013	0.318	0.038	0.314	0.037	0.303	0.035		
14	Potassium chloride	2927	0.204	0.033	0.257	0.038	−0.702	0.015	0.257	0.035	−0.723	0.016	0.253	0.039	0.231	0.035		
15	Potassium bicarbonate	2923	0.220	0.034	0.263	0.040	−0.723	0.016	0.237	0.036	0.220	0.039	−0.726	0.016	0.272	0.037		
16	Potassium carbonate	3021	0.135	0.033	0.172	0.037	−0.785	0.015	0.299	0.034	0.334	0.039	0.331	0.038	−0.642	0.014		

Table 2
Synthesis of percent success values obtained from the training of different three-field classification diagrams from the multi-variate discordant outlier – free database (7 hlr and 7 M conc models).

Classification type		number of samples used and the class name (100%)								number and % of correctly classified samples								number and % of incorrectly classified samples	
name or type[a]	model type	class 1		class 2		class 3		total		class 1		class 2		class 3		total		no.	%
		name	no.	name	no.	name	no.	name	no.	no.	%	no.	%	no.	%	no.	%		
1. Subdivision of Ca, Mg, and Na cations (Ca-Mg-Na classes)																			
Ca-Mg-Na	*7 hlr*	Ca	11,368	Mg	11,585	Na	11,559	Ca-Mg-Na	34,512	10,812	95.11	10,957	94.58	10,945	94.69	32,714	**94.79**	1798	5.21
Ca-Mg-Na	*7 M*	Ca	11,368	Mg	11,585	Na	11,559	Ca-Mg-Na	34,512	11,360	99.93	11,538	99.59	10,969	94.90	33,867	**98.13**	645	1.87
2. Subdivision of Ca, Mg, and K cations (Ca-Mg-K classes)																			
Ca-Mg-K	*7 hlr*	Ca	11,368	Mg	11,585	K	11,780	Ca-Mg-K	34,733	10,659	93.76	10,904	94.12	11,108	94.30	32,671	**94.06**	2062	5.94
Ca-Mg-K	*7 M*	Ca	11,368	Mg	11,585	K	11,780	Ca-Mg-K	34,733	11,352	99.86	11,531	99.53	11,220	98.19	34,103	**98.19**	630	1.81
3. Subdivision of Ca, Na, and K cations (Ca-Na-K classes)																			
Ca-Na-K	*7 hlr*	Ca	11,368	Na	11,559	K	11,780	Ca-Na-K	34,707	10,353	91.07	11,091	95.95	11,382	96.62	32,836	**94.58**	1881	5.42
Ca-Na-K	*7 M*	Ca	11,368	Na	11,559	K	11,780	Ca-Na-K	34,707	11,363	99.96	11,238	97.22	11,423	96.97	34,024	**98.03**	683	1.97
4. Subdivision of Mg, Na, and K cations (Mg-Na-K classes)																			
Mg-Na-K	*7 hlr*	Mg	11,585	Na	11,559	K	11,780	Mg-Na-K	34,924	10,587	91.39	11,185	96.76	11,451	97.21	33,223	**95.13**	1701	4.87
Mg-Na-K	*7 M*	Mg	11,585	Na	11,559	K	11,780	Mg-Na-K	34,924	11,564	99.82	11,229	97.15	11,520	97.79	34,313	**98.25**	611	1.75
5. Subdivision of SO_4, Cl, and HCO_3 anions (SO_4, Cl, and HCO_3 classes)																			
SO_4-Cl-HCO_3	*7 hlr*	SO_4	11,523	Cl	11,516	HCO_3	11,667	SO_4-Cl-HCO_3	34,706	10,689	92.76	11,084	96.25	11,310	96.94	33,083	**95.32**	1623	4.68
SO_4-Cl-HCO_3	*7 M*	SO_4	11,523	Cl	11,516	HCO_3	11,667	SO_4-Cl-HCO_3	34,706	11,520	99.97	11,144	96.77	11,273	96.62	33,937	**97.78**	769	2.22
6. Subdivision of SO_4, Cl, and CO_3 anions (SO_4, Cl, and CO_3 classes)																			
SO_4-Cl-CO_3	*7 hlr*	SO_4	11,523	Cl	11,516	CO_3	11,586	SO_4-Cl-CO_3	34,625	11,029	95.71	10,776	93.57	11,113	95.92	32,918	**95.07**	1707	4.93
SO_4-Cl-CO_3	*7 M*	SO_4	11,523	Cl	11,516	CO_3	11,586	SO_4-Cl-CO_3	34,625	11,501	99.81	10,806	93.83	11,586	100.00	33,873	**97.89**	732	2.11
7. Subdivision of SO_4, HCO_3, and CO_3 anions (SO_4, HCO_3, and CO_3 classes)																			
SO_4-HCO_3-CO_3	*7 hlr*	SO_4	11,523	HCO_3	11,667	CO_3	11,586	SO_4-HCO_3-CO_3	34,776	10,985	95.33	11,031	94.55	11,083	95.66	33,099	**95.18**	1677	4.82
SO_4-HCO_3-CO_3	*7 M*	SO_4	11,523	HCO_3	11,667	CO_3	11,586	SO_4-HCO_3-CO_3	34,776	11,492	99.73	10,964	93.97	11,577	99.92	34,033	**97.86**	743	2.14
8. Subdivision of Cl, HCO_3, and CO_3 anions (Cl, HCO_3, and CO_3 classes)																			
Cl-HCO_3-CO_3	*7 hlr*	Cl	11,516	HCO_3	11,667	CO_3	11,586	Cl-HCO_3-CO_3	34,769	11,094	96.34	11,306	96.91	10,761	92.88	33,161	**95.38**	1608	4.62
Cl-HCO_3-CO_3	*7 M*	Cl	11,516	HCO_3	11,667	CO_3	11,586	Cl-HCO_3-CO_3	34,769	11,114	96.51	11,292	96.79	11,582	99.97	33,988	**97.75**	781	2.25

[a] The numbers 1 to 8 in the subheadings refer to the DF1-DF2 diagrams or probability estimates for the decision on correct classification.

and dispersion parameters (e.g., Bevington, 1969; Bevington and Robinson, 2003; Barnett and Lewis, 1994; Castells and Castillo, 2000; Miller and Miller, 2010; Verma, 2020a), it is highly recommended to evaluate their effects on the 3 classification schemes. The analytical data are not free from errors or uncertainties. Even the atomic weights have their experimental uncertainties (see IUPAC database). On the other hand, water may have interacted with some specific minerals, whose effects could be different on these classification schemes. Therefore, it was considered indispensable to evaluate these effects.

At this stage, it should be clear that it is difficult to use and evaluate the classification schemes without a suitable computer program. Therefore, a computer program *WaterMClaSys_LDA* (Figs. S1–S3) was written in Java ZK Framework, which requires an input file in Excel® and provides an output Excel® file. The complex program structure can be visualized in three parts. The program starts with the data validation (Fig. S1). The Excel® datafile, containing concentration data (in mg/L) for 8 elements (4 cations and 4 anions), is first evaluated for any typographical errors and if present, the user is asked to correct them in Excel® before reentering *WaterMClaSys_LDA*, which provides the application of all 3 classification schemes (Fig. S2). The water nomenclature is achieved from probability calculations for the competing fields in all "three at a time" diagrams and their comparison (Fig. S2).

Besides the application of all three classification models (*7 hlr*, *7 M conc*, and *Greater molar conc*; Fig. S2), the program facilitates the evaluation of robustness of both analytical errors or uncertainties and compositional changes in the field (Fig. S3). Therefore, we can simulate such errors to evaluate the robustness of the multidimensional proposal for water classification. The uncertainty propagation module simulates 2200 new compositions from Monte Carlo simulations (Verma and Quiroz-Ruiz, 2006; Verma, 2020a) in terms of the mean composition and related analytical uncertainties (Fig. S3). All 2200 analyses are processed in *WaterMClaSys_LDA* for the 16 corresponding water class(es). The program then counts the samples identified as the original water class and calculates the percent success or correct classification as the percent ratio of correctly classified samples (as the original composition) and total number of samples being 2200 (Fig. S3). For robustness against compositional changes in the field, the program evaluates the effects of small steps of gain or loss of a few or all elements on a given original water composition (provided by the user in an Excel® file) and evaluates the total number of steps in which the original water class is maintained. The simulation stops when the water class is changed or any of the compositional conditions, such as any chemical concentration becomes <0.002 mg/L, is indicated. The total number of steps is reported for each of the three models.

The multidimensional *7 hlr model* was trained with 16 initial water classes and the application is strictly based on the probability concept. It is possible to achieve a final combined main and hybrid classification of many more, theoretically up to a total of 256 water classes. Let us suppose that p_i is the probability for a cation or anion, where the subscript i varies from 1 to 4 for 4 cations or anions. Of these 4 p_i's, let p_m be the highest probability and p_n be the next to the highest probability. The conditions that should be fulfilled for the basic nomenclature were as follows: $p_m \geq 0.50$ and $(p_m - p_n) \geq 0.25$ and $p_n \leq 0.25$; otherwise, a hybrid (two water types, the highest probability class name followed by the next highest name) nomenclature is assigned. Thus, for the cations and anions separately, 4 basic and 12 hybrid classes can be achieved. Their combination would provide 16×16, i.e., 256 classes in total.

WaterMClaSys_LDA provides a detailed final report as well as a synthesis report in an Excel® file, The detailed report contains the following parameters for each sample: (**1**) Original input concentrations (mg/L) of the 4 cations and 4 anions; (**2**) their conversion to milliequivalent (mEq/L) concentrations; (3) the conversion to milliMoles (mM/L) concentrations; (**4**) the original primary classification based on the Greater molar concentration criterion (*Greater molar conc model*); (**5**) the hlr transformed values for the *7 hlr model*; (**6**) the *DF*1-*DF*2 functions for the hlr and the respective probabilities for 4 "three at a time" diagrams for cations; (**7**) the same information for anions as in **6**; (**8**) the sum of the probabilities for each of the 4 cations in the 4 diagrams and the probability-based decisions of single cation nomenclature, including the hybrid types of two cation names instead of one; and (**11**) ionic charge imbalance (ici %) and acceptability (arbitrary criteria used: acceptable when ici ≤ 10%; otherwise, imbalance too high). The report of the *7 M conc* model was eliminated from the Recommended procedure of the final version of *WaterMClaSys_LDA* because of its significantly lesser robustness as compared to the *7 hlr model* (see Section 3.2).

3. Results and discussion

3.1. Training of the multidimensional schemes

We present the multidimensional solution for water classification as 16 classes (Table 2) from a total of 8 sets of probability calculations. Table 2 summarizes the training of two schemes: *7 hlr* and *7 M conc*. Although the plotting of samples in a diagram is not required, we will present them for those (almost everybody) who wish to see the plots. The 7 hlr based DF1-DF2 type classification was achieved in 8 diagrams (Fig. 1a-h; Table 2). Similarly, the 7 M concentrations based DF1-DF2 type classification was also achieved in 8 diagrams (Fig. 2a-h; Table 2).

For the first set of hlr based classification (*7 hlr model*; subdivision of Ca, Mg, and Na cations; Fig. 1a; Table 2), the 7 hlr transformed variables provided success values of 95.11% for Ca (10,812 samples correctly classified out of 11,368 samples), 94.58% for Mg (10,957 correctly classified out of 11,585 samples), and 94.69% for Na (10,945 correctly classified out of 11,559 samples). The overall percent success was 94.79% (32,714 correctly classified out of 34,512 samples), with the consequence that the incorrect classification was only about 5.21% (*7 hlr model*; Table 2). All data corresponding to the Ca, Mg, and Na classes were plotted, along with the respective centroids (Fig. 1a).

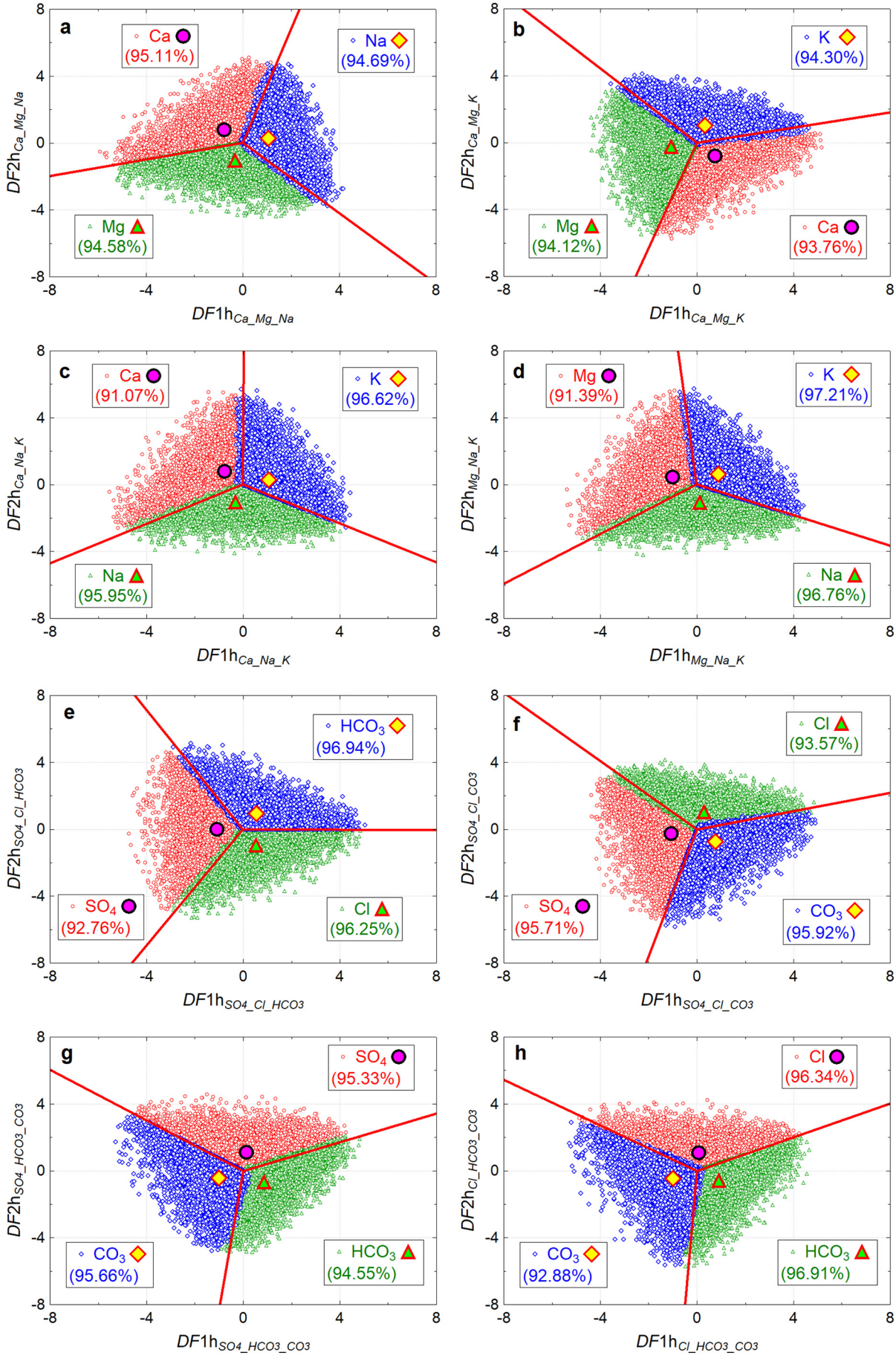
For the cations Ca-Mg-K, the percent success values were between 93.76% and 94.30% for the *7 hlr model* (Fig. 1b; Table 2). The respective DF functions were constructed from equations constructed from the general Eq. (1).

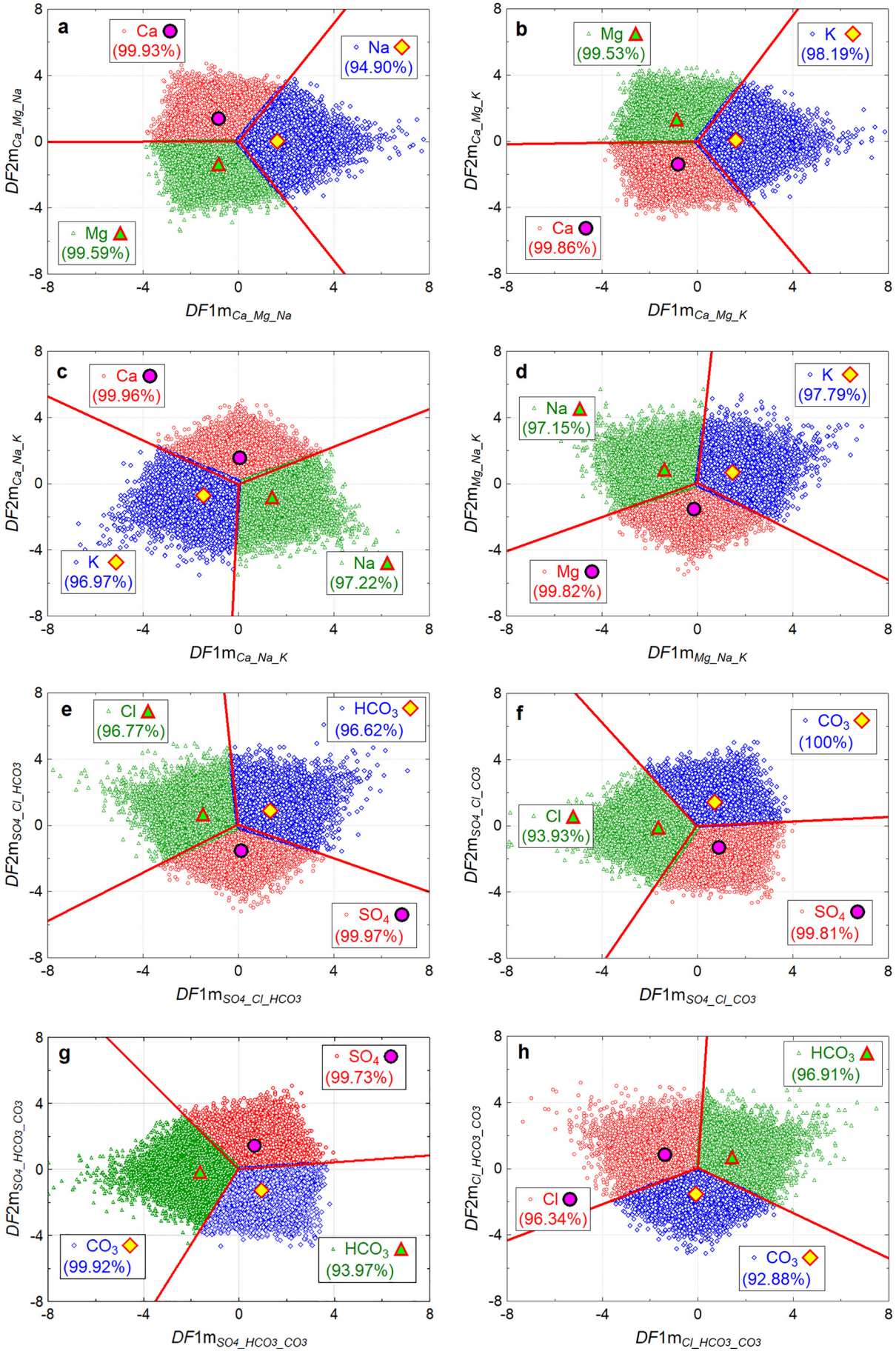
Similarly, the remaining two combinations of cations of three at a time, Ca-Na-K and Mg-Na-K, high success values of 91.07%–95.95%–96.62% (Fig. 1c) and 91.39%–96.76%–97.21% (Fig. 1d), respectively, were obtained (*7 hlr model*; Table 2). For the 4 sets of 3 cations at a time, the overall high percent success values for the *7 hlr model* varied from 94.06% to 95.13% (Table 2).

The classification of anions (SO_4, Cl, HCO_3, and CO_3) was similarly achieved from 4 diagrams of "three at a time" type (Fig. 1e–h; *7 hlr model*; Table 2). The success values were like those for cations. In summary, for the 4 sets of 3 anions at a time, the overall high percent success values for the *7 hlr model* varied from 95.07% to 95.38% (Table 2).

The cation and anion classification from the 7 original molar concentrations without log-ratio transformations (Figs. 2a-h; *7 M conc model*; Table 2) provided even higher success values than the *7 hlr model*. For example, for the subdivision of Ca-Mg-Na (Fig. 2a; Table 2), the *7 M conc model* gave overall percent success values of 98.13% (33,867 correctly classified out of 34,512 samples). Fig. 2a shows these samples plotted in a DF1-DF2 type diagram, specifically $DF1m_{Ca\_Mg\_Na} - DF2m_{Ca\_Mg\_Na}$ type, for which the functions were calculated, respectively, from Eq. (4).

Fig. 1. Discriminant function (*DF*1h-*DF*2h) diagrams for the hybrid log-ratios (hlr) of the training database corresponding to the cation (**a** to **d**) and anion (**e** to **h**) classes; the symbols and percent success values are explained in insets (the centroids are shown with larger filled symbols). (**a**) classification of Ca, Mg, and Na; (**b**) classification of Ca, Mg, and K; (**c**) classification of Ca, Na, and K; (**d**) classification of Mg, Na, and K; (**e**) classification of SO_4, Cl, and HCO_3; (**f**) classification of SO_4, Cl, and CO_3; (**g**) classification of SO_4, HCO_3, and CO_3; and (**h**) classification of Cl, HCO_3, and CO_3.


a
Ca
(95.11%)
Na
(94.69%)
Mg
(94.58%)
DF2hCa_Mg_Na
DF1hCa_Mg_Na
b
K
(94.30%)
Mg
(94.12%)
Ca
(93.76%)
DF2hCa_Mg_K
DF1hCa_Mg_K
c
Ca
(91.07%)
K
(96.62%)
Na
(95.95%)
DF2hCa_Na_K
DF1hCa_Na_K
d
Mg
(91.39%)
K
(97.21%)
Na
(96.76%)
DF2hMg_Na_K
DF1hMg_Na_K
e
HCO3
(96.94%)
SO4
(92.76%)
Cl
(96.25%)
DF2hSO4_Cl_HCO3
DF1hSO4_Cl_HCO3
f
Cl
(93.57%)
SO4
(95.71%)
CO3
(95.92%)
DF2hSO4_Cl_CO3
DF1hSO4_Cl_CO3
g
SO4
(95.33%)
CO3
(95.66%)
HCO3
(94.55%)
DF2hSO4_HCO3_CO3
DF1hSO4_HCO3_CO3
h
Cl
(96.34%)
CO3
(92.88%)
HCO3
(96.91%)
DF2hCl_HCO3_CO3
DF1hCl_HCO3_CO3


a
Ca (99.93%)
Na (94.90%)
Mg (99.59%)
DF2m Ca_Mg_Na
DF1m Ca_Mg_Na
b
Mg (99.53%)
K (98.19%)
Ca (99.86%)
DF2m Ca_Mg_K
DF1m Ca_Mg_K
c
Ca (99.96%)
K (96.97%)
Na (97.22%)
DF2m Ca_Na_K
DF1m Ca_Na_K
d
K (97.79%)
Na (97.15%)
Mg (99.82%)
DF2m Mg_Na_K
DF1m Mg_Na_K
e
Cl (96.77%)
HCO3 (96.62%)
SO4 (99.97%)
DF2m SO4_Cl_HCO3
DF1m SO4_Cl_HCO3
f
CO3 (100%)
Cl (93.93%)
SO4 (99.81%)
DF2m SO4_Cl_CO3
DF1m SO4_Cl_CO3
g
SO4 (99.73%)
CO3 (99.92%)
HCO3 (93.97%)
DF2m SO4_HCO3_CO3
DF1m SO4_HCO3_CO3
h
HCO3 (96.91%)
Cl (96.34%)
CO3 (92.88%)
DF2m Cl_HCO3_CO3
DF1m Cl_HCO3_CO3

The individual success values were as follows (Fig. 2a; *7 M conc model*; Table 2): 99.93% (11,360 correctly classified out of 11,368 samples) for Ca; 99.59% (11,538 out of 11,585) for Mg; and 94.90% (10,969 out of 11,559) for Na. Similarly, very high success values were obtained for the other, three at a time, classifications (Fig. 2b-h; *7 M conc model*; Table 2). In summary, the cations and anions were classified with overall success values of 98.03%–98.25% and 97.75%–97.89%, respectively (*7 M conc model*; Table 2).

Thus, the multidimensional classification of 16 classes of water (Table 2) was achieved with percent success values ranging from 94.06% to 95.38% from the hlr (*7 hlr model*) and from 97.75% to 98.25% for the molar concentrations (*7 M conc model*). The classification from the third scheme (*Greater molar conc model*), by definition, will give success values of 100%. Nevertheless, it is extremely important to evaluate the robustness of all three schemes to decide which one should be preferred for the water nomenclature of field samples. This should be done especially for the "Analytical errors or uncertainties" concept, irrespective of the basic or combined basic and hybrid water nomenclature.

