"El saber de mis hijos hará mi grandeza"

Sexto Congreso Nacional de Riego, Drenaje y Biosistemas

COMEII- 2021 / Hermosillo, Sonora

Detección de acuíferos sobreexplotados mediante técnicas de aprendizaje automático no supervisado

Alberto González Sánchez (IMTA)
Miguel Antonio Vega Castro (UPEMOR)
Ronald Ernesto Ontiveros Capurata (IMTA-Cátedras CONACyT)

IMTA
INSTITUTO MEXICANO DE TECNOLOGÍA DEL AGUA

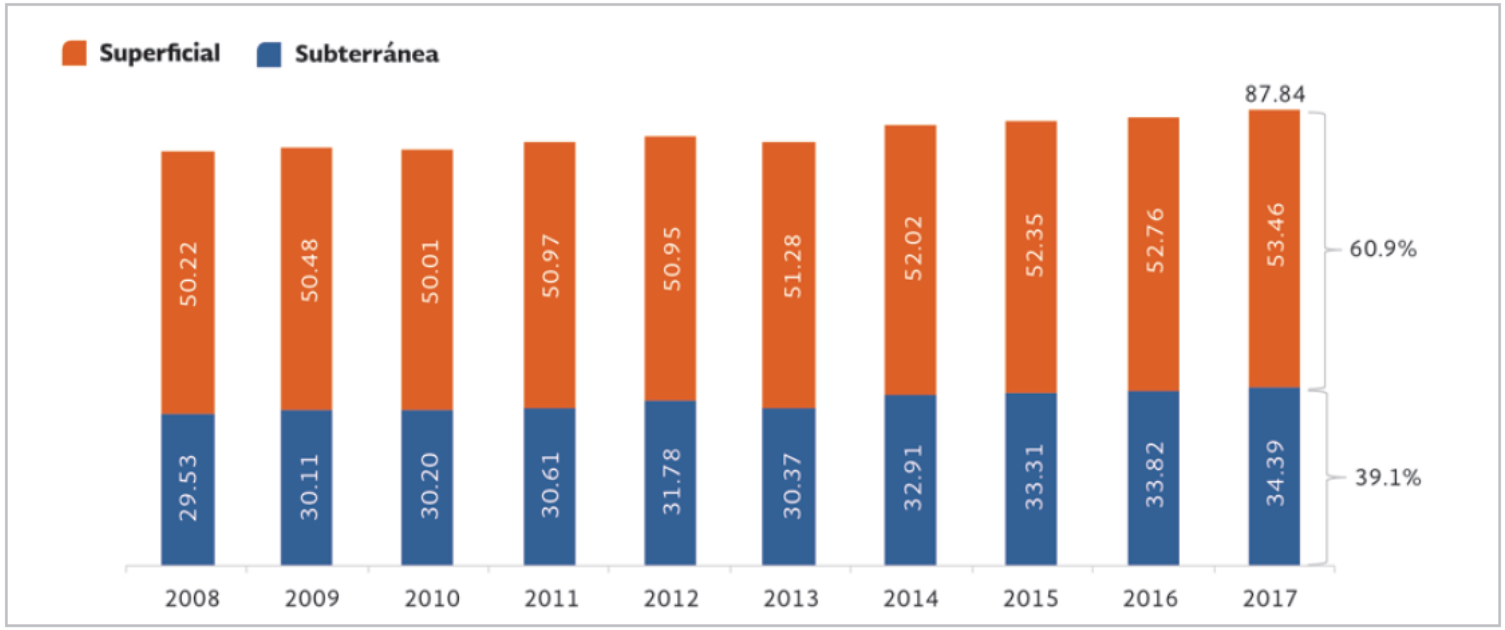
Fecha de presentación: 9 de Junio de 2021

Contenido

- Introducción
- Materiales y métodos
- Resultados
- Conclusiones

Introducción

- En México, los acuíferos aportan 36.41% del volumen anual utilizado en la agricultura y 39.1% del volumen concesionado para usos consuntivos.
- Su explotación ha estado en aumento
- Se requiere un uso sustentable del recurso para evitar su agotamiento.