3.2. Robustness tests

3.2.1. Analytical errors or uncertainties

Analytical errors or uncertainties on individual geochemical data are seldom if ever reported although, as recently documented by Verma et al. (2019), it is entirely feasible to do so. Thus, the reports of water analysis seldom, if ever, include analytical errors on individual measurements. For establishing probable uncertainties on individual elements, we noted that the mass imbalance errors of about 9% reported by Xue et al. (2019) may be indicative of individual partial errors of a few % in each component, whereas total uncertainties may be even higher around several %. Busico et al. (2020) estimated about 7% (partial) errors for their water chemistry data. Verma (2013) evaluated the analytical data from the International Atomic Energy Agency inter-laboratory comparison program of synthetic geothermal-like water samples and documented very high systematic errors (underestimations) of −5% for Na, −1% for K, −13% for Ca, −54% for Mg, −13% for Cl, and − 91% for SO_4. Again, these errors are only systematic errors, which should be avoided, although unfortunately no workers are really evaluating and eliminating them in water analysis. Based on this discussion and for illustration purposes, we assumed ±40% RSD (relative standard deviation) values for all elements.

In the absence of total errors or uncertainties in water analysis, we used our training database to estimate 16 compositions of the highest success probability (generally close to about the highest theoretical probability of 0.75 (equivalent to 75%) for a given field for 4 sets of "three at a time" subdivision) for all 16 water types (Table S3). The input concentrations were arbitrarily reported to two decimal places in Table S3. The Robustness module for uncertainty propagation was used to simulate the effects of analytical uncertainties on the compositions of Table S3. The results of all three models for the *7 hlr* and *7 M conc* initial compositions are summarized in Fig. 3a and b, respectively.

For the initial compositions corresponding to the highest 0.75 probability in the *7 hlr model* (Fig. 3a), the multidimensional *7 hlr model*, as compared to the other two models (*7 M conc* and *Greater molar conc*), showed higher percent success values for 9 (class numbered 1, 2, 4, 7, 10, 11, 12, 13, and 14) out of 16 classes, lower for only 1 class (numbered 6), and similar for the remaining 6 classes. Similarly, the comparison of the *7 M conc* and *Greater molar conc* models showed that the multidimensional 7 M conc model showed higher success for only 2 classes (numbered 1 and 12), lower for 4 classes (2, 10, 11, and 14), and similar for the remaining 10 classes (Fig. 3a). Note that all success values in these robustness tests were high from 91.4% to 97.9% (Fig. 3a).

For the initial compositions corresponding to the highest probability in the *7 M conc* model (Fig. 3b), the *7 hlr model* depicted higher success for 11 classes (numbered 1, 2, 3, 4, 6, 7, 11, 12, 13, 14, and 15) than the *7 M conc* model, lower for only 1 class (numbered 10), and similar for the remaining 4 classes. The multidimensional *7 M conc* model showed higher success for 6 classes (1, 3, 8, 10, 12, and 16), lower for 4 classes (2, 11, 14, and 15), and similar for the remaining 6 classes, as compared to the *Greater molar conc* model. For these cases also (Fig. 3b), all success values were high from 85.9% to 96.7%.

We repeated the simulations for the initial compositions (Table S4) corresponding to the 0.50 probability (lesser than 0.75 in Fig. 3c and d) for a given water class. For compositions corresponding to the *7 hlr model*, the results (Fig. 3c) for the *7 hlr model* showed clearly higher robustness (percent success) for 8 classes (numbered 2, 5–7, 9, and 13–15), lower for 2 classes (numbered 3 and 16), and similar for the remaining 6 cases, than the other two models (*7 M conc* and *Greater molar conc*). For the other set of initial compositions (*7 M conc*; Table S4), the robustness (Fig. 3d) was higher for the *7 hlr model* corresponding to 4 classes, lower for 5 and similar for 7 classes than the other two models.

The success values for the cases of Fig. 3c varied between 35.1% and 72.5%), whereas for Fig. 3d, they were between 14.0% and 44.8%. These values (Fig. 3c and d) are lower than those obtained in Fig. 3a and b, because the initial compositions correspond to a lesser probability of 0.50 (Table S4) for the respective classified field, instead of 0.75 (Table S3).

The three models can be similarly evaluated from any other combination of composition and uncertainty values. Thus, for the nomenclature of water, the significantly greater stability or robustness of the multidimensional *7 hlr model* against analytical uncertainty propagation is clear from this robustness analysis. Therefore, we can propose that the *7 hlr model* be used for the routine nomenclature of water samples.

3.2.2. Compositional changes in the field

The dissolution of minerals will contribute to the input of different elements. For example, plagioclase feldspar will contribute to Ca and Na, whereas potassium feldspar will provide mainly K. Gypsum is likely to add Ca and SO_4. Wide varieties of mica can contribute all four cations Ca, Mg, Na, and K, in addition to several less abundant cations, to the water. If interacted with halite, water will be enriched in Na and Cl. Such waters can be easily classified from field sampling and analyzing their chemistry. *WaterMClaSys_LDA* can be used to evaluate the robustness of the three classification schemes. We raise the following question: Is it possible to recognize the type of water that existed before the mixing process? It is more likely that the multidimensional *7 hlr model* could extract this information because of its generally higher robustness.

As an example, we present the effect of dissolution of plagioclase feldspar in small steps of +1% (gain) of both Ca and Na in the composition of all 16 types of water (Table S3). The results for the nomenclature of the *7 hlr* and *7 M conc* compositions (Table S3) are presented in Fig. 4a and b, respectively. The y-scale gives the maximum percentage of mass gain before the water sample changed its name in each scheme although for the stability of *7 hlr* and *Greater molar conc* models for two water classes (10 and 11) were higher than the y-scale of Fig. 4a. High stability of the nomenclature in all three schemes is clear (Fig. 4a, b). For the compositions of 75% probability for the *7 hlr model* (Table S3), the robustness of the nomenclature of *7 hlr model* was mostly higher than that of the *7 M conc model* (exception of only 1 case with the *7 hlr model* showing slightly lower robustness), whereas the *7 hlr* and *Greater molar conc* models

Fig. 2. Discriminant function (*DF*1m-*DF*2m) diagrams for the molar concentrations of the training database corresponding to the cation (**a** to **d**) and anion (**e** to **h**) classes; the symbols and percent success values are explained in insets (the centroids are shown with larger filled symbols). (**a**) classification of Ca, Mg, and Na; (**b**) classification of Ca, Mg, and K; (**c**) classification of Ca, Na, and K; (**d**) classification of Mg, Na, and K; (**e**) classification of SO_4, Cl, and HCO_3; (**f**) classification of SO_4, Cl, and CO_3; (**g**) classification of SO_4, HCO_3, and CO_3; and (**h**) classification of Cl, HCO_3, and CO_3.

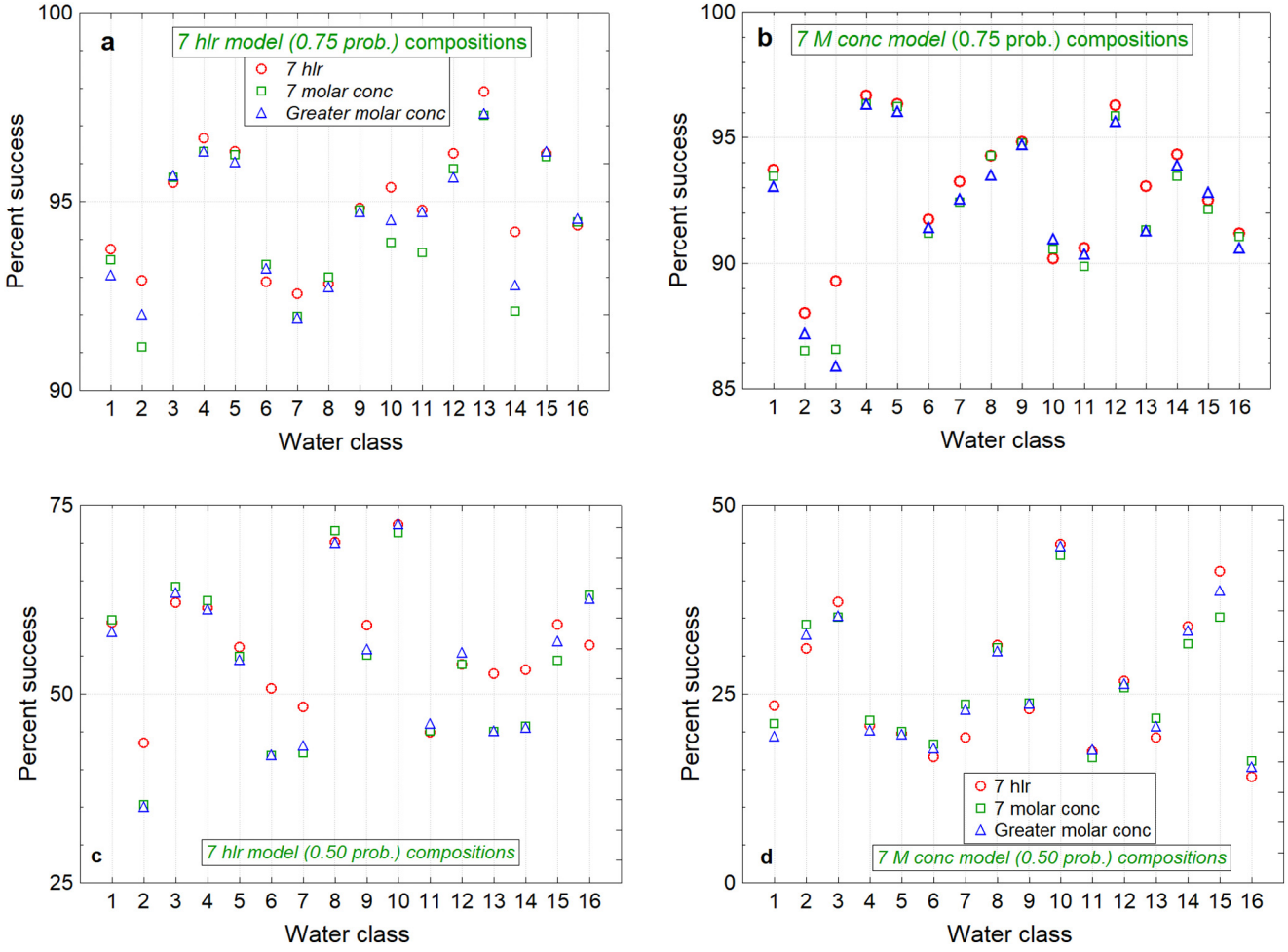


Fig. 3. Robustness tests for the selected compositions for any given cation or anion in terms of ±40% RSD (relative standard deviation); the symbols are explained as an inset in **a**. For water classes 1 to 16, see Table 1; note different y-scales required for different diagrams. (**a**) compositions of the *7 hlr* for compositions corresponding to the highest 0.75 probability for each water class; and (**b**) compositions of the *7 M conc* for compositions corresponding to the highest 0.75 probability for each water class; (**c**) compositions of the *7 hlr* for compositions corresponding to the 0.50 probability for each water class; and (**d**) compositions of the *7 M conc* for compositions corresponding to the 0.50 probability for each water class.

showed similar robustness (Fig. 4a). For the compositions of 75% probability for the *7 M conc model* (Table S3) also, the robustness of the *7 hlr model* as compared to the and *7 M conc model* were as follows: higher for 10 out of 16 cases, lower for 3 cases, and similar for the remaining 3 cases (Fig. 4b). The robustness of the *7 hlr model* was generally like the *Greater molar conc model* (Fig. 4a and b).

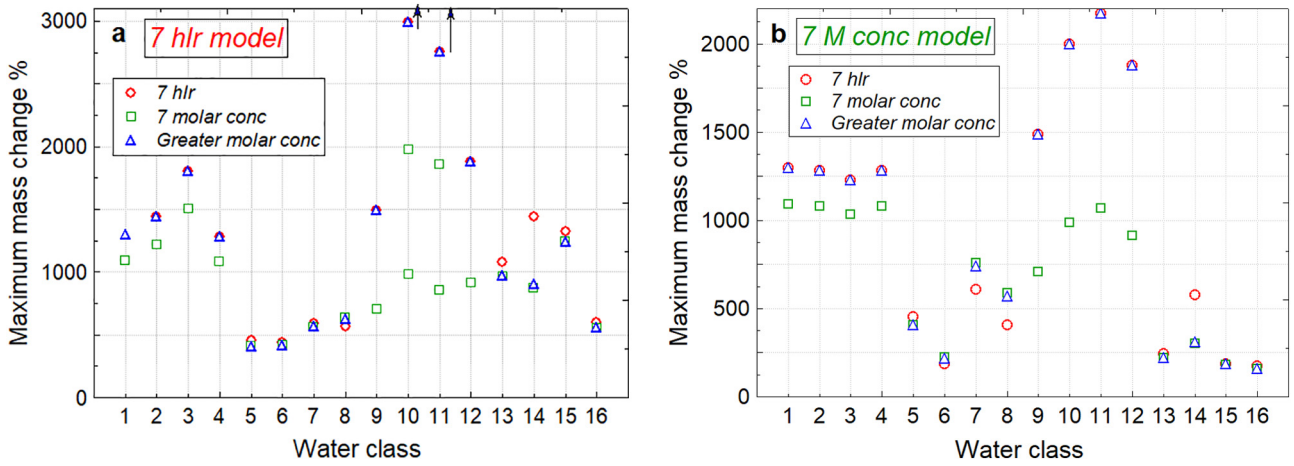


Fig. 4. Robustness tests for the selected compositions as a result of the dissolution of plagioclase feldspar in small steps of +1% (gain) of both Ca and Na; the symbols are explained as an inset in **a**. For water classes 1 to 16, see Table S1; note different y-scales required for the diagrams. (**a**) Initial compositions of the 75% probability for the *7 hlr model* for each water class; and (**b**) compositions of the 75% probability for the *7 M conc* model for each water class.

Table 3
Synthesis of percent success values obtained from the evaluation of different three-field classification diagrams from the testing database (7 hlr and 7 M conc models).

Classification type		number of samples used and the class name (100%)								number and % of correctly classified samples								number and % of incorrectly classified samples	
name or type[a]	model type	class 1		class 2		class 3		total		class 1		class 2		class 3		total		no.	%
		name	no.	name	no.	name	no.	name	no.	no.	%	no.	%	no.	%	no.	%		
1. Subdivision of Ca, Mg, and Na cations (Ca-Mg-Na classes)																			
Ca-Mg-Na	*7 hlr*	Ca	1873	Mg	1818	Na	1834	Ca-Mg-Na	5525	1760	94.0	1726	94.9	1728	94.2	5214	**94.4**	311	5.6
Ca-Mg-Na	*7 M*	Ca	1873	Mg	1818	Na	1834	Ca-Mg-Na	5525	1868	99.7	1810	99.6	1726	94.1	5404	**97.8**	121	2.2
2. Subdivision of Ca, Mg, and K cations (Ca-Mg-K classes)																			
Ca-Mg-K	*7 hlr*	Ca	1873	Mg	1818	K	1878	Ca-Mg-K	5569	1744	93.1	1714	94.3	1752	93.3	5210	**93.6**	359	6.4
Ca-Mg-K	*7 M*	Ca	1873	Mg	1818	K	1878	Ca-Mg-K	5569	1870	99.8	1809	99.5	1762	93.8	5441	**97.7**	128	2.3
3. Subdivision of Ca, Na, and K cations (Ca-Na-K classes)																			
Ca-Na-K	*7 hlr*	Ca	1873	Na	1834	K	1878	Ca-Na-K	5585	1684	89.9	1751	95.5	1813	96.5	5248	**94.0**	337	0.6
Ca-Na-K	*7 M*	Ca	1873	Na	1834	K	1878	Ca-Na-K	5585	1873	100	1778	97.0	1816	96.7	5467	**97.9**	118	2.1
4. Subdivision of Mg, Na, and K cations (Mg-Na-K classes)																			
Mg-Na-K	*7 hlr*	Mg	1818	Na	1834	K	1878	Mg-Na-K	5530	1661	91.4	1766	96.3	1825	97.2	5252	**95.0**	278	0.5
Mg-Na-K	*7 M*	Mg	1818	Na	1834	K	1878	Mg-Na-K	5530	1813	99.7	1785	97.3	1829	97.4	5427	**98.1**	103	1.9
5. Subdivision of SO_4, Cl, and HCO_3 anions (SO_4, Cl, and HCO_3 classes)																			
SO_4-Cl-HCO_3	*7 hlr*	SO_4	1825	Cl	1867	HCO_3	1877	SO_4-Cl-HCO_3	5569	1728	94.7	1800	96.4	1818	96.9	5346	**96.0**	223	0.4
SO_4-Cl-HCO_3	*7 M*	SO_4	1825	Cl	1867	HCO_3	1877	SO_4-Cl-HCO_3	5569	1825	100	1801	96.5	1809	96.4	5435	**97.6**	134	2.4
6. Subdivision of SO_4, Cl, and CO_3 anions (SO_4, Cl, and CO_3 classes)																			
SO_4-Cl-CO_3	*7 hlr*	SO_4	1825	Cl	1867	CO_3	1834	SO_4-Cl-CO_3	5526	1759	96.4	1731	92.7	1741	94.9	5231	**94.7**	295	5.3
SO_4-Cl-CO_3	*7 M*	SO_4	1825	Cl	1867	CO_3	1834	SO_4-Cl-CO_3	5526	1818	99.6	1757	94.1	1834	100	5409	**97.9**	117	2.1
7. Subdivision of SO_4, HCO_3, and CO_3 anions (SO_4, HCO_3, and CO_3 classes)																			
SO_4-HCO_3-CO_3	*7 hlr*	SO_4	1825	HCO_3	1877	CO_3	1834	SO_4-HCO_3-CO_3	5536	1765	96.7	1779	94.8	1738	94.8	5282	**95.4**	254	4.6
SO_4-HCO_3-CO_3	*7 M*	SO_4	1825	HCO_3	1877	CO_3	1834	SO_4-HCO_3-CO_3	5536	1814	99.4	1747	93.1	1834	100	5395	**97.5**	141	2.5
8. Subdivision of Cl, HCO_3, and CO_3 anions (Cl, HCO_3, and CO_3 classes)																			
Cl-HCO_3-CO_3	*7 hlr*	Cl	1867	HCO_3	1877	CO_3	1834	Cl-HCO_3-CO_3	5578	1796	96.2	1813	96.6	1691	92.2	5300	**95.0**	278	5.0
Cl-HCO_3-CO_3	*7 M*	Cl	1867	HCO_3	1877	CO_3	1834	Cl-HCO_3-CO_3	5578	1815	97.2	1803	96.1	1834	100	5452	**97.7**	126	2.3

[a] The numbers 1 to 8 in the subheadings refer to the DF1-DF2 diagrams or probability estimates for the decision on correct classification.

3.3. Testing of the new classification scheme

Although it is a common practice to subdivide randomly the complete database into two parts, one each for training and testing, we preferred to use the complete database for the training and to randomly generate afresh an independent database for testing. Our procedure assures that the training set samples comply with the basic assumption of multi-normality required for the LDA and canonical analysis (e.g., Morrison, 1990). The generation of the testing database was done following the same Monte Carlo simulation procedure as for the training set but from different seeds so that the testing data were also random and independent of the training set. This would provide an unbiased testing of the proposed classification scheme. These data were processed in DOMuDaF (Verma et al., 2016) for multivariate discordancy and then in *WaterMClaSys_LDA*. The classification results for 7403 samples are summarized in Table 3 as well as in Figs. 5a-h and 6a-h for the *7 hlr* and *7 M conc* models, respectively. The percent success values are reported in the insets of each diagram.

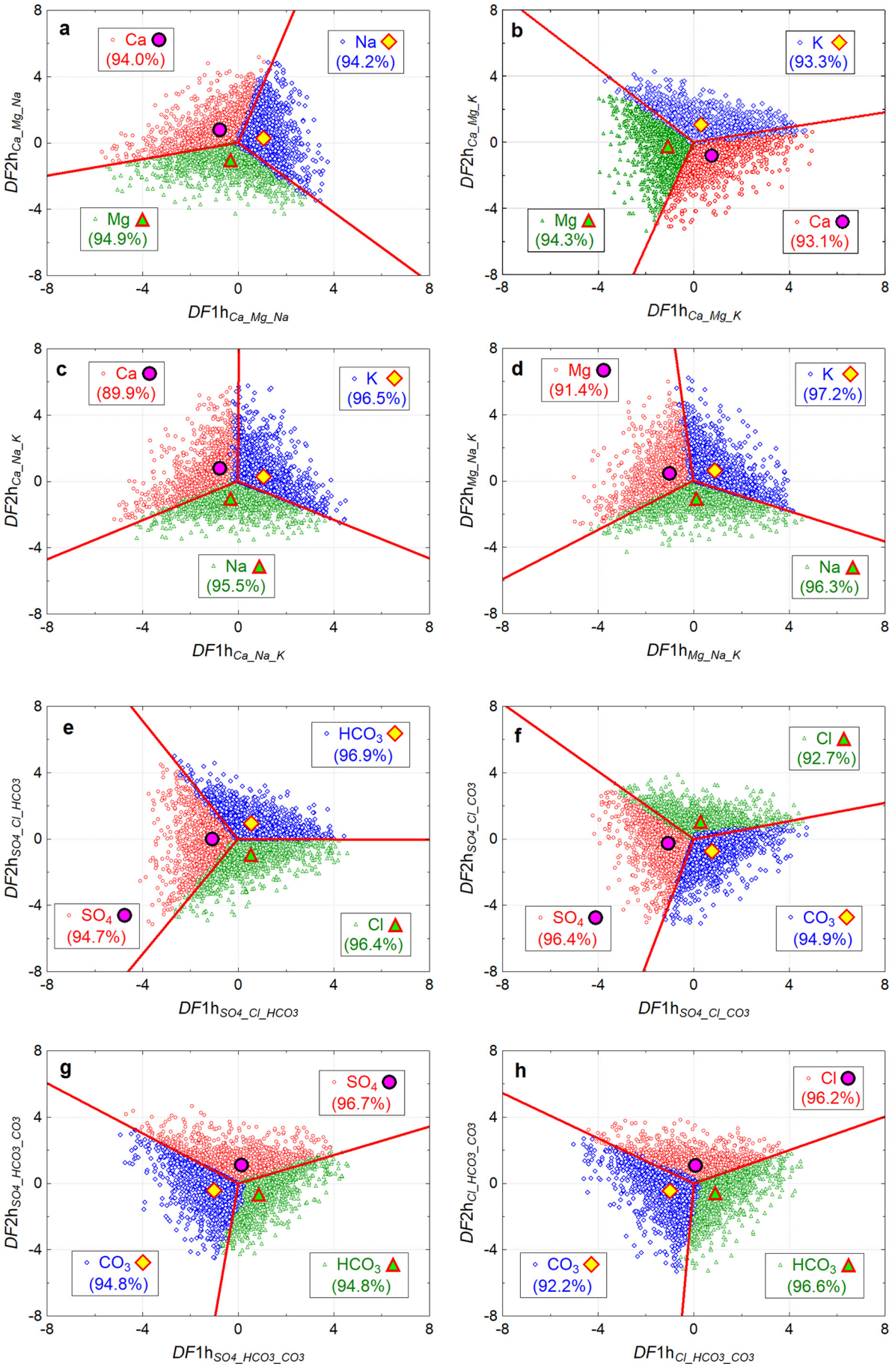
The overall percent success values for the 8 diagrams (Fig. 5a-h; Table 3; *7 hlr* model) were as follows: 94.4%, 93.6%, 94.0%, 95.0%, 96.0%, 94.7%, 95.4%, and 95.0%. These percent success values are comparable to those obtained for the training set samples (Fig. 1a-h; compare Tables 2 and 3). The overall percent success values for the *7 M conc* model (97.5% to 98.1%; weighted means of Fig. 6a-h; Table 3) were also high as for the training set samples (Fig. 2a-h; Table 2).

Therefore, from the high percent success values obtained in this work for both training and testing sets, we conclude that the classification scheme is free from bias. Our proposal was based on similar sample sizes of the training set samples for all classes. The canonical decision has used the concept of the same probability for all classes, irrespective of the sample sizes. Hence, we can infer that there was no overfitting in our classification scheme. The very high success values of over 90% are genuine outcome of our robust proposal.

3.4. Other multivariate techniques in addition to the LDA

In this work, we opted for the use of LDA and canonical analysis as the supervised multivariate technique for the water nomenclature. Several reviews are available on the multivariate techniques (e.g., Mardia et al., 1979; Hand, 1981, 1997; Breiman et al., 1983; Marini, 2010; Miller and Miller, 2010; Lavine and Mirjankar, 2012; Loh, 2014; Verma, 2020a). Classification and regression trees (Breiman et al., 1983; Timofeev, 2004; Loh, 2014) and random forests (Breiman, 2001) techniques have also been used for multivariate data. However, the classification trees do not handle the multivariate data in a statistically coherent way as documented by Agrawal and Verma (2007). The basic problem with the trees resides in the handling of compositional data as univariate, which are otherwise multivariate and could be best handled by the LDA and canonical analysis (e.g., Aitchison, 1986; Agrawal and Verma, 2007; Verma, 2020a).