Volumen concesionado para usos consuntivos (miles de hm^3)
Fuente: CONAGUA (2018)

Introducción

- Desde el 2001, la CONAGUA ha realizado estudios para determinar la Disponibilidad Media Anual (DMA) de agua en los acuíferos (NOM-011-CONAGUA-2000)
- CONAGUA publica periódicamente la DMA de los acuíferos desde el año 2009
- El volumen debe servir como referencia para el otorgamiento o limitación de las concesiones
- Sin embargo, en las mismas publicaciones se observa que el problema va en aumento.

Año de publicación	Acuíferos (653)	
	En déficit	Con disponibilidad
2009-2011	174	479
2013	193	460
2015	203	450
2018	245	408
2020	275	378

Introducción

- El aprovechamiento sustentable de los acuíferos requiere del análisis de una gran cantidad de variables involucradas.
  - El fin de contar con información para realizar un balance hídrico preciso
  - Conocer con anticipación el estado en el que pueden caer los acuíferos a futuro
- El estudio de fenómenos de naturaleza compleja normalmente se aborda con modelos físicos
- Sin embargo
  - Los acuíferos presentan una naturaleza difícil de modelar, responden a cambios en el uso de la tierra, el clima (temperatura y la precipitación), la recarga y las extracciones (Wang et al., 2018)
  - La recarga resulta muy difícil de pronosticar, ya que no puede ser medida directamente (Crosbie, Davies, Harrington, & Lamontagne, 2015; Gao, Connor, & Dillon, 2014)
  - Son costosos por la gran demanda de información, que en muchas ocasiones requieren medición de variables directamente en campo para su calibración, lo que también demanda mucho tiempo (Coulibaly, Anctil, Aravena, & Bobée, 2001)

Introducción

- Alternativa: uso de modelos basados en los datos, que se construyen a partir de algoritmos de aprendizaje automático (*machine learning*).

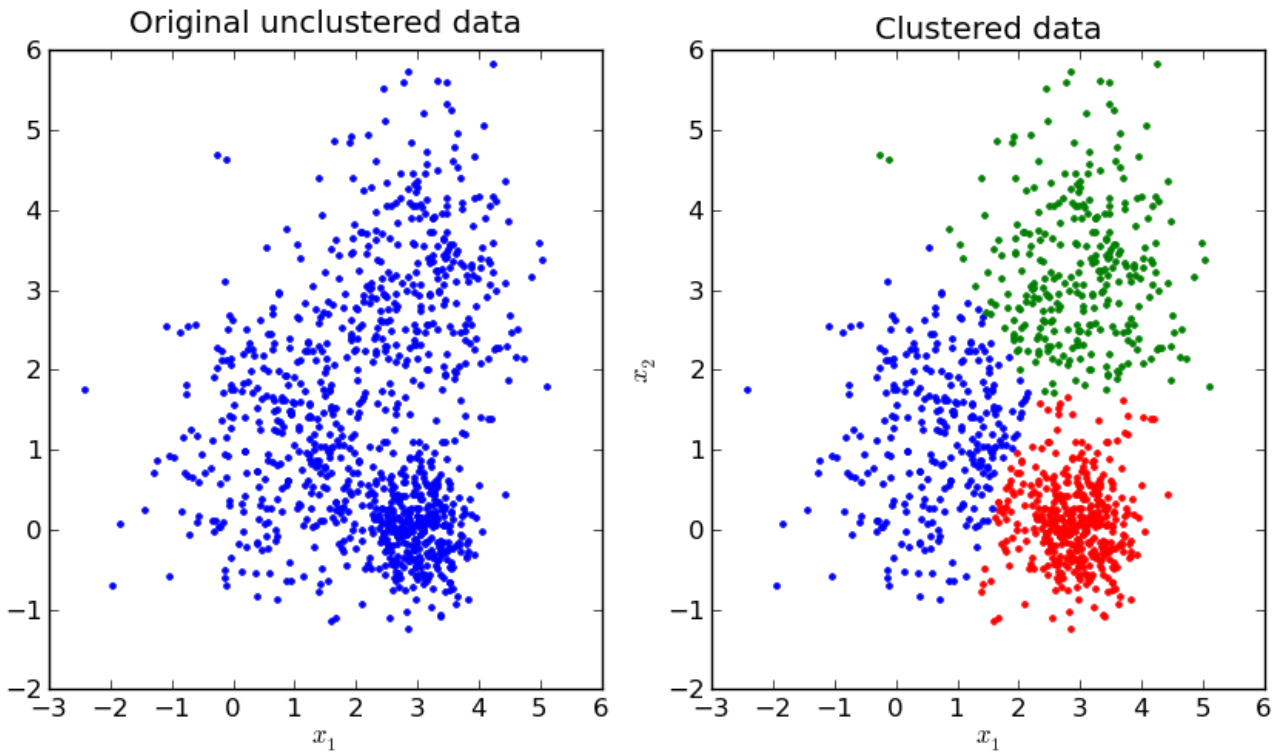
Steyn, (2018): El aprendizaje automático puede identificar patrones directamente desde la información, capturar tendencias y comportamientos de variables sin el conocimiento profundo de los atributos físicos subyacentes de los modelos de flujo de agua subterránea.