McNeil et al. (2005) applied the cluster analysis (CA) for the classification of a large, irregular dataset of approximately 34,000 surface water samples from Australia to identify 9 water types. Azhar et al. (2015) used the CA, principal component analysis (PCA), and LDA for the classification of river water quality from 9 stations in the Muda River basin (Malaysia). From the CA and PCA, these authors were able to identify two clusters in their study area. The LDA allowed the authors


a
Ca (94.0%)
Na (94.2%)
Mg (94.9%)
DF2h Ca_Mg_Na
DF1h Ca_Mg_Na
b
K (93.3%)
Mg (94.3%)
Ca (93.1%)
DF2h Ca_Mg_K
DF1h Ca_Mg_K
c
Ca (89.9%)
K (96.5%)
Na (95.5%)
DF2h Ca_Na_K
DF1h Ca_Na_K
d
Mg (91.4%)
K (97.2%)
Na (96.3%)
DF2h Mg_Na_K
DF1h Mg_Na_K
e
HCO3 (96.9%)
SO4 (94.7%)
Cl (96.4%)
DF2h SO4_Cl_HCO3
DF1h SO4_Cl_HCO3
f
Cl (92.7%)
SO4 (96.4%)
CO3 (94.9%)
DF2h SO4_Cl_CO3
DF1h SO4_Cl_CO3
g
SO4 (96.7%)
CO3 (94.8%)
HCO3 (94.8%)
DF2h SO4_HCO3_CO3
DF1h SO4_HCO3_CO3
h
Cl (96.2%)
CO3 (92.2%)
HCO3 (96.6%)
DF2h Cl_HCO3_CO3
DF1h Cl_HCO3_CO3

to propose one discriminant function for discriminating these two classes from only one water quality variable (NH_3—N).

The main distinction between the CA and PCA, on one hand, and the LDA, on the other is that the former methods belong to unsupervised or exploratory pattern recognition techniques, whereas the latter is a supervised and, consequently, a more powerful method (e.g., Reyment and Savazzi, 1999; Miller and Miller, 2010; Verma, 2020a). Thus, the LDA is a powerful multivariate technique, provided the basic assumption of multivariate normality is fulfilled (e.g., Morrison, 1990; Verma, 2020a). This has been precisely the case in the present work. The LDA has been extensively used for the multidimensional discrimination of tectonic settings from igneous and sedimentary rocks as well as for the classification of altered igneous rocks (see Verma, 2020a for dozens of references).

Other newer computational techniques, such as artificial neural network (ANN), regression-based decision tree classifiers, advanced supervised machine learning models, support vector machines (SVM) and multilevel boosting machine learning algorithms (categorical boosting CATBoost, extreme gradient boosting XGBoost, etc.) should be applied in future and compared with the traditional LDA and canonical analysis. For now, we have presented ample arguments (balanced or equal sized classes, all classes free from discordant multivariate outliers, i.e., the fulfilment of the multi-normality assumption for the LDA, option followed to assign the same probability for all classes independent of their sizes, similarly high percent success values from both training and independently generated testing sets, the relevant mathematics known for around 100 years, see, e.g., Morrison, 1990) to show that the LDA is free from bias and overfitting.

3.5. Advantages and shortcomings of the present multidimensional scheme

The usefulness of the classification scheme *WaterMClaSys_LDA* proposed from 16 water types is enhanced by the possibility of inferring both basic and hybrid water types. Although in theory, we may have up to 256 (16 basic and 240 hybrid) water types, in practice the water types will be limited in any given application. Nevertheless, the hybrid water types will be useful for deciphering important geological and environmental processes, such as the mixing of two types of waters and the effects of water-rock interaction. We provide, in the following, some examples where our classification scheme will be helpful.

Giggenbach and Glover (1992) documented the contribution of different plumes of high temperature Cl-rich and lower temperature, lower Cl, but high HCO_3 waters in the Rotorua geothermal field (New Zealand), associated to an extensional tectonic regime. Brombacch et al. (2000) studied the volcanic-hydrothermal system of the high-risk volcano La Soufrière and the geothermal area of Bouillante, Basse-Terre Island, Guadeloupe, Lesser Antilles. These authors identified thermal springs of the Na-Cl, Na-HCO_3, Ca-SO_4, Ca-Na-HCO_3, Ca-Na-Cl, and Ca-SO_4 types. Sadashivaiah et al. (2008) evaluated the groundwater quality from the chemistry data of water samples collected from a large area of over 1000 km^2 in the Karnataka State of South India. Besides the conventional Hill-Piper ternary diagrams, these authors used several chemical criteria for evaluating the water quality for irrigation purposes. Our new classification scheme of a total of 256 water types could help in deciphering geological processes in the volcanic hydrothermal system, such as mixing, as well as the identification of the endmembers in all kinds of waters. The hybrid-types of water classification will be able to handle and assist in the study of such a complex geological situation in a geothermal field or elsewhere for groundwaters.

Golekar et al. (2017) documented the geochemical characteristics of water for drinking and irrigation use in and around Warnanagar area of Kolhapur District (Maharashtra) India. For evaluating the water quality, they used several ratio parameters (all ions in mEq/L), such as sodium adsorption ratio ($\left\{Na/\sqrt{Ca+Mg}/2\right\}$), sodium percentage ($\{Na/(Na + K + Ca + Mg)\}$), and corrosivity ratio ($\{0.5 \times (Cl + SO_4)/(HCO_3 + CO_3)\}$), among others. These variables will be reflected in different hybrid water types and, therefore, these quality parameters could eventually be replaced by the multidimensional hybrid water types.

Although a new robust multidimensional scheme for water classification has been achieved, there is still need for increasing the dimensions of the solution, which should constitute a future study.

3.6. Higher dimensional classification schemes

From the results and discussion of Sections 3.1 and 3.2, it is clear to us that the multidimensional classification not only can be applied to environmental water samples but also can be extended to higher dimensions by including additional variables, such as salinity, pH, As, Cr, Li, Fe, Al, SiO_2, Pb, Se, Hg, P, S, NH_3, NH_4, NO_x, PO_x, H_xC_x, and probably isotopic compositions, among others. Total dissolved solids (TDS) will be automatically represented by so many new and existing chemical parameters. We would have to extend the present Monte Carlo methodology and the LDA and canonical analysis to extend the scheme and thus increase its applicability to all kinds of environmental problems. This would not be a simple job but could be handled following the recent methodology put forth by Verma (2020b), who was able to solve the problem of 29 classes in a 19-step process.

The proposed future work should be helpful in environmental studies, such as the presence of formaldehyde (CH_2O) in rain water in Mexico city atmosphere sampled during 1981 and 1982 (Baez et al., 1984) and the spill of geothermal fluids at the geothermal field of the Los Azufres geothermal field, Michoacán, Mexico (Birkle and Merkel, 2000).

On the other hand, the quadratic discriminant analysis (QDA; Srivastava et al., 2007) could probably be a more powerful technique than the LDA used here. Unfortunately, no commercial software is available to efficiently handle QDA of many (thousands of) samples. Therefore, this should constitute a future effort in an extended higher dimensional water classification scheme.

3.7. Applications for field samples

We present specific cases to show the nomenclature of actual water samples from the *7 hlr model* as well as the *7 M conc model* and the initial *Greater molar conc model* obtained from the online computer program *WaterMClaSys_LDA* available at http://tlaloc.ier.unam.mx. The related Excel® files can be downloaded from this web portal after registration and log-in.

3.7.1. Groundwater from India

Kumar (2013) reported compositional data of 12 water samples from the Tamil Nadu state in South India and used Hill-Piper and Chadha's diagrams for the nomenclature. These data were compiled and processed in *WaterMClaSys_LDA*. The compositional data and the resulting nomenclature are summarized in Table 4. The initial *Greater molar conc* based nomenclature indicated that the Tamil Nadu water samples were distributed as follows: sodium bicarbonate (5 samples), sodium chloride (6 samples), and magnesium bicarbonate (1 sample). The multidimensional *7 hlr model* ("basic" option) suggested the

Fig. 5. Discriminant function (*DF*1h-*DF*2h) diagrams for the molar concentrations of the testing database corresponding to the cation (**a** to **d**) and anion (**e** to **h**) classes; the symbols and percent success values are explained in insets (the centroids of the training set samples are shown with larger filled symbols). (**a**) classification of Ca, Mg, and Na; (**b**) classification of Ca, Mg, and K; (**c**) classification of Ca, Na, and K; (**d**) classification of Mg, Na, and K; (**e**) classification of SO_4, Cl, and HCO_3; (**f**) classification of SO_4, Cl, and CO_3; (**g**) classification of SO_4, HCO_3, and CO_3; and (**h**) classification of Cl, HCO_3, and CO_3.

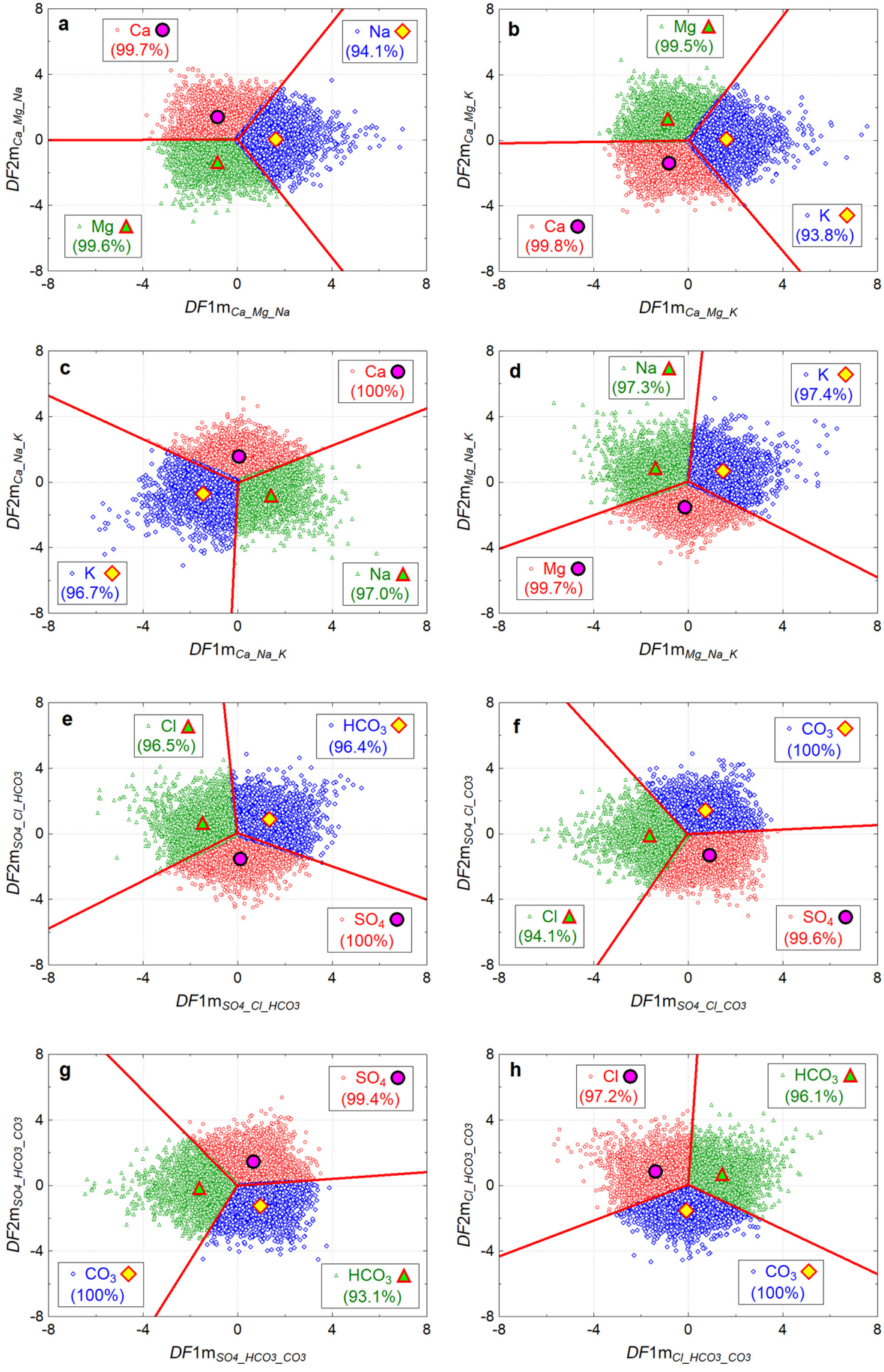

a
Ca (99.7%)
Na (94.1%)
Mg (99.6%)
DF2m Ca_Mg_Na
DF1m Ca_Mg_Na
b
Mg (99.5%)
K (93.8%)
Ca (99.8%)
DF2m Ca_Mg_K
DF1m Ca_Mg_K
c
Ca (100%)
K (96.7%)
Na (97.0%)
DF2m Ca_Na_K
DF1m Ca_Na_K
d
Na (97.3%)
K (97.4%)
Mg (99.7%)
DF2m Mg_Na_K
DF1m Mg_Na_K
e
Cl (96.5%)
HCO3 (96.4%)
SO4 (100%)
DF2m SO4_Cl_HCO3
DF1m SO4_Cl_HCO3
f
CO3 (100%)
Cl (94.1%)
SO4 (99.6%)
DF2m SO4_Cl_CO3
DF1m SO4_Cl_CO3
g
SO4 (99.4%)
CO3 (100%)
HCO3 (93.1%)
DF2m SO4_HCO3_CO3
DF1m SO4_HCO3_CO3
h
Cl (97.2%)
HCO3 (96.1%)
CO3 (100%)
DF2m Cl_HCO3_CO3
DF1m Cl_HCO3_CO3

Table 4
Application of the online computer program *WaterMClaSys_LDA* for the nomenclature of the groundwater samples from Tamil Nadu, India (the chemical data from Kumar, 2013).

Sample		Chemical composition (mg/L)[a]								Water nomenclature[b]		
number	identification	Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	*Greater molar conc model*	*7 hlr model* (basic)	*7 hlr model* ("basic + hybrid")
1	GW1	40	36	258	4	132	188	336	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-chloride
2	GW2	38	53	161	86	12	138	573	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-chloride
3	GW3	60	100	115	18	194	284	165	1	sodium chloride	sodium chloride	sodium-magnesium chloride
4	GW4	40	86	182	1	187	273	110	1	sodium chloride	sodium chloride	sodium-magnesium chloride
5	GW5	34	102	161	113	29	372	329	24	sodium chloride	sodium chloride	sodium chloride
6	GW6	22	80	133	3	14	184	311	24	sodium chloride	sodium bicarbonate	sodium-magnesium bicarbonate-chloride
7	GW7	50	52	145	26	78	213	140	1	sodium chloride	sodium chloride	sodium chloride-bicarbonate
8	GW8	32	32	113	6	6	53	361	14	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-chloride
9	GW9	36	74	107	10	52	142	348	1	sodium bicarbonate	sodium bicarbonate	sodium-magnesium bicarbonate-chloride
10	GW10	28	87	322	34	58	369	580	1	sodium chloride	sodium bicarbonate	sodium bicarbonate-chloride
11	GW11	30	29	14	2	7	21	183	2	magnesium bicarbonate	magnesium bicarbonate	magnesium-sodium bicarbonate-chloride
12	GW12	20	77	271	4	91	230	427	42	sodium bicarbonate	sodium bicarbonate	sodium-magnesium bicarbonate-chloride

[a] The unrealistic 0 concentrations were replaced by a probable lower limit of detection assumed to be 1 mg/L.
[b] The *Greater molar conc* would provide the water nomenclature of up to 16 classes, the *7 hlr* (basic) nomenclature refers to a total of 16 classes, and the *7 hlr* ("basic + hybrid") can provide up to 256 water classes.

following water types: sodium bicarbonate (7 samples), sodium chloride (4 samples), and magnesium bicarbonate (1 sample).

The recommended *7 hlr model* ("basic+hybrid" default option) suggested the following water classification (Table 4): sodium bicarbonate-chloride (4 samples), sodium-magnesium chloride (2 samples), sodium chloride (1 sample), sodium-magnesium bicarbonate-chloride (3 samples), sodium chloride-bicarbonate (1 sample), and magnesium-sodium bicarbonate-chloride (1 sample). Given the wide geographical area (of over 100,000 km^2) from which these groundwater samples were collected (12 samples from individual sites in the entire state of Tamil Nadu; Kumar, 2013), the "basic+hybrid" classification should be accepted as more likely. In other words, the water samples are likely to have diverse names as inferred in the last column of Table 4. This example clearly supports the routine use of the *7 hlr* "basic+hybrid" *model* classification for natural water samples.

3.7.2. Lake water from Mongolia

Chemical compositions of 31 samples from 14 saline lakes of the Gobi Desert region, Western Mongolia, were presented by Bayanmunkh et al. (2017). These authors used Hill-Piper diagram (Piper, 1944) without providing any chemical nomenclature for their water samples. The data of relevant elements and all three classification results are presented in Table 5. The recommended classification (*7 hlr* "basic+hybrid" *model*) can be summarized as follows: calcium-magnesium sulfate (1 sample); magnesium carbonate (2); magnesium-sodium carbonate (2); sodium carbonate (10); sodium carbonate-chloride (2); sodium chloride-sulfate (3); sodium sulfate (3); sodium sulfate-carbonate (5); and sodium sulfate-chloride (3).

Bayanmunkh et al. (2017) mentioned that 14 lakes were sampled as follows (we will use only the ID and not the lake names): B1, B2 and B3 brackish; B4, B5, B6, S1 and S2 saline; and S3, S4, H1, H2, H3 and H4 brine (synthesis according to their Table 1 general description). There was some discrepancy in the description provided by these authors, because later (according to their Table 2 actual salinity analysis of individual samples) they assigned letters B, S and H, respectively, to brackish, saline and brine. We will use this assignment in out discussion.

Twelve brackish water (having salinity between fresh water and seawater) samples (identified by letter B in Table 5) were sampled from 6 different lakes (Bayanmunkh et al., 2017). Two samples from the lake identified as B1 in Table 5 (B1–0 and B1–1) were both magnesium carbonate; 2 from B2 sodium chloride-sulfate; 2 from B3 sodium carbonate; two from B4 magnesium-sodium carbonate; 2 from B5 calcium-magnesium sulfate and sodium sulfate; and 2 from B6 sodium sulfate. Thus, the nomenclature showed great consistency for 5 lakes (the same names for duplicate samples), with the exception of lake B5 (different names for duplicate samples although sulfate was common to both of them; besides, these 2 samples B5–0 and B5–1 have drastically different Na concentrations; Table 5).

The nomenclature of saline waters (letter S in Table 5) was fully mutually consistent. All 5 samples from lake S1 were sodium sulfate-carbonate; 3 from S2 sodium carbonate; 2 from S3 sodium sulfate-chloride; and 2 from S4 sodium carbonate.

The brine waters (letter H in Table 5) were also classified consistently as follows: 2 from lake H1 as sodium carbonate; 1 from H2 sodium carbonate; 2 from H3 sodium chloride-sulfate; and 2 from H4 as sodium carbonate-chloride and sodium chloride-carbonate (different names in only this case obey relatively different proportions of anions Cl and CO_3 as observed in Table 5 for these chemical variables).

From this brief discussion, we can infer that the recommended *7 hlr* ("basic+hybrid") *model* provided fully consistent nomenclature. Once again, we stress the use of this model for the nomenclature of natural water samples.

3.7.3. Geothermal water from Turkey

Bayram and Gultekin (2010) presented geochemical data for 60 geothermal water samples from thermal springs and wells in western Turkey. The geochemical data and results of water nomenclature are summarized in Table 6. We comment only on the recommended nomenclature (*7 hlr* "basic + hybrid" *model*) presented in the last column of Table 6. The locations of the sampled thermal springs are widely distributed in western Turkey. The 45 samples were classified as follows (in descending order): calcium bicarbonate (10 samples), calcium-magnesium bicarbonate (10), sodium-potassium bicarbonate (7), calcium-sodium bicarbonate (5), sodium bicarbonate (3), sodium-calcium bicarbonate (3), sodium-calcium bicarbonate-chloride (2), calcium-sodium bicarbonate-carbonate (1), calcium-sodium bicarbonate-sulphate (1), magnesium-calcium bicarbonate (1), sodium bicarbonate-sulphate (1), and sodium-calcium chloride-bicarbonate (1). The 15 well water samples were distributed as follows: sodium bicarbonate-sulphate (8), sodium bicarbonate (6), and calcium

Fig. 6. Discriminant function (*DF*1m-*DF*2m) diagrams for the molar concentrations of the testing database corresponding to the cation (**a** to **d**) and anion (**e** to **h**) classes; the symbols and percent success values are explained in insets (the centroids of the training set samples are shown with larger filled symbols). (**a**) classification of Ca, Mg, and Na; (**b**) classification of Ca, Mg, and K; (**c**) classification of Ca, Na, and K; (**d**) classification of Mg, Na, and K; (**e**) classification of SO_4, Cl, and HCO_3; (**f**) classification of SO_4, Cl, and CO_3; (**g**) classification of SO_4, HCO_3, and CO_3; and (**h**) classification of Cl, HCO_3, and CO_3.

Table 5
Application of the online computer program *WaterMClaSys_LDA* for the nomenclature of saline lakes of the Gobi Desert region, Mongolia (the chemical data from Bayanmunkh et al., 2017).

Sample		Chemical composition (mg/L)[a]								Water nomenclature[b]		
number	identification	Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	*Greater molar conc model*	*7 hlr model* (basic)	*7 hlr model* ("basic + hybrid")
1	B1–0	30	130	40	10	100	30	560	12,000	magnesium carbonate	magnesium carbonate	magnesium carbonate
2	B1–1	30	130	40	10	100	30	570	39,000	magnesium carbonate	magnesium carbonate	magnesium carbonate
3	B2–0	110	80	280	10	400	380	190	1	sodium chloride	sodium chloride	sodium chloride-sulfate
4	B2–1	110	50	330	10	400	350	170	1	sodium chloride	sodium chloride	sodium chloride-sulfate
5	B3–0	40	160	660	30	700	30	260	6000	sodium carbonate	sodium carbonate	sodium carbonate
6	B3–1	50	160	760	30	600	30	280	6000	sodium carbonate	sodium carbonate	sodium carbonate
7	B4–0	1	330	740	60	500	30	1810	126,000	sodium carbonate	magnesium carbonate	magnesium-sodium carbonate
8	B4–1	1	350	760	80	300	30	1700	147,000	sodium carbonate	magnesium carbonate	magnesium-sodium carbonate
9	B5–0	480	240	100	40	2000	80	120	1	calcium sulfate	calcium sulfate	calcium-magnesium sulfate
10	B5–1	470	250	900	40	2000	70	100	1	sodium sulfate	sodium sulfate	sodium sulfate
11	B6–0	360	260	1300	30	2000	100	130	1	sodium sulfate	sodium sulfate	sodium sulfate
12	B6–1	350	270	1200	30	2000	100	130	1	sodium sulfate	sodium sulfate	sodium sulfate
13	S1–0	500	1220	6700	110	10,000	740	290	6000	sodium sulfate	sodium sulfate	sodium sulfate-carbonate
14	S1–1	500	1370	6400	110	10,000	780	40	6000	sodium sulfate	sodium sulfate	sodium sulfate-carbonate
15	S1–2	500	1370	6300	110	10,000	820	160	6000	sodium sulfate	sodium sulfate	sodium sulfate-carbonate
16	S1–3	250	1370	6800	110	12,000	760	160	6000	sodium sulfate	sodium sulfate	sodium sulfate-carbonate
17	S1–4	500	1220	7300	130	11,000	800	160	6000	sodium sulfate	sodium sulfate	sodium sulfate-carbonate
18	S2–0	70	510	12,000	340	16,000	1950	400	41,000	sodium carbonate	sodium carbonate	sodium carbonate
19	S2–1	80	500	13,000	350	15,000	1930	420	35,000	sodium carbonate	sodium carbonate	sodium carbonate
20	S2–2	80	510	13,000	320	16,000	1840	400	39,000	sodium carbonate	sodium carbonate	sodium carbonate
21	S3–0	1250	2280	14,000	90	6000	3810	60	1	sodium chloride	sodium sulfate	sodium sulfate-chloride
22	S3–1	1000	2130	15,000	100	6000	3790	60	1	sodium chloride	sodium sulfate	sodium sulfate-chloride
23	S4–0	750	9120	25,000	560	26,000	3230	390	54,000	sodium carbonate	sodium carbonate	sodium carbonate
24	S4–1	750	8970	92,000	4100	26,000	3230	390	63,000	sodium carbonate	sodium carbonate	sodium carbonate
25	H1–0	250	10,790	45,000	1100	28,000	8860	450	63,000	sodium carbonate	sodium carbonate	sodium carbonate
26	H1–1	250	15,960	66,000	1600	20,000	13,260	580	81,000	sodium carbonate	sodium carbonate	sodium carbonate
27	H2–0	500	17,020	79,000	980	24,000	16,840	700	135,000	sodium carbonate	sodium carbonate	sodium carbonate
28	H3–0	1000	31,310	25,000	550	21,000	30,130	780	1	magnesium chloride	sodium chloride	sodium chloride-sulfate
29	H3–1	2000	32,830	110,000	4500	21,000	30,200	750	1	sodium chloride	sodium sulfate	sodium sulfate-chloride
30	H4–0	500	12,770	83,000	2400	26,000	19,850	420	36,000	sodium carbonate	sodium carbonate	sodium carbonate-chloride
31	H4–1	500	14,590	84,000	2500	26,000	22,510	490	24,000	sodium chloride	sodium chloride	sodium chloride-carbonate

[a] The unrealistic 0 concentrations were replaced by a probable lower limit of detection assumed to be 1 mg/L.
[b] The *Greater molar conc* would provide the water nomenclature of up to 16 classes, the *7 hlr* (basic) nomenclature refers to a total of 16 classes, and the *7 hlr* ("basic + hybrid") can provide up to 256 water classes.

bicarbonate (1). The bicarbonate seems to be the most common anion specie in geothermal waters of western Turkey. We can, therefore, conclude that the *7 hlr* ("basic+hybrid") *model* should constitute the recommended procedure for the water nomenclature.

4. Conclusions

Refined Monte Carlo simulation procedure was used for successfully generating thousands of synthetic water samples having the composition of 8 elements (4 cations Ca, Mg, Na, and K and 4 anions SO_4, Cl, HCO_3, and CO_3; all in mg/L). All sample compositions were converted to mEq/L and then adjusted for ionic charge imbalance almost perfectly at better than ±0.00004%. The initial class assignment of 16 classes was based on the perfectly charge-balanced molar concentrations (mM/L) of simulated water samples. The multi-normality required by the LDA and canonical analysis was achieved in terms of 7 hybrid log-ratio (hlr) transformed variables. A new multidimensional water classification scheme was successfully proposed from the LDA and canonical analysis. This multivariate technique was also applied to the original molar concentrations of 7 elements (without transformation). In summary, three classification models (*7 hlr model*, *7 M conc*, and initial class assignment called *Greater molar conc model*) were proposed and evaluated. The high percent success values of the classification of 16 water classes from the multidimensional *7 hlr model* varied from 94.06% to 95.38%. The *7 M conc model* provided still higher percent success of 97.75% to 98.25%. The robustness tests against analytical error or uncertainty propagation and mineral-water interaction showed generally higher robustness of the *7 hlr model*, which constitutes the recommended model for the water nomenclature. The consideration of the overall probabilities for 4 cations and 4 anions further facilitated to put forth additional conditions for identifying 16 basic (or primary) water types as well as additional hybrid water classes. Thus, from the *7 hlr* ("basic+hybrid") *model*, a total of 256 water classes can be recognized for the first time in the literature.

The possible bias and overfitting were evaluated from independent random testing database. The classification scheme was shown to be free from bias and overfitting because similarly high percent success values were obtained from independent testing as from training. Applications of the new classification scheme to groundwaters, lake waters, and geothermal waters showed a consistent water nomenclature from the *7 hlr* ("basic+hybrid") *model*. A new online computer program WaterMClaSys_*LDA* available at our web portal http://tlaloc.ier.unam.mx, will facilitate the use of the new complex multidimensional scheme for the water nomenclature as 256 classes. Although this scheme can be applied to classify environmental water samples, the present multidimensional scheme should be extended in future to higher dimensions by including more variables in the LDA and canonical analysis.

CRediT authorship contribution statement

Surendra P. Verma: Conceptualization, Methodology, Figures and Tables Preparation, Writing-Original and Revised Manuscript Preparation, Literature Search and Evaluation, Reviewing and Editing of the Original and Revised Manuscript.

Oscar Alejandro Uscanga-Junco: Monte Carlo Simulations, Data Processing, Figures and Tables Preparation, Software Development and Validation, Literature Search and Evaluation, Reviewing and Editing of the Original and Revised Manuscript.

Lorena Díaz-González: Conceptualization, Software Testing and Validation, Literature Search, Applications Search, Figures and

Table 6
Application of the online computer program *WaterMClaSys_LDA* for the nomenclature of geothermal waters from Turkey (the chemical data from Bayram and Gultekin, 2010).

Sample		Chemical composition (mg/L)[a]								Water nomenclature[b]		
number	identification	Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	*Greater molar conc model*	*7 hlr model* (basic)	*7 hlr model* ("basic + hybrid")
1	Cold_1	50.41	22.34	13.2	13.37	27	8	157.8	21	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
2	Cold_2	37.24	6.59	13.8	13.12	17	7	126.2	10.5	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate
3	Cold_3	6.57	1	14.64	12.62	6	7	47.3	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
4	Cold_4	8.77	5.26	15.48	13.45	8	4	47.3	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
5	Cold_5	70.15	17.06	14.64	12.62	25	7	200	10.5	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
6	Cold_6	87.69	42.05	13.37	12.62	42	8	284	31.5	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
7	Cold_7	6.57	2.62	21.39	14.28	6	8	57.8	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
8	Cold_8	2.2	2.62	15.48	15.52	5	6	42	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
9	Cold_9	2.2	2.62	14.64	17.18	2.5	4	42	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
10	Cold_10	2.2	2.62	16.33	13.53	5	4	42	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
11	Cold_11	120.52	49.92	24.17	18.42	90	17	352.4	42	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
12	Cold_12	65.75	6.58	69.91	16.76	15	16	200	52.6	sodium bicarbonate	sodium bicarbonate	sodium-calcium bicarbonate
13	Cold_13	65.75	26.28	32.36	13.62	66	10	184	42	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate
14	Cold_14	65.75	32.83	12.95	12.62	16	10	210	21	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
15	Cold_15	59.18	14.45	27.3	12.62	16	10	121	31.6	calcium bicarbonate	sodium bicarbonate	sodium-calcium bicarbonate
16	Cold_16	35.08	11.81	34.05	14.03	12	10	115.7	21	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
17	Cold_17	87.69	3.94	14.64	14.86	25	7	136.7	42	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
18	Cold_18	15.34	10.51	26.45	14.03	35	12	42	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
19	Cold_19	105.23	35.47	34.05	18.5	17	13	363	21	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate
20	Cold_20	92.09	30.19	23.08	13.12	40	11	221	52.6	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
21	Cold_21	87.69	22.32	18.01	13.7	16	18	231	21	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
22	Cold_22	26.31	15.77	31.52	14.03	5	12	100	21	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
23	Cold_23	113.99	3.84	38.27	26.87	80	30	152.5	115.7	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate-carbonate
24	Cold_24	65.75	26.28	50.08	16.93	19	53	131.5	52.6	sodium bicarbonate	sodium bicarbonate	sodium-calcium bicarbonate-chloride
25	Cold_25	65.75	3.94	26.45	13.78	6	10	152.5	21	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate
26	Cold_26	94.25	19.68	28.14	12.71	33	12	294.5	21	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate
27	Cold_27	35.08	9.22	24.26	14.36	27	17	115.7	10.5	sodium bicarbonate	sodium bicarbonate	sodium-calcium bicarbonate
28	Cold_28	98.7	72.29	18.52	13.12	168	13	273.5	73.6	magnesium bicarbonate	magnesium bicarbonate	magnesium-calcium bicarbonate
29	Cold_29	0.56	0.31	16.33	12.62	22	10	52.6	1	sodium bicarbonate	sodium bicarbonate	sodium-potassium bicarbonate
30	Cold_30	41.64	3.94	18.86	13.04	165	8	126.2	21	calcium bicarbonate	calcium bicarbonate	calcium-sodium bicarbonate-sulphate
31	Cold_31	21.9	5.26	23.08	13.87	11	100	73.6	10.5	sodium chloride	sodium chloride	sodium-calcium chloride-bicarbonate
32	Cold_40	86	19	14.97	9	22.48	30.13	307.44	18.3	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
33	Cold_41	97	29	8	8	28.05	10.64	387.35	18.3	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
34	Cold_42	87	14	14.3	7	8.99	23.04	270.23	18.3	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
35	Cold_46	57	31	3.5	8	29.07	3.55	322.93	1	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
36	Cold_47	42	11	2.5	7	13.35	1.77	173.42	1	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
37	Cold_48	29	4	3.8	6	7.72	3.55	113.64	1	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
38	Cold_49	90	27	9.2	4	42.22	5.32	370.75	1	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
39	Cold_51	15.34	13.13	29.41	14.28	19	13	63	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
40	Cold_52	61.38	28.9	12.95	12.95	15	10	205	21	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
41	Cold_54	48	11	2.3	10	10.8	3.55	160.43	12	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
42	Cold_55	68	31	4.4	9	27.69	8.86	283.04	18.3	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
43	Cold_56	36	6	7.1	10	8.1	13.47	117.12	1	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
44	Cold_60	48	19	29.1	12	35.73	58.49	179.4	1	sodium bicarbonate	sodium bicarbonate	sodium-calcium bicarbonate-chloride
45	Cold_61	92	26	8.1	7	37.95	5.32	394.73	1	calcium bicarbonate	calcium bicarbonate	calcium-magnesium bicarbonate
46	Eynal_37	25	10	450	75	308.85	83.31	498.98	84.9	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
47	Eynal_43	25	7.5	380.1	42.5	377.82	63.81	552.78	64.71	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
48	Eynal_44	39	2	2.3	5	7.55	1.77	125.6	1	calcium bicarbonate	calcium bicarbonate	calcium bicarbonate
49	Eynal_50	53	5	510	40	448.52	81.53	744.99	34.62	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
50	Eynal_57	57.5	10	300	57.5	350.49	58.49	556.2	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
51	Eynal_62	65	7.5	520	40	529.53	77.99	592.43	121.17	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
52	Eynal_63	127.5	7.5	535	55	610.01	88.62	445.78	213.45	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
53	Citgol_39	35	5	360	35	285.14	69.13	403.21	84.9	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
54	Citgol_58	30	5	342.6	42.5	352.2	62.04	421.21	64.71	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
55	Citgol_64	54	7	277.5	20	323.54	53.17	416.44	28.86	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
56	Nasa_38	70	15	302.5	30	254.2	60.27	566.69	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
57	Nasa_45	60	10	325	35	324.87	54.55	604.02	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
58	Nasa_53	80	15	362.5	30	256.36	65.58	782.63	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate
59	Nasa_59	72.5	12.5	306.1	45	309.49	54.94	669.78	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate
60	Nasa_65	88	9	335	25	346.26	56.72	680.39	1	sodium bicarbonate	sodium bicarbonate	sodium bicarbonate-sulphate

[a] The unrealistic 0 concentrations were replaced by a probable lower limit of detection assumed to be 1 mg/L.
[b] The *Greater molar conc* would provide the water nomenclature of up to 16 classes, the *7 hlr* (basic) nomenclature refers to a total of 16 classes, and the *7 hlr* ("basic + hybrid") can provide up to 256 water classes.

Tables Preparation, Reviewing and Editing of the Original and Revised Manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors. The second author (OAU-J) thanks CONACyT for the scholarship awarded for his Master´s studies. We are grateful to the Associate Editor José Virgílio Cruz for efficiently handling our manuscript and three anonymous reviewers of the journal for numerous constructive comments on an earlier version of our presentation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.141704.

References

Agrawal, S., 1999. Geochemical discrimination diagrams: a simple way of replacing eye-fitted boundaries with probability based classifier surfaces. J. Geol. Soc. India 54, 335–346.
Agrawal, S., Verma, S.P., 2007. Comment on "Tectonic classification of basalts with classification trees" by Pieter Vermeesch (2006). Geochim. Cosmochim. Acta 71, 3388–3390.
Ahmad, N., Sen, Z., Ahmad, M., 2003. Ground water quality assessment using multi-rectangular diagrams. Groundwater 41, 828–832.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London, UK (416 p).
Al-Bassam, A.M., Khalil, A.R., 2012. DurovPwin: a new version to plot the expanded Durov diagram for hydro-chemical data analysis. Comput. Geosci. 42, 1–6.
Anderson, D.L., 1999. A theory of the earth: Hutton and Humpty Dumpty and Holmes. In: Craig, G.Y., Hull, J.H. (Eds.), James Hutton - Present and Future. vol. 150. Royal Geological Society, London, pp. 13–35.
Azhar, S.C., Aris, A.Z., Yusoff, M.K., Ramli, M.F., Juahir, H., 2015. Classification of river water quality using multivariate analysis. Procedia Environ. Sci. 30, 79–84.
Baez, A.P., Belmont, R., Rosas, I., 1984. Formaldehyde in rain water in Mexico city atmosphere. Geofis. Int. 23, 449–465.
Barnett, V., Lewis, T., 1994. Outliers in statistical data. 3rd edn. John Wiley & Sons, Chichester (584 p).
Bayanmunkh, B., Sen-Lin, T., Narangarvuu, D., Ochirkhuyag, B., Bolormaa, O., 2017. Physico-chemical composition of saline lakes of the Gobi Desert region, Western Mongolia. J. Earth Sci. Clim. Change 8. https://doi.org/10.4172/2157-7617.1000388.
Bayram, A.F., Gultekin, S.S., 2010. Classifying of the Simav geothermal waters with artificial neural network method. In proceedings world geothermal congress, Bali. Indonesia 25–29.
Bevington, P.R., 1969. Data Reduction and Error Analysis for the Physical Sciences. McGraw Hill Book Company, New York (336 p).
Bevington, P.R., Robinson, D.K., 2003. Data Reduction and Error Analysis for the Physical Sciences. Third edition. McGraw Hill, Boston (320 p).
Birkle, P., Merkel, B., 2000. Environmental impact by spill of geothermal fluids at the geothermal field of Los Azufres, Michoacán. Mexico. Water Air Soil Pollut 124, 371–410.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C., 1983. Classification and Regression Trees. Chapman & Hall/CRC, Boca Raton (357 p).
Brombacch, T., Marini, L., Hunziker, J.C., 2000. Geochemistry of the thermal springs and fumaroles of Basse-Terre island, Guadeloupe, Lesser Antilles. Bull. Volcanol. 61, 477–490.
Busico, G., Kazakis, N., Cuoco, E., Colombani, N., Tedesco, D., Voudouris, K., Mastrocicco, M., 2020. Novel hybrid method of specific vulnerability to anthropogenic pollution using multivariate statistical and regression analyses. Water Res. 171, 115386.
Butler, J.C., 1979. Trends in ternary petrologic variation diagrams - fact or fantasy? Am. Mineral. 64, 1115–1121.
Castells, R.C., Castillo, M.A., 2000. Systematic errors: detection and correction by means of standard calibration, Youden calibration and standard addition method in conjunction with a method response model. Anal. Chim. Acta 423, 179–185.
Chadha, D.K., 1999. A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. Hydrogeol. J. 7, 431–439.
Chayes, F., 1960. On correlation between variables of constant sum. J. Geophys. Res. 65, 4185–4193.
Chayes, F., 1971. Ratio Correlation. A Manual for Students of Petrology and Geochemistry. The University of Chicago Press, Chicago and London (108 p).
D'Amore, F., Scandiffio, G., Panichi, C., 1983. Some observations on the chemical classification of ground waters. Geothermics 12, 141–148.
Durov, S.A., 1948. Natural waters and graphic representation of their compositions. Dokl. Akad. Nauk SSSR 59, 87–90.
Elhag, A.B., 2016. New diagram useful for classification of groundwater quality. British Journal of Earth Sciences Research 4, 49–54.
Giggenbach, W.F., Glover, R.B., 1992. Tectonic regime and major processes governing the chemistry of water and gas discharges from the Rotorua geothermal field, New Zealand. Geothermics 21, 121–140.
Giménez-Forcada, E., 2010. Dynamic of sea water interface using hydrochemical facies evolution diagram. Ground Water 48, 212–216.
Golekar, R.B., Akshay, M., Shubham, J., Shubham, A., Patil, Y.M., 2017. Geochemical characteristics of water and its suitability for drinking and irrigation use in and around Warnanagar area of Kolhapur District (Maharashtra) India. Journal of Water Resources and Pollution Studies 2, 1–12.
Güler, C., Thyne, G.D., McCray, J.E., Turner, A.K., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. 10, 455–474.
Hand, D.J., 1981. Discrimination and Classification. John Wiley & Sons, Chichester (218 p).
Hand, D.J., 1997. Construction and Assessment of Classification Rules. Wiley, New York (214 p).
Handa, B.K., 1965. Modified Hill-piper diagram for presentation of water analysis data. Curr. Sci. 34, 131–314.
Hill, R.A., 1940. Geochemical patterns in Coachella Valley. Trans. Am. Geophys. Union, Part I 21, 46–49.
Kemp, P.H., 1971. Chemistry of natural waters – VI classification of waters. Water Res. 5, 943–956.
Kumar, P.J.S., 2013. Interpretation of groundwater chemistry using Piper and Chadha's diagrams: a comparative study from Perambalur Taluk. Elixir Geosci. 54, 12208–12211.
Lavine, B.K., Mirjankar, N., 2012. Clustering and classification of analytical data. Encyclopedia of Analytical Chemistry. John Wiley & Sons https://doi.org/10.1002/9780470027318.a5204.pub2.
Law, A.M., Kelton, W.D., 2000. Simulation modeling and analysis. Third edition. McGraw Hill, Boston (760 p).
Lee, T.-C., 1998. LEEGRAM: a program for normalized Stiff diagrams and quantification of grouping hydrochemical data. Comput. Geosci. 24, 523–529.
Lloyd, J.W., 1965. The hydrochemistry of the aquifers of northeastern Jordan. J. Hydrol. 3, 319–330.
Loh, W.-Y., 2014. Fifty years of classification and regression trees. Int. Stat. Rev. 82, 329–348.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, San Diego, CA (518 p).
Marini, F., 2010. Classification methods in chemometrics. Curr. Anal. Chem. 6, 72–79.
McNeil, V.H., Cox, M.E., Preda, M., 2005. Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia. J. Hydrol. 310, 181–200.
Miller, J.N., Miller, J.C., 2010. Statistics and Chemometrics for Analytical Chemistry. 6th edn. Pearson Prentice Hall, Essex CM20 2JE, England (271 p).
Morrison, D.F., 1990. Multivariate Statistical Methods. Third edn. McGraw-Hill Publishing Co, New York (495 p).
Pérez-Espinosa, R., Pandarinath, K., Hernández-Campos, F.J., 2019. CCWater - a computer program for chemical classification of geothermal waters. Geosci. J. 23, 261–635.
Piper, A.M., 1944. A graphic procedure in the geochemical interpretation of water-analyses. Trans. Am. Geophys. Union 25, 914–923.
Rao, N.S., 1998. MHPT.BAS: a computer program for modified Hill-Piper diagram for classification of ground water. Comput. Geosci. 24, 991–1008.
Ray, R.K., Mukherjee, R., 2008. Reproducing the Piper trilinear diagram in rectangular coordinates. Groundwater 46, 893–896.
Reyment, R.A., Savazzi, E., 1999. Aspects of Multivariate Statistical Analysis in Geology. Elsevier, Amsterdam (285 p).
Romani, S., 1981. A new diagram for classification of natural waters and interpretation of chemical analyses data. In Quality of Groundwater, Proceedings of an International Symposium (eds W. van Duijvenbooden, P. Glaebergen, H. van Lelyveld), (Noordwijkerhout, The Netherlands).
Sadashivaiah, C., Ramakrishnaiah, C.R., Ranganna, G., 2008. Hydrochemical analysis and evaluation of groundwater quality in Tumkur Taluk, Karnataka state, India. Int. J. Environ. Res. Public Health 5, 158–164.
Shelton, J.L., Englea, M.A., Buccianti, A., Blondes, M.S., 2018. The isometric log-ratio (ilr)-ion plot: a proposed alternative to the Piper diagram. J. Geochem. Explor. 190, 130–141.
Sheth, H.C., Torres-Alvarado, I.S., Verma, S.P., 2002. What is the "calc-alkaline rock series"? Int. Geol. Rev. 44, 686–701.
Shterev, K.D., 1973. Genetic-substantial classification of the exogenic mineral waters (hydromineral solutions) (actual pattern, interpreted in space and time). Sofia Chart Report (unpublished).
Srivastava, S., Gupta, M.R., Frigyik, B.A., 2007. Bayesian quadratic discriminant analysis. J. Mach. Learn. Res. 8, 1277–1305.
Stiff Jr, H. A., 1951. The interpretation of chemical water analysis by means of patterns. J. Petrol. Technol. 3(10), 15–3. DOI: https://doi.org/10.2118/951376-G.
Stuyfzand, P.J., 1989. A new hydrochemical classification of water types. Regional characterization of water quality. Proceedings of the Baltimore Symposium. vol. 182. International Association of Hydrological Sciences, pp. 89–98.
Teng, W.C., Fong, K.L., Shenkar, D., Wilson, J.A., Foo, D.C.Y., 2016. Piper diagram – a novel visualisation tool for process design. Chem. Eng. Res. Des. 112, 132–145.

Timofeev, R., 2004. Classification and Regression Trees (CART) Theory and Applications. M.A. thesis. Humboldt University, Berlin, Berlin (39 p).
Verma, S.P., 2012a. Geochemometrics. Rev. Mex. Cienc. Geol 29, 276–298.
Verma, S.P., 2012b. Application of multi-dimensional discrimination diagrams and probability calculations to acid rocks from Portugal and Spain. Comput. Geol. 99, 79–93.
Verma, M.P., 2013. IAEA inter-laboratory comparisons of geothermal water chemistry: critiques on analytical uncertainty, accuracy, and geothermal reservoir modeling of Los Azufres, Mexico. J. Iber. Geol. 31, 57–72.
Verma, S.P., 2015. Monte Carlo comparison of conventional ternary diagrams with new log-ratio bivariate diagrams and an example of tectonic discrimination. Geochem. J. 49, 393–412.
Verma, S.P., 2020a. Road from Geochemistry to Geochemometrics. Springer, Singapore (669 p).
Verma, S.P., 2020b. Comprehensive multidimensional tectonomagmatic discrimination from log-ratio transformed major and trace elements. Lithos 362–363, 105476.
Verma, S.P., Quiroz-Ruiz, A., 2006. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. Rev. Mex. Cienc. Geol 23, 133–161.
Verma, S.P., Rivera-Gómez, M.A., Díaz-González, L., Quiroz-Ruiz, A., 2016. Log-ratio transformed major-element based multidimensional classification for altered High-Mg igneous rocks. Geochem. Geophys. Geosys. 17, 4955–4972.
Verma, S.P., Rosales-Rivera, M., Rivera-Gómez, M.A., Verma, S.K., 2019. Comparison of matrix-effect corrections for ordinary and uncertainty weighted linear regressions and determination of major element mean concentrations and total uncertainties of 62 international geochemical reference materials from wavelength-dispersive X-ray fluorescence spectrometry. Spectrochim. Acta Part B 162, 105714.
Xue, X., Li, J., Xie, X., Qian, K., Wang, Y., 2019. Impacts of sediment compaction on iodine enrichment in deep aquifers of the North China Plain. Water Res. 159, 480–489.

Apéndice B. Comparison of machine learning models for water multidimensional classification

Journal of Hydrology 598 (2021) 126234

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Development and comparison of machine learning models for water multidimensional classification

Lorena Díaz-González [a,*], Oscar Alejandro Uscanga-Junco [b], Mauricio Rosales-Rivera [a]

[a] Centro de Investigación en Ciencias, Universidad Autónoma de Estado de Morelos, Cuernavaca, Morelos 62209, Mexico
[b] Maestría en Ciencias, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma de Estado de Morelos, Cuernavaca, Morelos 62209, Mexico

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
Machine learning
Gradient boosting
Support vector machines
Hill-Piper diagram
Molar concentrations
Monte Carlo simulation
Groundwater samples

ABSTRACT

We proposed four new models (WClassCB, WClassVL WClassVP WClassVR) for water classification using Categorical Boosting (CatBoost) and Support Vector Machines (SVM) with three kernels: linear, polynomial, and radial basis function. The new models were compared with the recently proposed WClassHLR (7 hybrid log-ratio) model based on linear discriminant analysis and canonical analysis techniques. A training database (50,000 samples) and another independent validation database (8,000 samples) of ionic charge-balanced concentrations of 4 cations (Ca, Mg, Na, and K) and 4 anions (SO_4, Cl, HCO_3, and CO_3) were generated through Monte Carlo simulations. The initial 16 classes were assigned from the highest cation and anion molar concentrations (GMC criteria, i.e. greater molar concentration model). Seven hybrid log-ratio transformations were used as features for training and external validation of the multidimensional classification models. These models generate probability values for each of the output classes allowing us to determine hybrid water types improving the possible water types to 256. WClassCB model showed the best accuracy values in the training set. However, WClassVL model is the recommended procedure because it generalizes better than other models in the external validation set. The new models outperform the recently proposed WClassHLR with up to a 7% difference. The usefulness of all models (WClassHLR, WClassCB, WClassVL, WClassVP, WClassVR) is illustrated by four applications to groundwater samples from India and Nigeria. All models have difficulties in classifying real samples when there is more than one major cation or anion, but they can recover the classification suggesting hybrid water types. The new computer program *WaterClaSys_ML* has been developed for applying these new models.

1. Introduction

Water type based on hydrochemical facies evaluation is extremely useful in providing a preliminary idea about the complex hydrochemical processes at the subsurface (Sajil Kumar, 2013). Thus, there have been many attempts to identify their nomenclature and develop easy to use techniques (Durov, 1948; Handa, 1965; Romani, 1981; Chadha, 1999; Ahmad et al., 2003; Giménez-Forcada, 2010; Elhag, 2017; Shelton et al., 2018). Güler et al. (2002) evaluated graphical and multivariate statistical methods for water classification, such as Schoeller semi-logarithmic diagrams (Schoeller, 1955) and principal component analysis (PCA).

Even though multiple ideas have been explored, the most used tools to identify hydrochemical facies are derived from the popular Hill-Piper diagram (Piper, 1944). This technique is based on two ternary diagrams constructed from the normalized milliequivalent per liter (mEq/L) concentrations of the cations (Ca^{2+}, Mg^{2+}, $(Na^{+} + K^{+})$) and anions (SO_4^{-2}, Cl^{-}, $(HCO_3^{-} + CO_3^{2-})$) and then projected in a diamond field to identify hydrochemical facies.