- Con este enfoque se han aplicado RNAs, Random Forest y SVM (aprendizaje supervisado); usualmente, se utilizan para determinar los niveles a los que se encuentra el agua subterránea
- Limitantes: El aprendizaje supervisado demanda contar con datos previamente etiquetados y con suficiente soporte histórico.
- A diferencia, los métodos no supervisados no requieren información a priori de las etiquetas de clase de los objetos como entrada, y pueden inferir relaciones desconocidas a partir de las características de los objetos (Ej. k-Means (MacQueen, 1967) y Fuzzy cMeans Clustering (FCM) (Bezdek, Ehrlich, & Full, 1984))

Propuesta

- El algoritmo *nAcuifDef*, que utiliza la técnica de aprendizaje no supervisado *Fuzzy cMeans* para detectar los acuíferos que en un futuro cercano puedan caer en un estado de déficit

- La estimación realizada por *nAcuifDef* es comparada contra un pronóstico simple basado en el ordenamiento ascendente de la disponibilidad media anual



Materiales y métodos (1)

- **El primer paso fue la conformación de una base de datos con las principales características de los acuíferos**

Tipo de información	Datos de interés	Fuente de la información	Período de análisis o fecha de publicación	Formato
Climatología	Temperaturas (mínima, máxima y promedio anual en °C), precipitación y evapotranspiración potencial anual (mm)	Global Climate Monitor (GCM) (Research Climate Group, s/f)	Datos anuales del 2005 al 2009	Shapefile
Uso del suelo y vegetación (serie IV)	Delimitación nacional de uso del suelo agrupada por agricultura, asentamientos humanos, bosque, cuerpos de agua, selva, vegetación y otros. Superficie en metros cuadrados (m^2)	Instituto Nacional de Estadística, Geografía e Informática (INEGI) a través de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) (CONABIO, 2018)	Publicado en el 2009	Shapefile
Permeabilidad	Tipo de permeabilidad de la superficie (tomado de la capa de hidrogeología, puede ser baja, baja a media, media a alta o alta)	CONABIO (CONABIO, 2018)	Estudio realizado en 1990	Shapefile
Concesiones de uso de agua subterránea	Volumen concesionado de agua subterránea (en miles de m3)	CONAGUA (CONAGUA, 2017)	Actualizada a diciembre de 2019	CSV
Disponibilidad media anual de agua en los acuíferos años 2009-2011 (1ra act.), 2013 (2da act.), 2015 (3ra act.), 2018 (4ta act.) y 2020 (5ta act.)	Recarga total (R), volumen disponible de agua subterránea (DAS) o DÉFICIT (en hectómetros cúbicos). 5 series con todos los acuíferos ww $\text{POR\_DISP\_DEF}\ (\%) = \left(\frac{\text{DAS+DÉFICIT}}{R}\right) * 100$	Publicaciones de CONAGUA en el Diario Oficial de la Federación (SEGOB, s/f)	Publicaciones en distintas fechas: 28/08/2009 08/07/2010, 16/08/2010, 25/01/2011, 14/12/2011, 20/12/2013, 20/04/2015, 04/01/2018, 17/09/2020	HTML