Many other techniques such as Durov diagram (Durov, 1948) and Chadha diagram (Chadha, 1999) are also popular approaches to determine water classification, which use trilinear diagrams or percent values of the chemical composition. However, there are severe problems like distortion and amplification-reduction of analytical errors in these diagrams caused by closure and constant sum problems (Chayes, 1960, 1971; Aitchison, 1986). Because crude compositional variables represent a closed unit-sum constrained system and ternary diagrams impose a further unit-sum constraint on any experimental data, these diagrams become statistically unsuitable to handle experimental data (Verma, 2012). The ternary diagram problems were stressed by Aitchison (1986) and Verma (2012, 2015). Aitchison (1986) proposed solutions to overcome the constant sum difficulties through a multivariate approach by

* Corresponding author.
E-mail addresses: ldg@uaem.mx, rorm@uaem.mx (L. Díaz-González).

https://doi.org/10.1016/j.jhydrol.2021.126234
Received 11 February 2021; Received in revised form 11 March 2021; Accepted 17 March 2021
Available online 23 March 2021
0022-1694/© 2021 Elsevier B.V. All rights reserved.

Table 1
Statistical summary of simulated concentration data for training the classification models (WClassCB; WClassVL; WClassVP; and, WClassVR).

No	Water class	*n*	Ca			Mg			Na			K	
			min	max	*median*	min	max	*median*	min	max	*median*	min	max
Concentrations (mg/L)													
1	Ca-SO_4	3102	1143.3	5946.6	3240.1	0.5	2545.9	934.4	0.1	2401.4	897.9	0.7	4150.4
2	Ca-Cl	3033	716.9	5747.4	3045.5	1.3	2401.0	859.4	0.2	2427.4	814.6	1.3	4236.8
3	Ca-HCO_3	3131	891.7	5978.3	3060.4	0.5	2313.6	861.3	0.9	2561.8	831.1	0.3	4705.0
4	Ca-CO_3	3031	1070.9	5595.4	3266.2	0.1	2394.3	961.2	0.6	2655.8	890.5	0.9	4637.2
5	Mg-SO_4	3120	2.8	4145.2	1567.0	571.3	4056.4	1980.3	0.5	2756.2	866.4	0.9	4573.0
6	Mg-Cl	3143	1.4	4236.2	1432.2	600.6	3636.4	1860.3	0.3	2797.1	848.9	0.6	4409.7
7	Mg-HCO_3	3161	3.0	3903.0	1402.4	364.9	3943.8	1830.7	3.4	2606.1	827.3	0.7	4420.3
8	Mg-CO_3	3160	3.4	4079.4	1563.6	671.3	4063.3	1972.6	0.2	2600.3	870.8	0.4	4307.2
9	Na-SO_4	3082	0.2	4686.7	1671.2	0.2	2744.7	1035.6	559.3	5159.8	2000.6	0.7	6848.9
10	Na-Cl	3120	0.1	4642.9	1518.5	2.5	2668.0	940.0	563.9	5035.6	1874.5	0.3	5261.9
11	Na-HCO_3	3148	1.0	4454.1	1534.8	1.4	2515.8	961.5	525.9	4612.5	1843.8	0.6	6300.5
12	Na-CO_3	3134	0.1	4501.4	1693.7	0.6	3034.0	1013.3	822.0	5072.4	1995.5	2.1	5489.6
13	K-SO_4	3147	3.4	4574.4	1692.3	0.8	2737.2	1048.9	0.1	3371.6	945.0	1100.8	8598.4
14	K-Cl	3128	0.7	4356.2	1568.8	0.4	2623.7	948.6	0.2	3466.7	890.6	903.7	7977.3
15	K-HCO_3	3113	0.4	4816.2	1571.9	0.9	2646.9	928.6	0.4	3040.3	898.7	1133.6	8548.2
16	K-CO_3	3247	1.5	4802.8	1652.2	0.3	2990.7	1022.5	0.1	3491.2	970.6	978.9	8797.8
Concentrations (mM/L)													
1	Ca-SO_4	3102	28.5	148.4	80.8	0.0	104.7	38.4	0.0	104.5	39.1	0.0	106.2
2	Ca-Cl	3033	17.9	143.4	76.0	0.1	98.8	35.4	0.0	105.6	35.4	0.0	108.4
3	Ca-HCO_3	3131	22.2	149.2	76.4	0.0	95.2	35.4	0.0	111.4	36.2	0.0	120.3
4	Ca-CO_3	3031	26.7	139.6	81.5	0.0	98.5	39.5	0.0	115.5	38.7	0.0	118.6
5	Mg-SO_4	3120	0.1	103.4	39.1	23.5	166.9	81.5	0.0	119.9	37.7	0.0	117.0
6	Mg-Cl	3143	0.0	105.7	35.7	24.7	149.6	76.5	0.0	121.7	36.9	0.0	112.8
7	Mg-HCO_3	3161	0.1	97.4	35.0	15.0	162.3	75.3	0.1	113.4	36.0	0.0	113.1
8	Mg-CO_3	3160	0.1	101.8	39.0	27.6	167.2	81.2	0.0	113.1	37.9	0.0	110.2
9	Na-SO_4	3082	0.0	116.9	41.7	0.0	112.9	42.6	24.3	224.4	87.0	0.0	175.2
10	Na-Cl	3120	0.0	115.8	37.9	0.1	109.8	38.7	24.5	219.0	81.5	0.0	134.6
11	Na-HCO_3	3148	0.0	111.1	38.3	0.1	103.5	39.6	22.9	200.6	80.2	0.0	161.1
12	Na-CO_3	3134	0.0	112.3	42.3	0.0	124.8	41.7	35.8	220.6	86.8	0.1	140.4
13	K-SO_4	3147	0.1	114.1	42.2	0.0	112.6	43.2	0.0	146.7	41.1	28.2	219.9
14	K-Cl	3128	0.0	108.7	39.1	0.0	107.9	39.0	0.0	150.8	38.7	23.1	204.0
15	K-HCO_3	3113	0.0	120.2	39.2	0.0	108.9	38.2	0.0	132.2	39.1	29.0	218.6
16	K-CO_3	3247	0.0	119.8	41.2	0.0	123.1	42.1	0.0	151.9	42.2	25.0	225.0

calculation of ratios and log-ratio transformations instead of using crude compositions. This approach eliminates the compositional units and renders the compositions as simple numbers opening the space.

Error propagation through Monte Carlo simulations was reported for the first time by Verma (2012) to illustrate the inconvenience of using ternary diagrams for compositional data, and instead, a natural logarithm-transformed bivariate diagram was suggested. Afterward, Verma (2015) compared ternary diagrams against bivariate diagrams using Linear Discriminant and Canonical Analysis (LDCA), and also compared the performance of three types of log-ratio transformations –additive and centered proposed by Aitchison (1986); and isometric proposed by Egozcue et al. (2003)–, concluding that the bivariate diagrams were a better option to visualize and interpret compositional data because they do not show the analytical distortion error problems, moreover, the three log-ratios showed similar results. Later, Verma et al. (2016) and Verma and Armstrong-Altrin (2013) used a hybrid log-ratio (*hlr*) transformation, which differs from the other transformations (isometric, additive, and centered; defined in the section of *Methods | Hybrid log-ratio transformation*). However, the performance of the *hlr* is little known in the literature. Also, Verma et al. (2020) compared the isometric and *hlr* transformation, concluding that when both transformations were applied to the same database of major element compositions of igneous rocks, in conjunction with the LDCA, the same results were obtained. Similar conclusions were achieved earlier by Verma (2015) for the isometric, additive, and centered transformations. This consistency has shown that it does not matter which transformation is used for a multidimensional classification using major element concentrations.

Furthermore, according to Piper (1944), most natural water contains relatively few dissolved constituents, with cations (metals or bases) and anions (acid radicales) in chemical equilibrium with each other; the most abundant cation constituents are two *alkaline earth* Ca^{2+} and Mg^{2+}, and one *alkali* Na^{+}; K^{+} also occurs, but ordinarily is much less abundant than Na^{+}. Similarly, the most common anion-constituents are one *weak acid* HCO_3; and two *strong acids* $SO_4{}^{-2}$ and Cl^{-}. Other cations and anions occur in considerable quantities in highly concentrated waters. However, Piper (1944) suggested that all these less abundant constituents can be added with the major three constituents to which they are respectively related in chemical properties, e.g., Ca^{2+} with barium (Ba^{2+}) or strontium (Sr^{2+}); Na^{+} with K^{+}, cesium (Cs^{+}), rubidium (Rb^{+}), lithium (Li^{+}) or ammonium ($NH_4{}^{+}$); $HCO_3{}^{-}$ with $CO_3{}^{-}$ or Tetraborate ($B_4O_7{}^{-2}$); Cl^{-} with fluoride (F^{-}), nitrate ($NO_3{}^{-}$) or nitrite ($NO_2{}^{-}$). Thus, a water sample is treated on Hill-Piper ternary diagram substantially as though it contains only three cation-constituents and three anion-constituents. This explains why the cations *Na* and *K* are ordinarily combined as the anions HCO_3 and CO_3, and this approach reduces the number of possible hydrochemical facies that can be determined by ternary diagrams.

Therefore, since our scheme for water multidimensional classification is not based on ternary diagrams, we decided that it would be worthwhile to explore a new classification scheme without combining

K	SO_4			Cl			HCO_3			CO_3		
median	min	max	*median*	min	max	*median*	min	max	*median*	min	max	*median*
Concentrations (mg/L)												
1521.8	2084.9	14394.3	7818.5	1.3	3969.8	1368.0	0.2	7413.6	2331.6	0.2	6498.2	2339.7
1388.4	5.4	10457.1	3998.3	1109.0	8249.5	3076.5	2.0	8936.8	2558.2	0.4	7322.1	2584.3
1407.9	5.7	11913.4	4049.9	0.4	4516.2	1482.6	1955.8	13754.9	5259.8	2.3	7843.9	2583.9
1505.7	2.4	9547.0	3700.1	3.4	3736.2	1414.4	1.4	6686.5	2388.4	1497.8	9220.8	4839.8
1539.2	2581.2	14017.5	7782.7	0.8	4221.2	1382.9	0.7	6853.9	2417.9	1.6	6070.9	2321.1
1461.8	3.6	10762.5	4071.0	994.6	8291.8	3077.2	0.1	8321.0	2651.7	0.1	7402.5	2514.6
1376.6	2.9	10599.3	4029.5	0.1	5266.6	1458.4	1203.3	14570.4	5239.4	2.4	6477.8	2415.4
1485.3	1.9	9481.4	3684.8	0.7	4142.5	1356.7	2.1	6988.6	2385.2	1784.6	9436.9	4837.3
1653.8	1762.8	14189.0	7281.1	0.3	3969.2	1291.3	0.1	6857.3	2236.5	1.1	5585.3	2149.9
1521.2	6.6	10861.2	3709.5	809.7	7192.6	2844.2	4.6	7495.7	2373.2	0.9	7168.3	2283.9
1494.9	7.2	11315.7	3673.6	0.0	4755.5	1364.9	1427.5	12685.5	4903.7	0.5	6962.2	2351.3
1642.9	2.9	8941.6	3461.9	0.4	4068.3	1283.2	0.3	6649.0	2283.3	1754.0	8315.1	4557.4
3383.4	1908.3	13845.4	7296.9	1.1	3870.3	1305.1	3.0	6621.8	2207.5	1.6	6095.1	2174.7
3122.5	0.9	10535.3	3676.8	955.9	8294.0	2898.1	9.4	8006.4	2358.8	2.2	6846.6	2293.3
3157.3	1.6	10177.5	3728.4	2.4	5355.9	1476.7	1465.7	12264.1	4956.2	0.2	6589.5	2297.3
3403.3	1.0	9354.7	3482.1	1.0	4083.3	1286.1	0.2	6798.8	2180.1	1550.4	9325.8	4561.3
Concentrations (mM/L)												
38.9	21.7	149.9	81.4	0.0	112.0	38.6	0.0	121.5	38.2	0.0	108.3	39.0
35.5	0.1	108.9	41.6	31.3	232.7	86.8	0.0	146.5	41.9	0.0	122.0	43.1
36.0	0.1	124.0	42.2	0.0	127.4	41.8	32.1	225.4	86.2	0.0	130.7	43.1
38.5	0.0	99.4	38.5	0.1	105.4	39.9	0.0	109.6	39.1	25.0	153.7	80.7
39.4	26.9	145.9	81.0	0.0	119.1	39.0	0.0	112.3	39.6	0.0	101.2	38.7
37.4	0.0	112.0	42.4	28.1	233.9	86.8	0.0	136.4	43.5	0.0	123.4	41.9
35.2	0.0	110.3	41.9	0.0	148.6	41.1	19.7	238.8	85.9	0.0	107.9	40.3
38.0	0.0	98.7	38.4	0.0	116.9	38.3	0.0	114.5	39.1	29.7	157.3	80.6
42.3	18.4	147.7	75.8	0.0	112.0	36.4	0.0	112.4	36.7	0.0	93.1	35.8
38.9	0.1	113.1	38.6	22.8	202.9	80.2	0.1	122.8	38.9	0.0	119.5	38.1
38.2	0.1	117.8	38.2	0.0	134.1	38.5	23.4	207.9	80.4	0.0	116.0	39.2
42.0	0.0	93.1	36.0	0.0	114.8	36.2	0.0	109.0	37.4	29.2	138.6	75.9
86.5	19.9	144.1	76.0	0.0	109.2	36.8	0.0	108.5	36.2	0.0	101.6	36.2
79.9	0.0	109.7	38.3	27.0	234.0	81.8	0.2	131.2	38.7	0.0	114.1	38.2
80.8	0.0	106.0	38.8	0.1	151.1	41.7	24.0	201.0	81.2	0.0	109.8	38.3
87.0	0.0	97.4	36.3	0.0	115.2	36.3	0.0	111.4	35.7	25.8	155.4	76.0

any of these 8 ionic species.

The *7-hlr* model recently presented by Verma et al. (2021; denoted as WClassHLR), was developed using LDCA and the hybrid log-ratio (*hlr*) transformations of the milliMoles (mM/L) concentrations of 8 major ions, without combining any of them. For WClassHLR model, a representative database from Monte Carlo simulations (Law and Kelton, 2000; Verma and Quiroz-Ruiz, 2006) was generated (defined in section *Methods | Database simulation procedure*). This initial classification of each simulated sample was achieved from the greater molar concentration concept of each cation and anion, called the Greater Molar Concentration (GMC) criteria. The 16 classes were all set to similar sizes (minimum size = 3021 samples; maximum size = 3247 samples; see Table 1). Before undertaking the LDCA, classes were free censored from multivariate discordant outliers. Each class was processed through DOMuDaF (*Discordant Outlier from Multivariate Data through F-test of W;* Verma et al., 2016) program for the detection and separation of multivariate discordant outliers, because LDCA require that the individual classes be multi-normally distributed in terms of the features (*7-hlr* variables). DOMuDaF program detects multivariate discordant outliers through a transformation of the Wilks statistic W (Wilks, 1963) to the well-known F-test.

A total of 46,292 multivariate discordant outlier-free samples were obtained from DOMuDaF for training the LDCA. The statistical characteristics of outlier-free multivariate data are summarized in Table S1. Finally, the WClassHLR model was obtained using the 16 classes initially assigned, and from the probability concept, it was possible to identify hybrid water types, along with the basic types of water. Thus, WClassHLR model under the "*hybrid*" option (defined in section *Methods | Hybrid log-ratio transformation*) can classify as many as 256 different water classes.

Machine Learning (ML) techniques have been used in many aspects of hydrological science (e.g., Acharya et al., 2019), but mainly, to classify or estimate water quality (Muhammad et al., 2015; Wang et al., 2017; Gakii and Jepkoech, 2019; Abba et al., 2020; Gaya et al., 2020; Melesse et al., 2020; Lu and Ma, 2020; Liang et al., 2020; Banadkooki et al., 2020; Tung and Yaseen, 2020; Zhou, 2020). Bayram and Gultekin (2010) proposed an Artificial Neuron Network (ANN) model to classify water samples from Simav geothermal area in Western Turkey, this model classify the samples into 4 different classes (Eynal, Çitgöl, geothermal water, and cold water), the model indeed have good results, however, this classification is not a general nomenclature that could be used as hydrochemical facies. Also, ML techniques such as Support Vector Machines (SVM) and Random Forest (RF) have been used for flood susceptibility mapping (Nachappa et al., 2020); Ni et al. (2020) compared the following three ML models: extreme gradient boosting model with Gaussian mixture model (XGB), standalone XGBoost and SVM to provide streamflow forecasting; Wu et al. (2019) conducted a comparison of eight ML models (i.e., ANN; RF, Gradient Boosting Decision Tree – GBDT; XGBoost, Multivariate Adaptive Regression Spline – MARS; SVM; and, Kernel-based Nonlinear Extension of Arps Decline – KNEA model) for the estimation of monthly mean daily reference evapotranspiration, which is important in hydrology research, irrigation

scheduling design and water resources management; and, SVM and XGBoost comparison for streamflow forecasting (Yu et al., 2020). In general, all the ML techniques show good results over other techniques employed, particularly the SVM and XGB models.

In this work, our main aims are summarized as follows:

(i) to generate a training database ($n = 50,000$ samples) and another independent validation database ($n = 8,000$ samples) through Monte Carlo simulations (*Methods|Databases simulation procedure*); available in the public server (*https://github.com/usju/water-databases*);

(ii) to develop four new nonparametric models (WClassCB, WClassVL, WClassVP, WClassVR) for water nomenclature through the molar concentrations of 8 ions ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4^{-2}, Cl^{-}, HCO_3^{-}$, and CO_3^{2-}), from two machine learning techniques based on Gradient Boosting and Support Vector Machines (*Methods | Classification methods*), and *hlr* transformation (*Methods / Hybrid* log-*ratio transformation*);

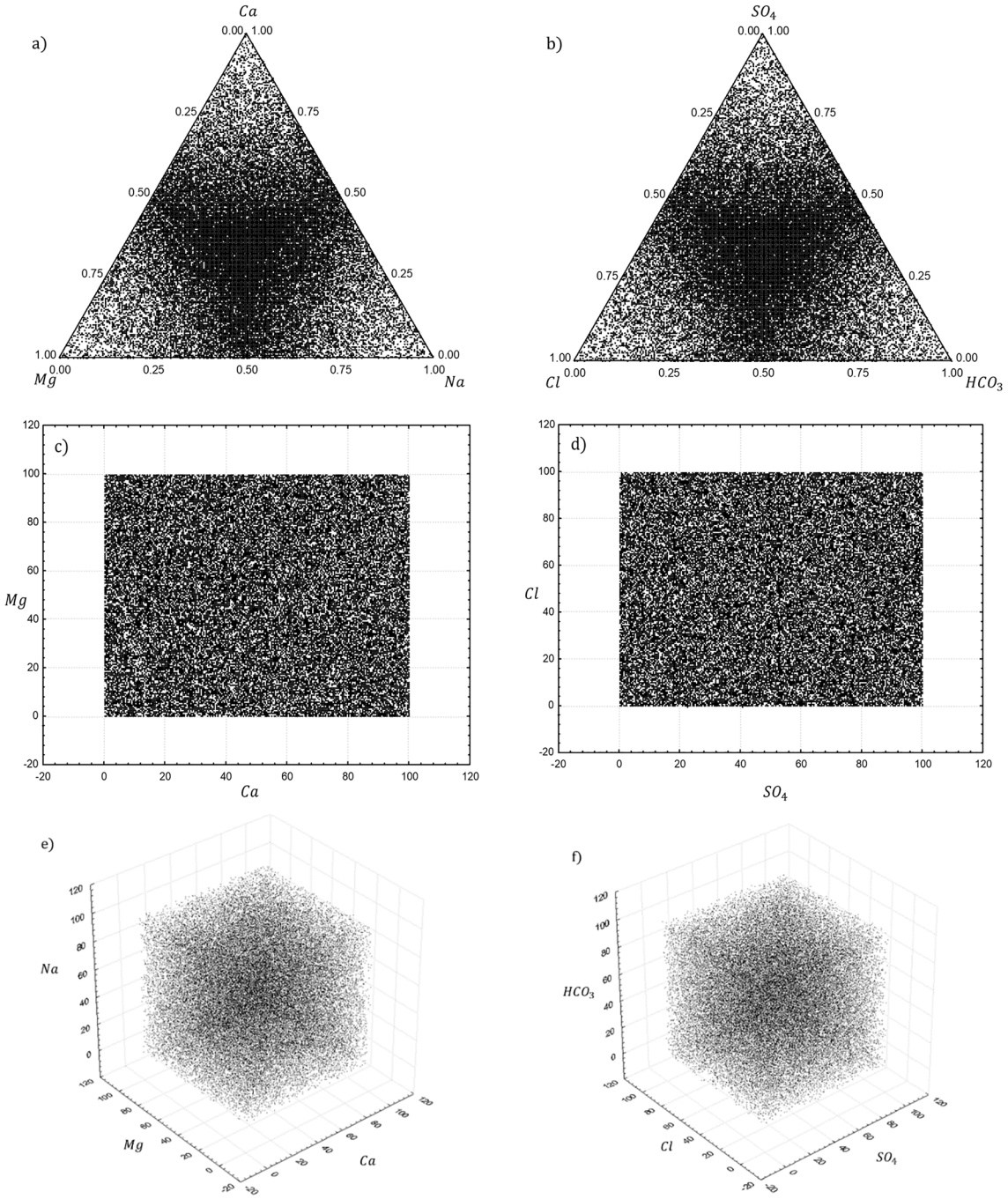


Fig. 1. Schematic representation of the training database (50,000 simulated samples). (a) and (b) Ternary diagrams cations (Ca^{2+}, Mg^{2+}, Na^{+}) and anions ($SO_4^{-2}, Cl^{-}, HCO_3^{-}$). (c) and (d) Two-dimensional structure plot of the simulated cations (Ca^{2+}, Mg^{2+}) and anions (SO_4^{-2}, Cl^{-}) data. The x- and y-axes are U_i and U_{i+1}, respectively, where $U_1, U_2, \ldots U_8$ is the sequence of random numbers generated for each element (Ca^{2+}, Mg^{2+}, $(Na^{+} + K^{+})$, SO_4^{-2}, Cl^{-}, $(HCO_3^{-} + CO_3^{2-})$). (e) and (f) Three-dimensional plot of the simulated cations and anions data. The x-, y-, and z-axes are the sequences of random numbers generated for each element. Note that the simulated data fill the space.

(iii) to develop the new models on Python following standard machine learning practices (*Methods | Training and validation of models*) such as hyperparameter optimization and training and evaluation metrics, which performance up to 99% was achieved; this Python program is available in the public server (*https://github.com/usju/water-training-test-models*);
(iv) to develop a post-processing program to identify hybrid water types in the four machine learning models, from probabilities generated for each of the output basic classes;
(v) to compare through machine learning metrics (section of *Methods | Models evaluation metrics*) the performance of these four nonparametric models with WClassHLR model (Verma et al., 2021) to find which technique performs better to determine hydrochemical facies in water samples;
(vi) to illustrate the usefulness of all the models by four applications on real groundwater samples from India and Nigeria; and,
(vii) to develop a program on Python for the effective use of the new four machine learning models, which is available in the public server (https://github.com/usju/water-classification-ML).

2. Methods

2.1. Training database

The training and the independent validation databases consist of 50,000 samples and 8,000 samples on molar concentrations (mM/L) with 4 cations ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$) and 4 anions ($SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$, $CO_3{}^{2-}$), respectively. The generated databases are described in the following section.

2.2. Databases simulation procedure

The Monte Carlo procedure suggested by Verma and Quiroz-Ruiz (2006) shows how to assess the randomness of uniformly distributed numbers (i.e., IID U(0,1)) and to generate normally distributed samples (i.e., IID N(0,1)) using different seeds to guaranteed the randomness of such values via numerical simulation.

Following these suggestions, our Monte Carlo sampling simulation can be summarizing as follows:

- Generate N individual uniformly distributed numbers IID U(0,1) for each element ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$, and$CO_3{}^{2-}$) using the Mersenne Twister pseudo-random number generator algorithm (Matsumoto and Nishimura, 1998); where N took values of 50,000 samples for the training dataset and 8,000 samples for the independent validation dataset.
- Compute a scalar value of 100 to cover the representability of the ternary diagrams (Fig. S2a). The representativeness of the *initial* database is illustrated by plotting the simulated in a cation ternary diagram (Ca^{2+}, Mg^{2+}, Na^{+}; Fig. 1a) and an anion ternary diagram ($SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$; Fig. 1b).
- Use the ICB equation (ionic charge-balance equation; Nicholson, 1933) as $ICB = \frac{\left|\sum cations + \sum anions\right|}{\left|\sum cations - \sum anions\right|}$, where $\sum cations$(or *anions*) are given on milliequivalent per liter (mEq/L) units. Also, a value of 0.00005% was established as the maximum unbalance. However, as the samples were randomly generated, we proceed to apply the following unbalance factor as $Factor = \frac{\left|\sum cations\right|}{\left|\sum anions\right|}$to assess the charge balance in our simulated database. This unbalance factor is applied to a pseudo-random increment from 0 to 10% for each element. The ionic charge-balance validation procedure is shown in Fig. S1, which enabled the selection of samples with $ICB < \pm 0.00005\%$ for a better dataset setting. A histogram of the database after the ICB procedure is presented in Fig. S2b.
- Determine the cross-combination of majoritarian cation and anions of the simulated dataset (Table 1).

The 16 classes achieved consist of cross-combinations of 4 cations and 4 anions primary classes, which were assigned from the highest anion and cation molar concentrations (Table 1; listed under the column "Water class"). The statistical characteristics of the multivariate database (number of samples n, median x , maximum and minimum values) are summarized in Table 1. This database has 16 balanced or equal-sized classes (minimum size = 3021 samples; maximum size = 3247 samples; see Table 1).

2.3. Hybrid log-ratio transformation

We decided to use *hlr* transformation based on the previous comparisons reported by Verma (2015, 2020) where it concluded that is not relevant which transformation is chosen for multidimensional classification using major element concentrations. The *hlr* transformation of the molar concentrations (mM/L) of 8 elements ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4{}^{-2}, Cl^{-}, HCO_3{}^{-}$, and $CO_3{}^{2-}$) were calculated using the following general equation:

$$hlr_{(i+1)} = ln\left(\frac{g(x_i, \cdots, x_n)^{1/n}}{x_{(i+1)}}\right), i = 1, 2, \cdots, (n-1)$$

where $g(\cdot)$ denotes the geometric mean, x_i represent the concentration of each element in the same order $(Ca \times Mg \times Na \times K \times SO_4 \times Cl \times HCO_3 \times CO_3)^{(\frac{1}{8})}$, $x_{(i+1)}$ denotes one ion at a time from second (Mg^{2+}) to last ($CO_3{}^{2-}$), and n subscript is the total number of major elements ($n = 8$). For more information about *hlr* transformation see *A1. Additional details on hybrid* log*-ratio transformation* section in Appendix A.

Thus, the seven *hlr* variables (hlr_2 to hlr_8; Eqs. S1 to S7) were used as features for training the multidimensional classification models. A histogram of the *hlr* transformations is presented in Fig. S2c.

According to Aitchison (1986) and Aitchison and Egozcue (2005), in any discipline, when a problem is compositional, we are recognizing that the sizes of our specimens are irrelevant. For example, a geologist talking about the composition of an object, such as the major oxide composition of a rock, is declaring as a dimensionless problem. There is no concern about whether the rock specimen weighs one gm or one lb. Compositions are concerned with relative values and ratios of components.

The geometric mean was used in the *clr* and *ilr* transformations, proposed by Aitchison (1986) and Egozcue et al. (2003), respectively, for the treatment of compositional data. Using the geometric mean on the components of a sample has the advantage of treating the parts symmetrically and is a reasonable way to measure the dependence between the composition parts.

In this work, as the water components are charge-balanced concentrations of 4 cations (*Ca*, *Mg*, *Na*, and *K*) and 4 anions (SO_4, *Cl*, HCO_3, and CO_3), the geometric mean was applied, considering that these components are chemically related. The geometric mean use relies on the fact that it considers the cumulative and compound effects (Spizman and Weinstein, 2008). Moreover, the geometric mean is closely related to the log-transformation in statistics, which is widely used for skewed data (Feng et al., 2013).

Furthermore, the geometric mean has been used in a very broad range of natural and social science disciplines like environmental monitoring, scientometrics, nuclear medicine, infometrics, economics, finance, ecology, surface and groundwater hydrology, geoscience, and geomechanics (Vogel, 2020).

2.4. Classification methods

2.4.1. WClassHLR model generated by LDCA

The *7-hlr* model proposed by Verma et al. (2021; denoted as WClassHLR) consists of an assembly of classifiers created by LDCA, which is a supervised classification technique that consists of finding a one-dimensional linear function that discriminates between the classes by the measure of maximum separation and serves as a basis for classifying samples of unknown classes. This methodology maximizes the distance between the classes and minimizes the distance between the samples for each class.

The main idea was to achieve a classification model for cations and another for anions. However, for 4 groups the LDCA would provide 3 discriminant functions, requiring a three-dimensional diagram to visualize them. Therefore, 3 groups were evaluated at the same time, which required making 4 2D-models for cations (Ca, Mg, Na, and K) and 4 2D-models for anions (SO_4, Cl, HCO_3, and CO_3). The 16 water types could thus be classified by cross combinations. Each of these models has the advantage of being visualized in 2D diagrams using two discriminant functions, which can be represented in a general form in Eq. S8. See *A2. Additional details on linear discriminant and canonical analysis (LDCA)* section in Appendix A.

The WClassHLR model consists of 8 sub-models (4 for cations and 4 for anions), containing 16 DFs, 128 coefficients (Table S2), and 48 centroids (Table S3). For example, the $DF1h_{Ca\_Mg\_Na}$ and $DF2h_{Ca\_Mg\_Na}$ required for the classification of $Ca-Mg-Na$ can be calculated from Eqs. S9 and S10. For the other cation sub-models of three at a time ($Ca-Mg-K$, $Ca-Na-K$, and $Mg-Na-K$), the discriminant functions can be similarly calculated (see coefficients in Table S2). The classification of anions (SO_4, Cl, HCO_3, and CO_3) was similarly achieved from 4 sub-models of the "three at a time" type. The final water nomenclature is achieved from probability calculations for the competing fields in all "three at a time" sub-models and their comparison.

2.4.2. Gradient boosting (Catboost) technique

CatBoost (i. e. categorial boosting) works by sequentially adding weak predictors to make a strong classifier, each predictor of the assemble is trained with the residual errors of its predecessor (Géron, 2019). CatBoost is a process of constructing an ensemble model (strong classifier) by performing gradient descent in a functional space (Prokhorenkova et al., 2018). Categorical Boosting is an open-source library that implements a modification of the standard algorithm, it uses binary decision trees as weak predictors that make up the assemble which divide the feature space into disjoint regions according to the values of some splitting attributes (Prokhorenkova et al., 2018). CatBoost is available as a Python library and uses scikit-learn (Pedregosa et al., 2011) framework.

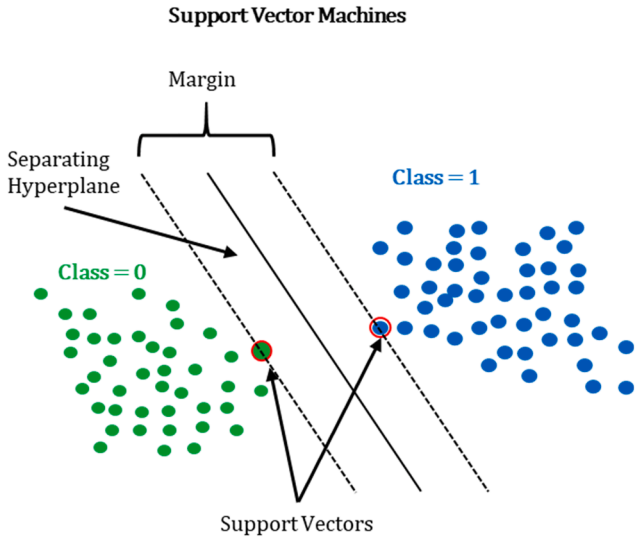


Fig. 2. A diagram to illustrate SVMs concepts visually.

2.4.3. Support Vector Machines (SVM) technique

SVM are supervised machine learning algorithms that are used for classification and regression purposes. When it comes to classification, samples of the form $\{(X_i, y_i)\}, i = 1 \cdots n$ can be classified by a *hyperplane* which optimizes the distances between itself and the closest vectors of each class. These vectors are called *support vectors* and are examples from the training set, the distance between the vectors and the classes is called *margin* (Géron, 2019**).** Fig. 2 diagram illustrates these concepts visually.

In SVM, our main objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. The hyperplane can be expressed as $f(x) = \omega\phi(x) + b$; where x is the input data; $\phi(x)$ represents a kernel function that projects x into the high-dimensional feature space; ω and b are the coefficients estimated by the SVM procedure (Huang et al., 2019).

SVMs use the so-called "*kernel-trick*" that allows using a kernel function denoted as $K(x, z)$, which makes it possible to map the dot product of vectors (x, z) in high dimensional space (Géron, 2019). When samples are projected in a higher dimensional space by the kernel function, the additional dimensions afford a greater opportunity to find a hyperplane that separates classes (Wadkar et al., 2019). SVMs are recommended when it is needed to perform classification on complex small or medium datasets, as in our dataset of 50,000 samples (Géron, 2019).

In this work, we use and evaluate SVMs with three different kernel functions: linear (L), polynomial (P), and radial basis function (RBF). SVMs are used through scikit-learn's SVC module (Pedregosa et al., 2011).

2.4.4. New water classification models (WClassCB, WClassVL, WClassVP, WClassVR)

To carry out the water classification models through CatBoost and SMV techniques, a general computational methodology was developed (Fig. 3). A total of 50,000 simulated analyses of ionic charge-balanced concentrations of 8 major elements (Ca, Mg, Na, K, SO_4, Cl, HCO_3, CO_3; mMol/L) were used for training models. The *hlr* transformations of the milliMoles (mM/L) concentrations of the 8 ions were calculated. After, two ML techniques based on Gradient Boosting and Support Vector Machines were applied.

2.4.4.1. Gradient Boosting (CatBoost) implementation. CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. We applied this algorithm from the catboost.CatBoostClassifier Python module. The multiclass support is handled according to a "one-vs-rest" scheme. Thus, to classify 4 classes of cations (Ca, Mg, Na, and K), 4 binary classifiers are constructed as follow: (i) Ca *vs* $(Mg-Na-K)$, (ii) Mg *vs* $(Ca-Na-K)$, (iii) Na *vs* $(Ca-Mg-K)$ and, (iv) K *vs* $(Ca-Mg-Na)$. The classification of anions (SO_4, Cl, HCO_3, and CO_3) was similarly completed from 4 sub-models of the "one-to-rest" type (Fig. 3).

The decision trees were used for these classifiers, each of the trees corresponds to a partition of the feature (input variables: hlr_2 to hlr_8) space and the output value. A decision rule is used for each level of the tree acting as the "splitting criterion". Each decision rule can be conceptualized as a pair $r = (k, v)$, that contains a feature index $r = 1, \cdots, m$ and a threshold value $v \in R$. Thus, a set of feature vectors X can be split into two disjoint subsets of X^L and X^R. (Kang et al., 2019) Then, for each $x = (x^1, \cdots, x^n) \in X$ we have that $x \in X^L$ *if* $(x^k \leq v)$ or $x \in X^R$ *if* $(x^k > v)$. Then, applying the decision rule to s disjoint sets $X_{(1)}, \cdots, X_{(s)} \in R^m$, the total number of disjoint sets is $2s$; $X^L_{(1)}, X^R_{(1)}, \cdots, X^L_{(s)}, X^R_{(s)}$.

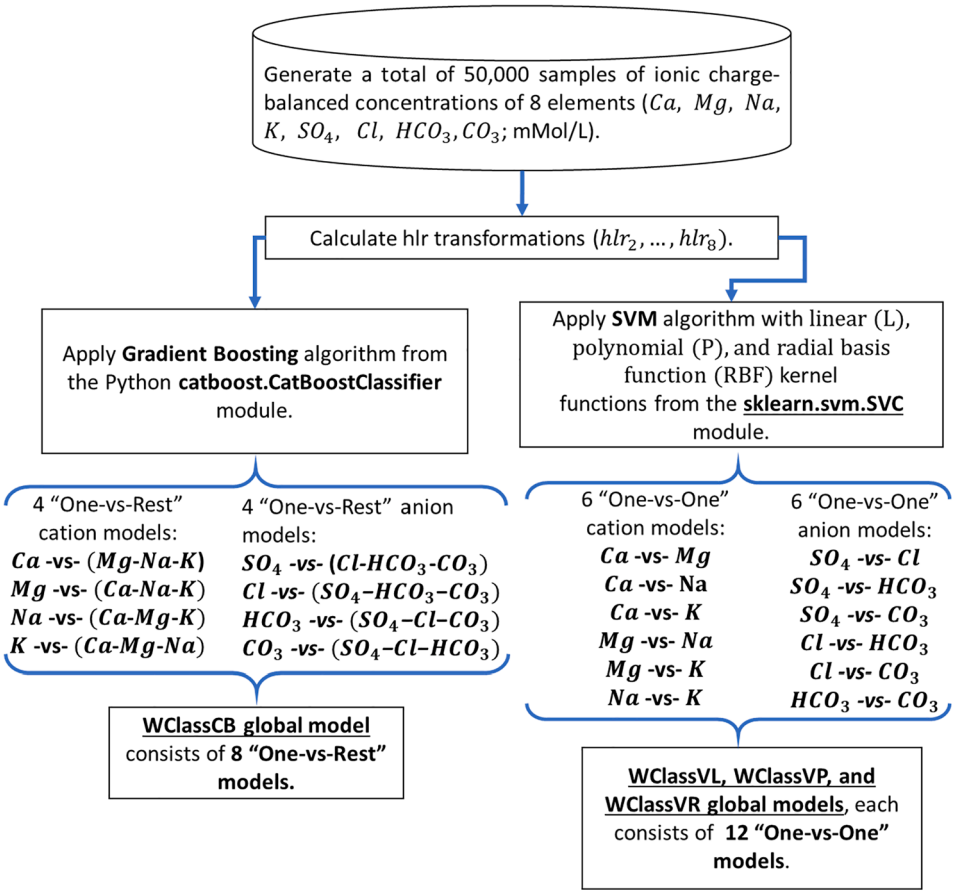


Fig. 3. Schematic diagram of the development of new water classification models (WClassCB, WClassVL, WClassVP, and WClassVR).

(Kang et al., 2019).

Once the models were trained, we used the "*predict*" and "*predict_proba*" functions of the *sklearn.svm.SVC* library to perform classification on unknown water samples. The "predict" function applies the model to the given dataset, and the "*predict_proba*" function generates the probability that the object belongs to the given classes.

The CatBoost based model, named WClassCB, can be used in the executable program available at the public server (https://github.com/usju/water-classification-ML).

2.4.4.2. Support Vector Machines (SVM) implementation. For SVM, the *sklearn.svm.SVC* library (Pedregosa et al., 2011) was used. The multiclass support is handled according to a "one-vs-one" scheme. Thus, to classify 4 classes of cations (*Ca*, *Mg*, *Na*, and *K*), 6 binary classifiers ($n_{classes}$*($n_{classes}-1$)/2) as follow: (i) *Ca vs Mg*, (ii) *Ca vs Na*, (iii) *Ca vs K*, (iv) *Mg vs Na*, (v) *Mg vs K*, and, (vi) *Na vs K*. The classification of anions (SO_4, *Cl*, HCO_3, and CO_3) was similarly achieved. The final water nomenclature is achieved from probabilities for the competing fields in all sub-models. We applied linear (WClassVL model), polynomial (WClassVP model), and radial basis function (WClassVR model) kernel functions. For details on the precise mathematical formulation of the kernels, see the corresponding documentation provided by Pedregosa et al. (2011). In the linear kernel, the hyperplane or decision function for discriminates of $c1$ and $c2$ classes can be constructed as $hp_{c1\_c2} = (w_1 \times hlr_2) + (w_2 \times hlr_3) + (w_3 \times hlr_4) + (w_4 \times hlr_5) + (w_5 \times hlr_6) + (w_6 \times hlr_7) + (w_7 \times hlr_8) + b$; where hp_{c1_c2} is the hyperplane equation for discriminates of $c1$and $c2$ classes, hlr_2-hlr_8 are input variables, w_1 to w_7 and b are the coefficients or adjustable parameters of the decision function.

For an unknown sample, the hyperplane equation generates a positive or negative value for a sample classified as $c1$ or $c2$ class, respectively. Once the SVM models have been trained, their coefficients and intercept values can be extracted, through the functions "coef_" and "intercept_", respectively, from *sklearn.svm.SVC* library. These coefficients are listed in Table S3 and are used to construct the hyperplane functions (DF). The *WClassVL* model consists of 12 "one-to-one" submodels (6 for cations and 6 for anions) and 96 coefficients (Table S3). The *WClassVL* model final water nomenclature is achieved from probability calculations for the competing fields in all sub-models.

When the kernel is not linear, the *sklearn.svm.SVC* library provides the coefficients ("*dual_coef_*") and support vectors ("*support_vectors_*") that are required to construct the hyperplane or decision function. To keep this work short, we only present the SVL-L model coefficients. However, we used the "predict" function from *sklearn.svm.SVC* library to perform classification on unknown samples. This function predicts the results using the trained model. The proposed SVM-based models WClassVL (linear kernel), WClassVP (polynomial kernel), and WCLassVR (radial basis function kernel) can be used through the *WaterClaSys_ML* executable program available at the public server (https://github.com/usju/water-classification-ML).

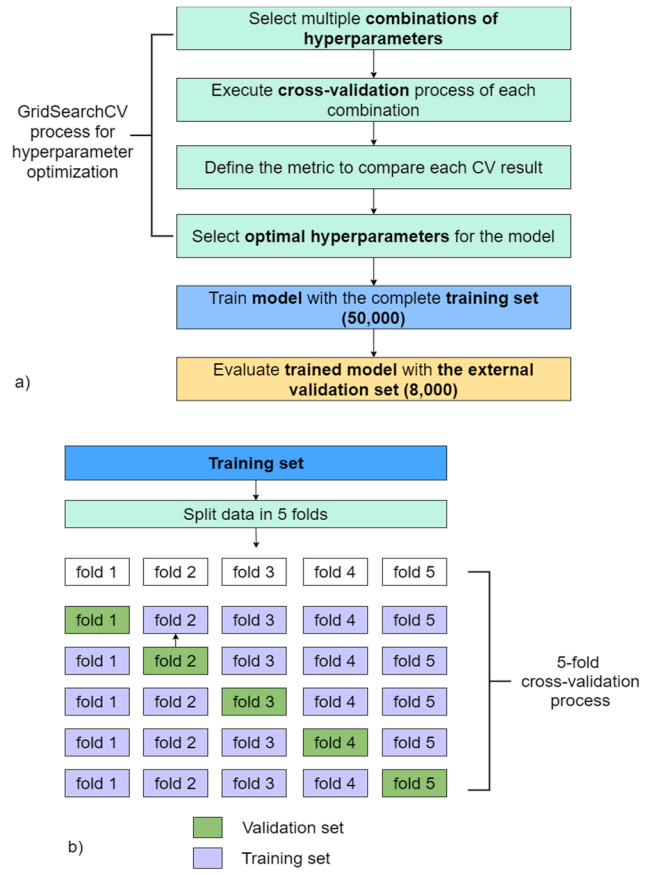


Fig. 4. Schematic diagram of the following processes: a) GridSearchCV process for hyperparameter optimization, training, and validation of machine learning models; b) 5-fold cross-validation process (Pedregosa et al., 2011).

2.5. Training and validation of models

2.5.1. Hyperparameter tuning

Since ML algorithms have many hyper-parameters to tune, this process has been automated. Hyperparameter tuning was done through *GridSearchCV* module from scikit-learn (Pedregosa et al., 2011), which for given values, exhaustively considers all parameter combinations. GridSearchCV uses a 5-fold cross-validation procedure, of which 80% (four subsets) is used to train and the remaining 20% (one fold) is used to validate the model. The process makes combinations of the folds where all subsets are used as a validation set at least once. The cross-validation was executed per combination of given hyperparameters, the best hyperparameters were selected by looking at the average accuracy of each cross-validation (Heung et al., 2016). Once a set of hyper-parameters is selected, it is used to train the model with the complete training dataset. The general procedure for the training and evaluation of the ML models (WClassCB, WClassVL, WClassVP, WClassVR) and the cross-validation procedure are shown schematically in Fig. 4.

2.5.2. Evaluation metrics

To compare the performance of the models, we use the following four metrics: (i) *classification accuracy* (Eq. S11); (ii) *classification precision* (Eq. S12); (iii) *recall* (Eq. S13); and the number of false positives (Eq. S14). These metrics are calculated for all possible classification thresholds (Géron, 2019). An ideal model with high precision and high recall will return many results, with all results classified correctly, and would have an *AUC* close to 1 that tells us that the classifier is not making random decisions (Géron, 2019). To achieve the prediction of multiple classes by the classifiers, we compute the *macro-average precision* (Eq. S15), *macro-average recall* (Eq. S16), and *macro-average AUC* (Eq. S17) using the "one vs rest" approach. See *A3. Additional details evaluation metrics* section in Appendix A.

2.6. Application of the models

2.6.1. Preprocessing of real samples

Before applying the models we used L2 normalization –defined as $X_{new} = \frac{X}{\sqrt{X_1^2+X_2^2+X_3^2\cdots+X_n^2}}$, where X is a sample vector and X_{new} is the scaled vector– to scale field water samples before the application of the classification models. L2 norm is a standard method to compute the length of a vector in Euclidean space. L2 norm of a vector is defined as the square root of the sum of the squares of the values in each dimension. The features or input variables (hlr_2 to hlr_8) of field water samples (non-simulated water samples) could have different scales and contain some extreme values, which could degrade the predictive performance of the machine learning algorithms (Pedregosa et al., 2011).

2.6.2. WaterClaSys_ML program

It should be clear that it is difficult to use and evaluate these new classification schemes without a suitable computer program. Therefore, a computer program *WaterClaSys_ML* was written in Python, which is available on the public server (https://github.com/usju/water-classification-ML) and requires an input file and provides an output file. For the proper use of this program, the user must provide the input samples in mg/L concentrations of 4 cations (*Ca*, *Mg*, *Na*, and *K*) and 4 anions (SO_4, *Cl*, HCO_3, and CO_3).

2.6.3. Hybrid water types

Since these multidimensional models are based on the probability concept, it is possible to find a final combined basic and hybrid types, theoretically up to a total of 256 different water classes. Let us suppose that P_i is the probability for a cation or anion, where the subscript i varies from 1 to 4 for 4 cations or anions. Of these 4 probabilities, let $P_m$be the highest probability and P_n be the second-highest probability. The conditions that define if the nomenclature is basic (just a water type) are as follows: if $(P_m \geq 0.5)$and $((P_m - P_n) \geq 0.25)$ and $(P_n \leq 0.25)$; otherwise, a hybrid (two water types, the highest probability class name followed by the next highest name) nomenclature is assigned. This condition allows us to determine up to 256 hybrid water types (16*16), given that for the 4 cations and anions separately, 4 basic and 12 hybrid classes can be achieved (Verma et al., 2021).

3. Results and discussion

In this work, we used the mMol/L units as was suggested by Verma et al. (2021), because the initial assignment for 16 classes was achieved from the greater molar concentration concept of each cation and anion. Also, the number of atoms of the chemical elements proved to be a feasible way to determine water types.

3.1. Training of the new multidimensional models

The 7 variables (hlr_2 to hlr_8) calculated from the molar concentrations (mM/L) of 4 cations (Ca^{2+},Mg^{2+},Na^{+},K^{+}) and 4 anions ($SO_4{}^{-2}$,Cl^{-}, $HCO_3{}^{-}$, $CO_3{}^{2-}$) of the training set were used to train the WClassCB, WClassVL, WClassVP, and WClassVR models. The training database has 12297, 12584, 12484, 12,635 samples respectively for each cation class and 12451, 12424, 12553, 12,572 samples for each anion class, both parts of the classification were determined by the greater cation and

Table 2
Selected hyperparameters for classification models and their accuracy values obtained from the cross-validation process.

		Cation classification			Anion Classification		
		Selected hyperparameters	*cross-validation accuracy* ($\bar{x} \pm s$)		Selected hyperparameters	*cross-validation accuracy* ($\bar{x} \pm s$)	
No	Model		*training set*	*validation set*		*training set*	*validation set*
1	WClassCB	depth = 4, learning rate = 0.1, iterations = 2000	99.16 ± 0.02	96.94 ± 0.30	depth = 4, learning rate = 0.1, iterations = 2000	98.98 ± 0.04	98.60 ± 0.08
2	WClassVL	C = 10	99.64 ± 0.03	99.61 ± 0.02	C = 10	99.75 ± 0.02	99.70 ± 0.08
3	WClassVP	C = 10, degree = 3	98.15 ± 0.05	97.83 ± 0.07	C = 10, degree = 3	99.27 ± 0.08	98.95 ± 0.04
4	WClassVR	C = 10, gamma = 0.1	99.87 ± 0.04	97.88 ± 0.11	C = 10, gamma = 0.01	99.13 ± 0.03	98.98 ± 0.07

Table 3
Synthesis of percent success values obtained from the training of the selected models (WClassHLR; WClassCB; WClassVL; WClassVP; and, WClassVR).

Model	Number and % (classification accuracy) of correctly classified samples								Total number and % of classified samples			
									correctly		incorrectly	
	no.	classification accuracy	no.	classification accuracy	no.	classification accuracy	no.	classification accuracy	no.	total classification accuracy	no.	%
	Cation classes											
	Ca		Mg		Na		K		Cation classification (Ca, Mg, Na, K)			
WClassHLR	10,325	90.83	10,549	91.06	10,819	93.60	11,054	93.84	42,747	92.34	3545	7.66
WClassCB	12,235	99.50	12,557	99.79	12,450	99.73	12,612	99.82	49,854	99.71	146	0.29
WClassVL	12,252	99.63	12,566	99.86	12,433	99.59	12,583	99.59	49,834	99.67	166	0.33
WClassVP	12,098	98.38	12,412	98.63	12,232	97.98	12,430	98.38	49,172	98.34	828	1.66
WClassVR	12,111	98.49	12,476	99.14	12,376	99.13	12,517	99.07	49,480	98.96	520	1.04
	Anion classes											
	SO_4		Cl		HCO_3		CO_3		Anion classification (SO_4, Cl, HCO_3, CO_3)			
WClassHLR	10,688	92.75	10,687	92.80	10,938	93.75	10,756	92.84	43,069	93.04	3223	6.96
WClassCB	12,447	99.97	12,419	99.96	12,549	99.97	12,566	99.95	49,981	99.96	19	0.04
WClassVL	12,425	99.79	12,380	99.65	12,524	99.77	12,557	99.88	49,886	99.77	114	0.23
WClassVP	12,372	99.37	12,315	99.12	12,484	99.45	12,521	99.59	49,692	99.38	308	0.62
WClassVR	12,372	99.37	12,308	99.07	12,440	99.10	12,510	99.51	49,630	99.26	370	0.74

Table 4
Synthesis of m*acro-average precision* and *recall* values obtained from the training and external validation of the selected classification models (WClassHLR; WClassCB; WClassVL; WClassVP; and, WClassVR).

Model	Cation classification (Ca, Mg, Na, K)				Anion classification (SO_4, Cl, HCO_3, CO_3)			
	Training set		External validation set		Training set		External validation set	
	macro-average precision	*macro-average recall*	*macro-average precision*	*macro-average recall*	*macro-average precision*	*macro-average recall*	*macro-average precision*	*macro-average recall*
WClassHLR	92.4	92.33	90.81	90.76	93.05	93.04	92.05	92.05
WClassCB	99.71	99.71	96.13	96.15	99.96	99.96	98.69	98.69
WClassVL	99.67	99.67	99.51	99.51	99.77	99.77	99.84	99.84
WClassVP	98.35	98.34	97.85	97.81	99.39	99.38	99.02	99.03
WClassVR	98.96	98.96	97.82	97.81	99.26	99.26	98.9	98.9

anion. On the other hand, the outlier-free database used for training the WClassHLR model has 11368, 11585, 11559, 11,780 samples for each cation class, and 11523, 11516, 11667, 11,586 samples for the anion classes. This criterion is called GMC, which stands for "Greater molar concentration model" to designate initial water types (Verma et al., 2021).

From scikit-learn.GridSearchCV Python module (Pedregosa et al., 2011), we selected a set of hyperparameter for each model by searching the equilibrium point between good regularization and accuracy. The selected hyperparameters along with their accuracy values are listed in Table 2. The depth equal to 4 (Table 2 and Table S2) for the WClassCB model was selected, because very large depths could prone to overfit (Huang et al., 2019). The "C" regularization parameter equal to 10 for the WClassVL, WClassVP, and WClassVR models was selected, since small values achieved a better generalization (Géron, 2019). The tested parameter settings are listed in Tables S2 and S3. The training of the models was executed in an intel core i7-9750H CPU.

We present the results of the training process which shows the accuracy of each class in the selected classification methods. The WClassHLR model provided a success value of 90.83% for Ca (10,325 samples correctly classified), 91.06% for Mg (10,549 samples correctly classified), 93.60% for Na (10,819 samples correctly classified), and 93.84% for K (11,054 samples correctly classified. The overall success of the WClassHLR model is 92.34% that corresponds to 42,747 samples

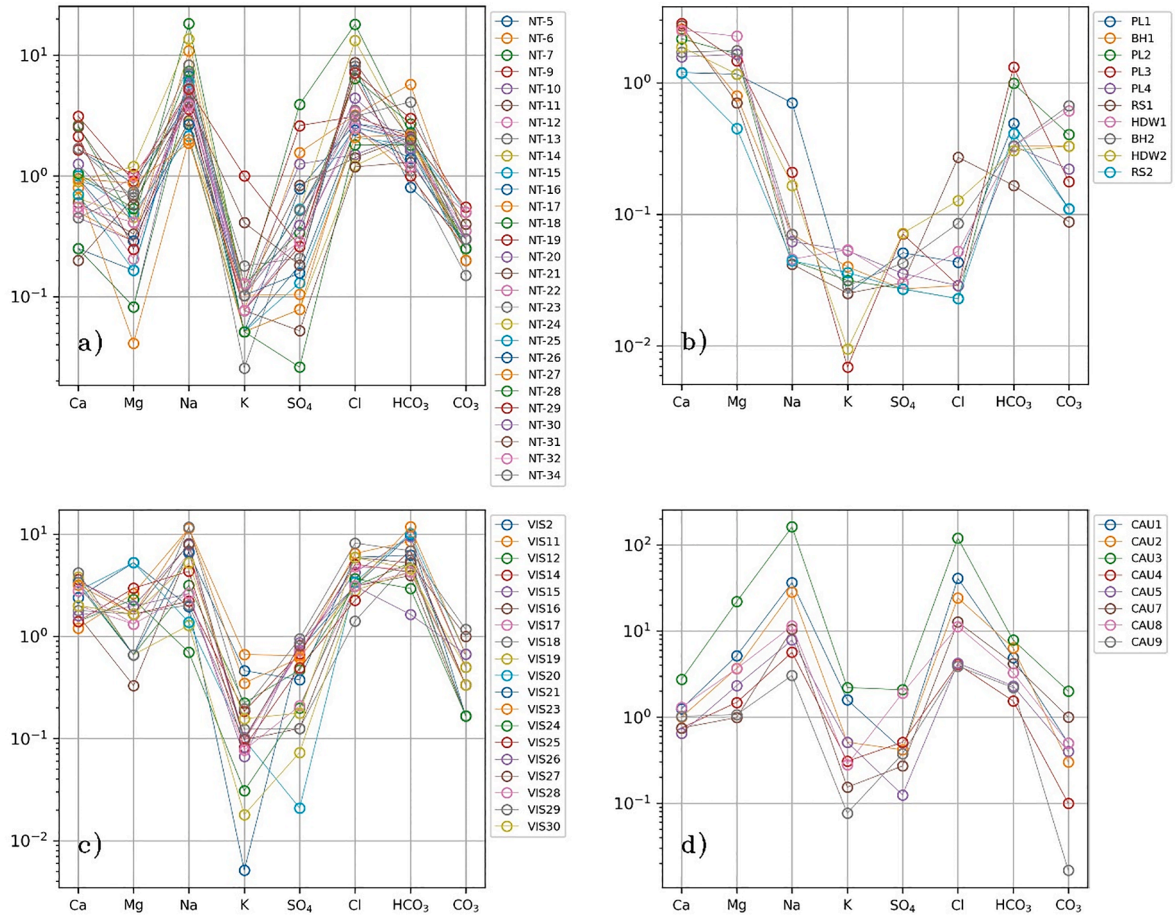


Fig. 5. Concentration diagrams (mM/L) in log scale to illustrate four application cases: (a) Nirmal province, India (Adimalla et al., 2019); (b) Ngbo, Ebonyi State, Nigeria (Ifediegwu et al., 2019); (c) Fluoride rich groundwater from Sattenapalle Region, Guntur district, Andhra Pradesh, India (Subba Rao et al., 2019); (d) Cauvery, India (Sajil Kumar et al., 2020).

correctly classified of a total of 46,292 samples of the outlier-free database (Verma et al., 2021); whereas a total of 3,545 samples incorrect classified, which corresponds to 7.66% of the total database. For the classification of anion classes (SO_4, Cl, HCO_3, CO_3) the individual class success goes from 92.75 to 93.75, providing an overall classification accuracy of 93.04.

The overall classification accuracy of the remaining models for predicting the cation and anion classes are 99.71 and 99.96 for WClassCB 99.67 and 99.77 for WClassVL 98.34 and 99.38 for WClassVP and 98.96 and 99.26 for WClassVR. Percent and number of samples incorrectly classified are also listed in Table 3.

3.2. External validation of all classification models

The external validation database contains a total of 8,000 samples on molar concentrations (mM/L), each consisting of 8 ions (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4^{-2}, Cl^{-}, HCO_3^{-}, CO_3^{2-}). This database is independent of the set used to train the WClassCB, WClassVL, WClassVP, and WClassVR models. Model WClassHLR was also applied to this external dataset.

In Table 4, we show the values of *classification accuracy*, *macro-average precision*, and *macro-average recall* for training and external validation sets for each model. For the training set, WClassCB presented the best overall results for both cations (99.71%) and anions (99.96%) models in all metrics. WClassVL is the second-best model to classify both cations (99.67%) and anions (99.77%) in this set. Further, WClassVR, WClassVP, and WClassHLR models are in third, fourth, and fifth place, respectively. On the other hand, for the external validation set, WClassVL presented the best overall results for both cations (99.51%) and anions (99.84%) models in all metrics; while WClassVP, WClassVR, WClassCB, and WClassHLR models are in the second, third, fourth and fifth place. The macro-average *AUC* scores for each model are presented in Table S6, all models have an *AUC* score equal to 1 in both training and external validation sets.

3.3. Application for field samples

In this section, we present four application cases for illustrating (Figs. 5 and 6) the use of these classification schemes: WClassHLR, WClassCB, WClassVL, WClassVP, and WClassVR models as well as the GMC model. These application cases are constituted by groundwater samples from India and Nigeria. Of each application case, only the samples with complete and non-zero concentrations in the 8 major ions were selected. We used L2 normalization to scale field water samples before applying the models based on CatBoost and SVM. This normalization was not used for the WClassHLR model, as it was not part of the original proposal by Verma et al. (2021). For these application cases, we also show the results obtained from CCWater program (Pérez-Espinosa et al., 2019) for the application of Hill-Piper diagrams.

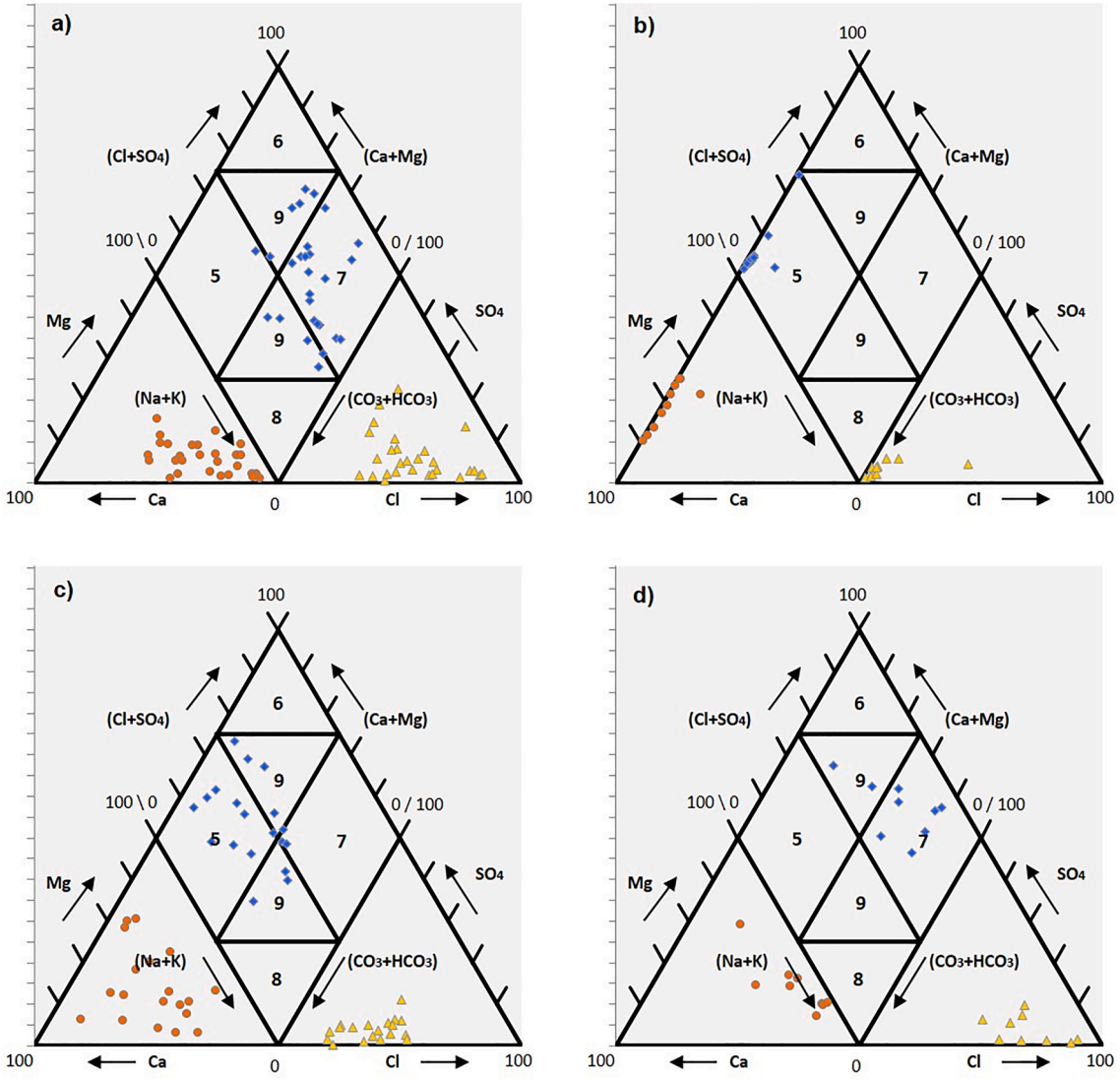


Fig. 6. Hill-Piper ternary diagrams to illustrate four application cases: (a) Nirmal province, India (Adimalla et al., 2019); (b) Ngbo, Ebonyi State, Nigeria (Ifediegwu et al., 2019); (c) Fluoride rich groundwater from Sattenapalle Region, Guntur district, Andhra Pradesh, India (Subba Rao et al., 2019); (d) Cauvery, India (Sajil Kumar et al., 2020).

3.3.1. *Groundwater from Nirmal Province, South India*

Chemical compositions of 34 groundwater samples of Nirmal Province in South India were presented by Adimalla et al. (2019). Only 28 samples (Fig. 5a) with complete and non-zero concentrations in the 8 major ions were selected. The chemical composition (mMol/L) and results of water classification obtained for all the models are presented in Table 5. The water types determined by the major ions criteria (GMC column in Table 5) are distributed only in two water types: *Na −Cl* (21 samples) and *Na −HCO_3* (7 samples). The Hill-Piper diagram through CCWater program (Pérez-Espinosa et al., 2019; Fig. 6b and Table 5) indicates that the samples are spread over in three zones as follow: 17 samples of zone 7–Noncarbonate alkali > 50% (alkalies and strong acids dominate)–, 10 samples of zone 9 –No cation–anion pair > 50%. –, and 1 sample of zone 5 –Carbonate hardness > 50% (alkaline earths and weak acids dominate)–. The "Water nomenclature" column provides the water nomenclature a total of 16 basic classes, and the "Hybrid water nomenclature" column can provide up to 256 hybrid water classes. All models performed well in classifying the water types, particularly the WClassHLR, WClassCB, and WClassVP models, which correctly predicted all water samples. The remaining models (WClassVL and WClassVR) incorrectly predicted only one sample (NT-18). The WClassHLR model, unlike the other models, obtained hybrid types of water for most of the samples (26 out of a total of 28); the WClassVP model identified 4 samples as hybrid types, and the remaining models identified only 1 sample as a hybrid type. A graphical representation of the mMol/L concentrations of these 28 groundwater samples is presented in Fig. 5a.

3.3.2. *Carbonate aquifers samples from Ngbo, Ebonyi State, Nigeria*

Ifediegwu et al. (2019) reported compositional data of 10 samples collected from pit lakes (PL), hang dug wells (HDW), boreholes (BH), and rivers (RS) from Ngbo and environs in Ebonyi State, southeastern, Nigeria, to ascertain the major ion chemistry and quality of waters for domestic and drinking uses. A graphical representation of the mMol/L concentrations of these 10 samples is presented in Fig. 5b; where each sample is identified by the initials of the place, where they were sampled. These samples were processed through all the classification models and the results of each model, including the "basic" and "hybrid" water nomenclature are summarized in Table 6.

These authors used the Hill-Piper diagram (Piper, 1944) to determine the water types (units in %mEq/l), they did not used the segmented diamond but part of the diagram to determine the major

Table 5
Application of the classification methods for the nomenclature of the groundwater samples from Nirmal Province, India (Adimalla et al., 2019).

Sample		Chemical composition (mMol/L)								Hill-Piper diagram (Fig. 6a)				
		Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	%Cl + SO4	%HCO3	%Na + K	%Ca + Mg	Class
1	NT-5	0.25	0.16	6.70	0.05	0.78	2.71	2.29	0.35	58.78	41.22	89.07	10.93	Zone 7
2	NT-6	0.50	0.29	10.87	0.08	1.56	3.19	5.74	0.30	49.9	50.1	87.43	12.57	Zone 9
3	NT-7	2.62	0.41	18.27	0.10	3.90	18.00	2.20	0.20	90.86	9.14	75.19	24.81	Zone 7
4	NT-9	3.12	0.91	5.22	0.13	2.60	3.10	2.10	0.55	72.21	27.79	39.92	60.08	Zone 9
5	NT-10	0.25	0.08	7.00	0.10	1.25	1.50	2.21	0.20	60.45	39.55	91.46	8.54	Zone 7
6	NT-11	0.20	0.66	5.26	0.10	0.83	1.41	2.10	0.25	54.22	45.78	75.77	24.23	Zone 7
7	NT-12	0.95	0.21	5.13	0.10	0.52	2.51	2.05	0.30	57.28	42.72	69.4	30.6	Zone 7
8	NT-13	0.90	0.70	4.44	0.03	0.52	3.50	1.61	0.25	68.31	31.69	58.27	41.73	Zone 7
9	NT-14	0.65	0.45	2.74	0.05	0.08	1.21	1.97	0.25	35.69	64.31	55.9	44.1	Zone 9
10	NT-15	1.00	0.49	2.13	0.05	0.13	2.09	1.80	0.20	51.59	48.41	42.25	57.75	Zone 9
11	NT-16	1.65	0.86	4.13	0.05	0.18	7.50	0.80	0.30	84.87	15.13	45.45	54.55	Zone 9
12	NT-17	0.80	0.04	2.00	0.05	0.08	2.51	1.00	0.25	64.01	35.99	55	45	Zone 7
13	NT-18	0.25	0.08	5.18	0.05	0.03	1.81	1.80	0.25	44.65	55.35	88.74	11.26	Zone 9
14	NT-19	1.65	1.03	3.57	0.08	0.26	6.40	1.00	0.25	82.2	17.8	40.51	59.49	Zone 9
15	NT-20	0.45	0.29	5.61	0.08	0.29	2.71	2.05	0.30	55.33	44.67	79.42	20.58	Zone 7
16	NT-21	0.60	0.33	2.87	0.08	0.05	1.18	1.29	0.30	40.49	59.51	61.36	38.64	Zone 9
17	NT-22	0.50	0.99	4.78	0.08	0.39	3.39	1.69	0.35	63.56	36.44	62.05	37.95	Zone 7
18	NT-23	1.70	0.66	5.22	0.18	0.21	8.10	1.10	0.15	85.89	14.11	53.41	46.59	Zone 7
19	NT-24	0.90	1.19	13.57	0.10	0.52	13.17	1.90	0.40	84.03	15.97	76.58	23.42	Zone 7
20	NT-25	0.70	0.16	6.05	0.10	0.53	2.51	1.97	0.50	54.63	45.37	78.08	21.92	Zone 7
21	NT-26	0.95	0.29	2.65	0.10	0.16	2.40	1.39	0.25	58.88	41.12	52.71	47.29	Zone 7
22	NT-27	0.95	0.86	1.87	0.10	0.10	2.09	2.20	0.20	46.93	53.07	35.25	64.75	Zone 5
23	NT-28	1.05	0.53	7.35	0.13	0.34	6.40	2.70	0.25	68.84	31.16	70.26	29.74	Zone 7
24	NT-29	2.12	0.25	5.22	1.00	0.26	7.11	3.00	0.50	65.61	34.39	56.76	43.24	Zone 7
25	NT-30	1.25	0.41	4.00	0.10	0.39	4.40	1.70	0.30	69.22	30.78	55.3	44.7	Zone 7
26	NT-31	2.55	0.58	4.09	0.41	0.18	8.69	2.00	0.40	76.38	23.62	41.88	58.12	Zone 9
27	NT-32	0.55	0.41	3.87	0.13	0.29	2.51	1.20	0.50	58.4	41.6	67.55	32.45	Zone 7
28	NT-34	0.45	0.74	8.31	0.10	0.52	3.10	4.10	0.30	46.87	53.13	77.95	22.05	Zone 9

cation and anion in all the samples. They determined that the main water types were $Ca-CO_3$ and $Mg-HCO_3$. The Hill-Piper diagram (CCWater program; Pérez-Espinosa et al., 2019; Fig. 6b and Table 6) indicates that all samples are of zone 5 (alkaline earths and weak acids dominate). The initial GMC criteria based nomenclature indicated that the carbonate aquifers from Ngbo samples were distributed as follows: $Ca-CO_3$ (3 samples), $Ca-Cl$ (1 sample), $Ca-HCO_3$ (4 samples), $Mg-CO_3$ (1 sample), and $Mg-HCO_3$ (1 sample). Considering the GMC criterion as a reference, the WClassVL model obtained the highest number of coincidences or correctly classified samples (with 9 out of 10 samples), followed by WClassHLR model (with 8 out of 10 samples), WClassVP (5 samples), WClassVR (2 samples), and WClassCB (2 samples). In this dataset, WClassVL was the best model (only 1 misclassification).

In the hybrid water nomenclature (Table 6), WClassHLR model obtained hybrid types of water for all samples; WClassVL and WClassVP predicted 3 samples as hybrid water types, WClassCB model predicted only 1, and S WClassVR did not identify any.

3.3.3. *Fluoride rich groundwater from Sattenapalle Region, Guntur district, Andhra Pradesh, India*

Chemical compositions of 30 fluoride-rich groundwater samples from Sattenapalle Region, Guntur district, Andhra Pradesh, India were reported by Subba Rao et al. (2019), of which, only 19 samples were processed through the classification models (Fig. 5c), and samples with missing concentrations were discarded. The chemical composition of each sample is presented in Table 7 along with the classification of each model.

The Hill-Piper diagram through CCWater program (Pérez-Espinosa et al., 2019) indicated that the water samples are spread over in three zones as follow (Fig. 6c and Table 7): 3 samples of zone 7 (alkalies and strong acids dominate), 9 sample of zone 5 (alkaline earths and weak

Table 6
Application of the classification methods for the nomenclature of the carbonate aquifers samples from Ngbo, Ebonyi State, Nigeria (Ifediegwu et al., 2019).

Sample		Chemical composition (mMol/L)								Hill-Piper diagram (Fig. 6b)			
		Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	%Na + K	%Ca + Mg	%Cl + SO4	%HCO3
1	PL1	1.20	1.16	0.70	0.03	0.05	0.04	0.49	0.11	16.94	83.06	13.33	86.67
2	BH1	2.56	0.79	0.07	0.04	0.03	0.03	0.33	0.33	7.73	92.27	1.62	98.38
3	PL2	2.16	1.65	0.04	0.03	0.03	0.02	0.99	0.40	4.12	95.88	0.99	99.01
4	PL3	2.81	1.47	0.21	0.01	0.07	0.03	1.31	0.18	9.27	90.73	2.46	97.54
5	PL4	1.57	1.65	0.06	0.05	0.04	0.03	0.33	0.22	11.46	88.54	1.76	98.24
6	RS1	2.71	0.70	0.04	0.03	0.03	0.27	0.17	0.09	49.23	50.77	0.97	99.03
7	HDW1	2.50	2.25	0.05	0.05	0.03	0.05	0.33	0.61	6.78	93.22	1.04	98.96
8	BH2	1.70	1.76	0.07	0.03	0.04	0.09	0.33	0.67	9.28	90.72	1.41	98.59
9	HDW2	1.87	1.16	0.17	0.01	0.07	0.13	0.30	0.33	21.93	78.07	2.81	97.19
10	RS2	1.19	0.45	0.04	0.04	0.03	0.02	0.41	0.11	10.87	89.13	2.4	97.6

Zone 5: carbonate hardness > 50%
(alkaline earths and weak acids dominate).

Water nomenclature						Hybrid water nomeclature				
GMC	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR
Na-Cl						Na-Cl-HCO3			Na-Cl-HCO3	
Na-HCO3						Na-HCO3-Cl				
Na-Cl						Na-Ca-Cl				
Na-Cl						Na-Ca-Cl-SO4				
Na-HCO3						Na-HCO3-Cl				
Na-HCO3						Na-HCO3-Cl				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl						Na-Cl-HCO3				
Na-HCO3						Na-HCO3-Cl				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl						Na-Ca-Cl				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl			Na-HCO3	Na-Cl	Na-HCO3	Na-Cl-HCO3				
Na-Cl						Na-Ca-Cl				
Na-Cl						Na-Cl-HCO3				
Na-HCO3						Na-HCO3-Cl			Na-HCO3-Cl	
Na-Cl						Na-Cl-HCO3				
Na-Cl										
Na-Cl										
Na-Cl						Na-Cl-HCO3				
Na-Cl						Na-Ca-Cl-HCO3			Na-HCO3-Cl	
Na-HCO3						Na-HCO3-Cl			Na-HCO3-Cl	
Na-Cl						Na-Cl-HCO3				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl						Na-Ca-Cl-HCO3				
Na-Cl						Na-Cl-HCO3				
Na-HCO3						Na-HCO3-Cl				

acids dominate), and **7 samples of zone 9** (No cation–anion pair > 50%). The initial GMC criteria based nomenclature indicated that these samples were distributed as follows: $Ca-Cl$ (1 sample), $Ca-HCO_3$ (3 samples), $Mg-Cl$ (1 sample), $Mg-HCO_3$ (2 samples), $Na-Cl$ (5 samples), and $Na-HCO_3$ (7 samples). Considering the GMC nomenclature as a reference, the WClassVL model obtained all correctly classified samples, followed by WClassHLR, WClassVP, and WClassVR models, which only had a single misclassified sample and the WClassCB model had only two mistakes out of a total of 19 samples. In the hybrid water nomenclature (Table 7), WClassHLR model obtained hybrid types for all samples; and WClassVP, WClassCB, WClassVL, and WClassVR models obtained hybrid types for 5, 3, 2, and 2 samples, respectively.

3.3.4. *Groundwater from point Calimere wetland in lower Cauvery region, India*

The chemical composition of 9 water samples was reported by Sajil Kumar et al. (2020) from Point Calimere wetland, which is in the Vedaranyam block of Nagapattinam district in India. These samples were collected to assess the impact and sources of saline intrusion on groundwater. These authors used the Hill-Piper diagram (Piper, 1944) and reported that 83% of the samples are $Na-Cl$ type, and the remaining samples are of $Ca-Na-HCO_3$ type. Only 8 samples (Fig. 5d) with complete data were processed by the classification models (Fig. 2d). The chemical composition (mMol/L) and water nomenclature determined by the models are shown in Table 8. According to the major ions criteria (GMC) all 8 samples are Na-Cl type. All models classified all 8 samples correctly and only WClassHLR model identified hybrid water types (7 samples). The Hill-Piper diagram through CCWater program (Pérez-Espinosa et al., 2019; Fig. 6d and Table 8) indicates that 7 samples belong to zone 7 (alkalies and strong acids dominate) and 1 sample to zone 9 (No cation–anion pair > 50%).

Hill-Piper diagram (Fig. 6b)	Water nomenclature						Hybrid water nomeclature				
Class	GMC	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR
Zone 5	Ca-HCO3	Mg-HCO3	Mg-HCO3	Ca-HCO3	Mg-HCO3	Mg-HCO3	Mg-Na-HCO3-CO3		Ca-Mg-HCO3		
Zone 5	Ca-CO3	Ca-CO3	Mg-HCO3	Ca-HCO3	Ca-CO3	Mg-HCO3	Ca-Mg-CO3-HCO3	Mg-HCO3-CO3	Ca-HCO3-CO3		
Zone 5	Ca-HCO3	Ca-HCO3	Mg-HCO3	Ca-HCO3	Mg-HCO3	Mg-HCO3	Ca-Mg-HCO3-CO3				
Zone 5	Ca-HCO3	Ca-HCO3	Mg-HCO3	Ca-HCO3	Mg-HCO3	Mg-HCO3	Ca-Mg-HCO3			Mg-Ca-HCO3	
Zone 5	Mg-HCO3	Mg-CO3	Mg-HCO3	Mg-HCO3	Mg-HCO3	Mg-HCO3	Mg-Ca-CO3-HCO3				
Zone 5	Ca-Cl	Ca-Cl	Mg-Cl	Ca-Cl	Ca-Cl	Mg-Cl	Ca-Mg-Cl-HCO3				
Zone 5	Ca-CO3	Ca-CO3	Mg-CO3	Ca-CO3	Mg-CO3	Mg-CO3	Ca-Mg-CO3-HCO3				
Zone 5	Mg-CO3			Mg-Ca-CO3-HCO3		Mg-Ca-CO3					
Zone 5	Ca-CO3	Ca-CO3	Mg-CO3	Ca-CO3	Mg-CO3	Mg-CO3	Ca-Mg-CO3-HCO3			Mg-CO3-HCO3	
Zone 5	Ca-HCO3	Ca-HCO3	Mg-HCO3	Ca-HCO3	Mg-CO3	Mg-HCO3	Ca-Mg-HCO3-CO3				

Table 7
Application of the classification methods for the nomenclature of Fluoride rich groundwater from Sattenapalle Region, Guntur district, Andhra Pradesh, India (Subba Rao et al., 2019).

Sample		Chemical composition (mMol/L)								Hill-Piper diagram (Fig. 6c)				
		Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	%Cl + SO4	%HCO3	%Na + K	%Ca + Mg	Class
1	VIS2	2.40	5.27	1.95	0.01	0.73	3.10	9.83	0.17	30.97	69.03	11.33	88.67	Zone 5
2	VIS11	1.20	2.63	11.44	0.66	0.65	6.49	8.52	0.50	44.96	55.04	61.24	38.76	Zone 9
3	VIS12	3.19	0.66	3.16	0.22	0.47	3.10	5.24	0.17	42	58	30.5	69.5	Zone 5
4	VIS14	3.59	1.65	2.21	0.18	0.68	5.08	4.26	0.33	56.62	43.38	18.63	81.37	Zone 9
5	VIS15	3.39	1.97	7.88	0.10	0.73	4.23	9.51	0.50	35.13	64.87	42.65	57.35	Zone 5
6	VIS16	1.60	0.33	6.86	0.10	0.12	3.10	3.93	0.17	44	56	64.35	35.65	Zone 9
7	VIS17	2.00	1.32	6.57	0.09	0.72	4.80	4.59	0.33	54.25	45.75	50.14	49.86	Zone 7
8	VIS18	1.80	1.65	2.03	0.12	0.12	1.41	5.57	0.17	21.94	78.06	23.83	76.17	Zone 5
9	VIS19	3.79	0.66	1.28	0.02	0.07	2.82	4.26	0.50	36.06	63.94	12.75	87.25	Zone 5
10	VIS20	2.79	5.27	1.37	0.10	0.02	3.39	10.16	0.67	22.97	77.03	8.32	91.68	Zone 5
11	VIS21	3.39	0.66	6.66	0.46	0.37	5.92	6.23	0.33	49.19	50.81	46.77	53.23	Zone 5
12	VIS23	3.19	1.65	11.51	0.35	0.61	3.10	11.80	0.67	24.8	75.2	55.05	44.95	Zone 9
13	VIS24	1.40	2.30	0.70	0.03	0.20	3.67	2.95	0.17	55.3	44.7	8.94	91.06	Zone 9
14	VIS25	1.40	2.96	4.33	0.08	0.49	2.26	6.88	0.33	30	70	33.61	66.39	Zone 5
15	VIS26	1.60	1.97	2.63	0.07	0.86	3.10	1.64	0.67	61.91	38.09	27.38	72.62	Zone 9
16	VIS27	2.00	1.65	8.07	0.10	0.81	5.92	5.24	1.00	51.03	48.97	52.87	47.13	Zone 7
17	VIS28	2.79	1.32	2.63	0.08	0.21	3.10	4.26	0.33	41.66	58.34	24.78	75.22	Zone 5
18	VIS29	4.19	0.66	11.79	0.19	0.95	8.18	6.88	1.17	52.23	47.77	55.26	44.74	Zone 7
19	VIS30	2.00	1.65	5.22	0.16	0.18	5.92	4.59	0.33	54.43	45.57	42.48	57.52	Zone 9

Zone 5: carbonate hardness > 50% (alkaline earths and weak acids dominate); zone 7: Noncarbonate alkali > 50% (alkalies and strong acids dominate); zone 9: No cation–anion pair > 50%.

Table 8
Application of the classification methods for the nomenclature of the groundwater samples from Cauvery, India (Sajil Kumar et al., 2020).

Sample		Chemical composition (mMol/L)								Hill-Piper diagram (Fig. 6d)			
		Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	%Cl + SO4	%HCO3	%Na + K	%Ca + Mg
1	CAU1	1.25	5.18	36.54	1.59	0.42	40.90	4.92	0.50	87.58	12.42	74.77	25.23
2	CAU2	1.00	3.70	28.27	0.51	0.42	24.26	6.39	0.30	78.21	21.79	75.38	24.62
3	CAU3	2.74	21.97	161.81	2.20	2.08	119.89	7.87	2.00	91.27	8.73	76.84	23.16
4	CAU4	0.75	1.48	5.65	0.31	0.51	4.09	1.54	0.10	74.59	25.41	57.21	42.79
5	CAU5	0.65	2.32	7.83	0.51	0.12	4.23	2.29	0.40	59.15	40.85	58.41	41.59
6	CAU7	0.75	0.99	10.44	0.15	0.27	12.69	4.20	1.00	68.12	31.88	75.31	24.69
7	CAU8	1.30	3.65	11.44	0.28	1.91	11.28	3.28	0.50	77.92	22.08	54.21	45.79
8	CAU9	1.02	1.07	3.04	0.08	0.37	3.89	2.20	0.02	67.56	32.44	42.72	57.28

Zone 7: Noncarbonate alkali > 50% (alkalies and strong acids dominate); zone 9: No cation–anion pair > 50%.

4. Discussion

The training and external validation datasets consist of synthetic samples that were generated from Monte Carlo simulations. All synthetic samples presented an ionic charge imbalance almost perfectly at better than ± 0.00004%. The initial 16 classes were assigned from the highest anion and cation molar concentrations. Seven variables (hlr_2 to hlr_8) previously calculated with the mMol/L concentrations of the 4 major cations ($Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}$) and the 4 major anions ($SO_4{}^{-2}, Cl^{-}$, $HCO_3{}^{-}, CO_3{}^{2-}$) were used as features for the training of classification models.

The four new multidimensional water classification models (WClassCB, WClassVL, WClassVP, and WClassVR) were successfully developed and were compared with the WClassHLR model proposed by Verma et al. (2021). All models were enabled to provide the type of hybrid water. WClassCB and WClassVL models had the highest performances on the training set. WClassVL model had the highest performance on the external validation set. The assemble model WClassHLR had the lowest performance both in the training and external validation set. All anion classification models have better performance than the cation classification ones. Finally, this work involves standardized practices for the development of machine learning models using simulated data.

The recently proposed WClassHLR (7-hlr model; Verma et al., 2021) has the advantage of generating a graphical output using the discriminant functions. The new models (WClassCB, WClassVL, WClassVP, WClassVR) can easily process up to 50,000 water samples on a single run. Also, the output probabilities from these models allow us to determine hybrid water types, which increase the number of hydrochemical facies that can be determined. This is impossible with traditional methods, such as Hill-Piper diagram and its derivatives. This is a significant improvement considering that classical approaches have fewer possible hydrochemical facies, resulting in ambiguous water types.

5. Conclusions

This work highlights the importance and applicability of machine learning models for water multidimensional classification, using rich volumes of data generated through Monte Carlo simulations. The developed water classification models on this work would afford better benefits than the traditional methods (e.g. Hill-Piper diagram), since our ML-models accurately offer a diversity of basic and hybrid water types. Over performing those traditional models which only provide 5 to 6 types (e.g. Hill-Piper diagram), and two of them (zone 9) are ambiguous and overlap several water compositions. We suggest that implementing these multidimensional models should replace the use of classical methods such as Hill-Piper diagrams. Because the usage of ternary

Water nomenclature							Hybrid water nomeclature				
GMC	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR		WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR
Mg-HCO3		Mg-HCO3-Cl									
Na-HCO3		Na-HCO3-Cl									
Ca-HCO3	Na-HCO3	Na-HCO3	Ca-HCO3	Na-HCO3	Na-HCO3		Na-Ca-HCO3-Cl		Ca-Na-HCO3	Na-Ca-HCO3	
Ca-Cl	Ca-Cl	Na-Cl	Ca-Cl	Ca-Cl	Ca-Cl		Ca-Na-Cl-HCO3				Ca-Na-Cl
Na-HCO3		Na-Ca-HCO3-Cl									
Na-HCO3		Na-Ca-HCO3-Cl									
Na-Cl		Na-Cl-HCO3	Na-Cl-HCO3	Na-Cl-HCO3	Na-Cl-HCO3	Na-Cl-HCO3					
Na-HCO3		Na-Ca-HCO3-Cl									
Ca-HCO3		Ca-HCO3-Cl	Ca-Na-HCO3								
Mg-HCO3		Mg-Ca-HCO3-Cl									
Na-HCO3		Na-Ca-HCO3-Cl	Na-HCO3-Cl		Na-HCO3-Cl						
Na-HCO3		Na-HCO3-Cl									
Mg-Cl		Mg-Ca-Cl-HCO3									
Na-HCO3		Na-Mg-HCO3-Cl									
Na-Cl		Na-Mg-Cl-HCO3									
Na-Cl		Na-Cl-HCO3			Na-Cl-HCO3						
Ca-HCO3	Na-HCO3	Na-HCO3	Ca-HCO3	Na-HCO3	Na-HCO3		Na-Ca-HCO3-Cl			Na-Ca-HCO3	
Na-Cl		Na-Ca-Cl-HCO3									
Na-Cl		Na-Cl-HCO3									

Zone 5: carbonate hardness > 50% (alkaline earths and weak acids dominate); zone 7: Noncarbonate alkali > 50% (alkalies and strong acids dominate); zone 9: No cation–anion pair > 50%.

Hill-Piper diagram (Fig. 6d)	Water nomenclature						Hybrid water nomeclature				
Class	GMC	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR	WClassHLR	WClassCB	WClassVL	WClassVP	WClassVR
Zone 7	Na-Cl						Na-Cl-HCO3				
Zone 7							Na-Cl-HCO3				
Zone 7											
Zone 7							Na-Cl-HCO3				
Zone 7							Na-Cl-HCO3				
Zone 7							Na-Cl-HCO3				
Zone 7							Na-Mg-Cl				
Zone 9							Na-Cl-HCO3				

diagrams had been shown severe problems like distortion and amplification-reduction of analytical errors, as was mentioned by many researchers before.

The usefulness of the new models (WClassCB; WClassVL; WClassVP; and WClassVR) and WClassHLR model is illustrated by applications to groundwater samples from India and Nigeria. Considering the performance of the models in real cases, we can see that all have difficulties in real samples when there is not a single major cation or anion, for example, NT-18 (see *Cl* and HCO_3) sample from Nirmal Province, India (Adimalla et al., 2019), and BH1 (see HCO_3 and CO_3) sample from Ngbo, Ebonyi State, Nigeria (Ifediegwu et al., 2019). When these cases occur, it is highly useful for the model to provide hybrid water types. In this context, one of the main advantages of the WClassHLR model is that identifies more "hybrid" types than other models. However, WClassVL is the best model overall because shows higher classification accuracy in the external validation dataset and identifies "basic" and "hybrid" water types when there are similar molar concentrations. Then, we highly recommended the use of WClassVL model with hybrid water nomenclature in any future applications for water multidimensional classification.

CRediT authorship contribution statement

Lorena Díaz-González: Supervision, Conceptualization, Methodology, Writing - review & editing. **Oscar Alejandro Uscanga-Junco:** Methodology, Resources, Software, Validation, Visualization. **Mauricio Rosales-Rivera:** Conceptualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.126234.

References

Abba, S.I., Pham, Q.B., Saini, G., Thuy Linh, N.T., Ahmed, A.N., Mohajane, M., Bach, Q.-V., 2020. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environ. Sci. Pollut. Res. 27, 41524–41539. https://doi.org/10.1007/s11356-020-09689-x.

Acharya, T.D., Subedi, A., Huang, H., Lee, D.H., 2019. Classification of surface water using machine learning methods from landsat data in Nepal. Proceedings, 4(1), 43. 10.3390/ecsa-5-05833.

Adimalla, N., Li, P., Qian, H., 2019. Evaluation of groundwater contamination for fluoride and nitrate in semi-arid region of Nirmal Province, South India: a special

emphasis on human health risk assessment (HHRA). Hum. Ecol. Risk Assess. 25 (5), 1107–1124. https://doi.org/10.1080/10807039.2018.1460579.
Ahmad, N., Sen, Z., Ahmad, M., 2003. Ground water quality asssessment using multi-rectangular diagrams. Groundwater 41, 828–832. https://doi.org/10.1111/j.1745-6584.2003.tb02423.x.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London, UK.
Aitchison, J., Egozcue, J.J., 2005. Compositional data analysis: where are we and where should we be heading? Math. Geol. 37 (7), 829–850. https://doi.org/10.1007/s11004-005-7383-7.
Banadkooki, F.B., Ehteram, M., Panahi, F., Sammen, S.S., Othman, F.B., Ahmed, E.S., 2020. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. Journal of Hydrology 587, 124989.
Bayram, A.F., Gultekin, S.S., 2010. Classifying of the Simav geothermal waters with artificial neural network method. In: Proceedings World Geothermal Congress, Bali Indonesia, pp. 25–29.
Chadha, D.K., 1999. A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. Hydrogeol. J. 7 (5), 431–439. https://doi.org/10.1007/s100400050216.
Chayes, F., 1960. On correlation between variables of constant sum. J. Geophys. Res. 65, 4185–4193. https://doi.org/10.1029/JZ065i012p04185.
Chayes, F., 1971. Ratio Correlation. A Manual for Students of Petrology and Geochemistry. The University of Chicago Press, Chicago and London.
Durov, S.A., 1948. Natural waters and graphic representation of their composition. Doklady Akademii Nauk SSSR 59 (3), 87–90.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35 (3), 279–300. https://doi.org/10.1023/A:1023818214614.
Elhag, A.B., 2017. New diagram useful for classification of groundwater quality. J. Geol. Geophys. 6, 279. https://doi.org/10.4172/2381-8719.1000279.
Feng, C., Wang, H., Tu, X.M., 2013. Geometric mean of nonnegative random variable. Commun. Statistics-Theory Methods 42 (15), 2714–2717. https://doi.org/10.1080/03610926.2011.615637.
Gakii, C., Jepkoech, J., 2019. A classification model for water quality analysis using decision tree. Eur. J. Comput. Sci. Inform. Technol. 7 (3), 1–8.
Gaya, M.S., Abba, S.I., Abdu, A.M., Tukur, A.I., Saleh, M.A., Esmaili, P., Wahab, N.A., 2020. Estimation of water quality index using artificial intelligence approaches and multi-linear regression. IAES Int. J. Artificial Intell. 9 (1), 126–134. https://doi.org/10.11591/ijai.v9.i1.pp126-134.
Géron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems, Second ed. O'Reilly Media, Canada.
Giménez-Forcada, E., 2010. Dynamic of sea water interface using hydrochemial facies evolution diagram. Ground Water 48, 212–216. https://doi.org/10.1111/j.1745-6584.2009.00649.x.
Güler, C., Thyne, G.D., McCray, J.E., Turner, A.K., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. 10, 455–474. https://doi.org/10.1007/s10040-002-0196-6.
Handa, B.K., 1965. Modified Hill-Piper diagram for classification of groundwater in arid and semi-arid regions. Geochem. Soc. India Bull. 1, 20–24.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.
Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zhou, H., 2019. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J. Hydrol. 574, 1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085.
Ifediegwu, S., Onyeabor, C.F., Nnamani, C., 2019. Geochemical evaluation of carbonate aquifers in Ngbo and environs, Ebonyi State, southeastern, Nigeria. Model. Earth Syst. Environ. 5, 1893–1909. https://doi.org/10.1007/s40808-019-00646-3.
Kang, P., Lin, Z., Teng, S., Zhang, G., Guo, L., Zhang, W., 2019. Catboost-based framework with additional user information for social media popularity prediction. In: 27th ACM International Conference, pp. 2677–2681. https://doi.org/10.1145/3343031.3356060.
Law, A.M., Kelton, W.D., 2000. Simulation Modeling and Analysis. McGraw Hill, Boston.
Liang, Z., Zou, R., Chen, X., Ren, T., Su, H., Liu, Y., 2020. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. Journal of Hydrology 581, 124432.
Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169.
Matsumoto, M., Nishimura, T., 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. 8 (1), 3–33. https://doi.org/10.1145/272991.272995.
Melesse, A.M., Khosravi, K., Tiefenbacher, J.P., Heddam, S., Kim, S., Mosavi, A., Pham, B. T., 2020. River water salinity prediction using hybrid machine learning models. Water 12, 2951. https://doi.org/10.3390/w12102951.
Muhammad, S.Y., Makhtar, M., Rozaimee, A., Abdul, A., Jamal, A.A., 2015. Classification model for water quality using machine learning techniques. International Journal of Software Engineering and Its Applications 9 (6), 45–52. https://doi.org/10.14257/ijseia.2015.9.6.05.
Nachappa, T.G., Piralilou, S.T., Gholamnia, K., Ghorbanzadeh, O., Rahmati, O., Blaschke, T., 2020. Flood susceptibility mapping with machine learning, multi-criteria decision analysis an ensemble using Dempster Shafer Theory. J. Hydrol. 590, 125275. https://doi.org/10.1016/j.jhydrol.2020.125275.
Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., Liu, J., 2020. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. J. Hydrol. 586, 124901. https://doi.org/10.1016/j.jhydrol.2020.124901.
Nicholson, K., 1933. Geothermal Fluids: Chemistry and Exploration Techniques. Springer-Verlag, Berlin Heidelberg, Germany.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Machine Learn. Res. 12, 2825–2830.
Pérez-Espinosa, R., Pandarinath, K., Hernández-Campos, F.J., 2019. CCWater-a computer program for chemical classification of geothermal waters. Geosci. J. 23 (4), 621–635. https://doi.org/10.1007/s12303-018-0064-6.
Piper, A.M., 1944. A graphic procedure in the geochemical interpretation of water analyses. Trans. Am. Geophys. Union 25, 914–923.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogoush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Neural Inform. Process. Syst. 6637–6647.
Romani, S., 1981. A new diagram for classification of natural waters and interpretation of chemical analyses data. In: Proceedings of Ground Water international Symposium Noordwijkerhout. Studies in environmental science. https://doi.org/10.1016/S0166-1116(08)71980-0.
Sajil Kumar, P.J., 2013. Interpretation of groundwater chemistry using piper and chadhás diagrams: a comparative study from perambalur taluk. Elixir Geosci. 54, 12208–12211.
Sajil Kumar, P.J., Jegathambal, P., Babu, B., Kokkat, A., James, E., 2020. A hydrogeochemical appraisal and multivariate statistical analysis of seawater intrusion in point calimere wetland, lower Cauvery region, India. Groundwater Sustain. Dev. 11, 100392. https://doi.org/10.1016/j.gsd.2020.100392.
Schoeller, H., 1955. Géochemie des eaux souterraines. Revue de l'Institut Francais du Petrole 10, 230–244.
Shelton, J.L., Englea, M.A., Buccianti, A., Blondes, M.S., 2018. The isometric log-ratio (ilr)-ion plot: a proposed alternative to the Piper diagram. J. Geochem. Explor. 190, 130–141. https://doi.org/10.1016/j.gexplo.2018.03.003.
Spizman, L., Weinstein, M.A., 2008. A Note on utilizing the geometric mean: when, why and how the forensic economist should employ the geometric mean. J. Legal Econ. 15 (1), 43–55.
Subba Rao, N., Srihari, C., Spandana Deepthi, B., Sravanthi, M., Kamalesh, T., Jayadeep, V.A., 2019. Comprehensive understanding of groundwater quality and hydrogeochemistry for the sustainable development of suburban area of Visakhapatnam, Andhra Pradesh, India. Hum. Ecol. Risk Assessment Int. J. 25 (1–2), 52–80. https://doi.org/10.1080/10807039.2019.1571403.
Tung, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology 585, 124670.
Verma, S.P., 2012. Geochemometrics. Revista Mexicana de Ciencias Geológicas 29, 276–298.
Verma, S.P., 2015. Monte Carlo comparison of conventional ternary diagrams with new log-ratio bivariate diagrams and an example of tectonic discrimination. Geochem. J. 49, 393–412.
Verma, S.P., 2020. Road from Geochemistry to Geochemometrics. Springer, Singapore.
Verma, S.P., Armstrong-Altrin, J.S., 2013. New multi-dimensional diagrams for tectonic discrimination of siliciclastic sediments and their application to Precambrian basins. Chem. Geol. 355, 117–133. https://doi.org/10.1016/j.chemgeo.2013.07.014.
Verma, S.P., Quiroz-Ruiz, A., 2006. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. Revista Mexicana de Ciencias Geológicas 23, 133–161.
Verma, S.P., Diaz-González, L., Rivera-Gómez, M.A., Rosales-Rivera, M., 2020. New multidimensional classification scheme of altered igneous rocks from performance comparison of isometric and modified log-ratio transformations of major elements. Earth Sci. Inf. 13, 1031–1064. https://doi.org/10.1007/s12145-020-00473-6.
Verma, S.P., Rivera-Gomez, M.A., Díáz-González, L., Quiroz-Ruiz, A., 2016. Log-ratio transformed major-element based multidimensional classification for altered high-Mg igneous rocks. Geochem. Geophys. Geosyst. 17, 4955–4972. https://doi.org/10.1002/2016GC006652.
Verma, S.P., Uscanga-Junco, O.A., Díáz-González, L., 2021. A statistically coherent robust multidimensional classification scheme for water. Sci. Total Environ. 750, 141704. https://doi.org/10.1016/j.scitotenv.2020.141704.
Vogel, R.M., 2020. The geometric mean? Commun. Stat. - Theory Methods. https://doi.org/10.1080/03610926.2020.1743313.
Wadkar, M., Di Troia, F., Stamp, M., 2019. Detecting malware evolution using support vector machines. Expert Syst. Appl. 143, 113022. https://doi.org/10.1016/j.eswa.2019.113022.
Wang, X., Zhang, F., Ding, J., 2017. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. Sci. Rep. 7, 12858–12877. https://doi.org/10.1038/s41598-017-12853-y.
Wilks, S.S., 1963. Multivariate statistical outliers. Sankhya 25, 407–426.
Wu, L., Peng, Y., Fan, J., Wang, Y., 2019. Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data. Hydrol. Res. 50 (6), 1730–1750. https://doi.org/10.2166/nh.2019.060.
Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., Qin, H., 2020. Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. J. Hydrol. 582, 124293. https://doi.org/10.1016/j.jhydrol.2019.124293.
Zhou, Y., 2020. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques. Journal of Hydrology 589, 125164.