Materiales y métodos (2)

- El algoritmo recibe el conjunto de acuíferos, genera grupos y los califica.
- La métrica de calificación puede seleccionarse.
- De cada grupo, se toma una proporción de acuíferos candidatos a caer en déficit, asociado a la proporción actual

```
# Regresa la información de los nA acuíferos próximos a caer en déficit
# agrupando en base a los atributos en datosAcuif y usando la cantidad
# nG de grupos para clasificación
nAcuifDef<-function(datosAcuif,nA,nG) {
  datosC=calificaAcuiferos(datosAcuif,nG,2)
  # damePropDef determina la proporción de acuíferos con déficit
  # en cada grupo
  propDefG=damePropDef(datosC,nG)
  nT=nA
  # difprop indica la proporción pendiente que no pudo
  # ser cubierta por los acuíferos con disponibilidad
  difprop=0
  # salida es una lista que almacena los acuíferos seleccionados
  salida=NULL
  g=1
  while (nA>0 && g<=nG) {
    # toma los acuíferos candidatos del grupo g
    # cuyo porcentaje de disponibilidad sea positivo
    candidatos=datosC[datosC$Grupo==g & datosC$POR_DISP_DEF>0,]
    candidatos=candidatos[order(candidatos$POR_DISP_DEF),]

    # El número de candidatos es proporcional a los acuíferos en déficit del cluster
    nc=round(nT*propDefG[g],0)
    if (nrow(candidatos)>=nc) { # hay suficientes candidatos, úsalos todos
      candidatos=candidatos[1:nc,]
    } else { # faltaron, calcula la diferencia de proporción pendiente
      difprop=propDefG[g]*(1-(nrow(candidatos)/nc))
      # distribuye la diferencia pendiente entre los grupos restantes
      if (difprop>0 && g<nG) {
        sumaprop=sum(propDefG[(g+1):nG])
        #actualiza la proporción para considerar la que no se cubrió pero de forma proporcional
        for (i in (g+1):nG) {
          propDefG[i]=propDefG[i]+(propDefG[i]/sumaprop)*difprop
        }
      }
    }

    if (is.null(salida)) {
      salida=candidatos
    } else
      salida=rbind(salida,candidatos)
    nA=nA-nrow(candidatos)
    g=g+1
  }
  salida=salida[order(salida$POR_DISP_DEF),]
  return (salida)
}
```

Materiales y métodos (3)

- La función que califica aplica el algoritmo *Fuzzy cMeans Clustering (library ("e1071")*

```
library("e1071")

# califica los acuíferos en datosAcuif asignando a Grupo los valores de 1 a nG
# dependiendo de la métrica de calificación utilizada. El sentido es ascendente,
# 1=el grupo con la peor calificación, nG el mejor.
# mcalif==1 utiliza la disponibilidad promedio del grupo
# mcalif==2 utiliza la proporción de acuíferos que no están en déficit (default)
calificaAcuiferos<-function(datosAcuif,nG, mcalif=2) {
  # elimina atributos a no considerar en el agrupamiento (la variable Grupo
  # puede existir si no es el primer llamado a calificaAcuif)
  tabla=datosAcuif[,!(names(datosAcuif) %in% c("CLAVE_ACUIFERO","POR_DISP_DEF","Grupo"))]
  cmres=cmeans(tabla,nG)
  dpc=rep(0,length(cmres$clusters))
  for (i in 1:nG) {
    if (mcalif==1) {# la disponibilidad promedio del grupo
      dpc[i]=mean(datosAcuif[cmres$cluster==i,]$POR_DISP_DEF)
    }
    if (mcalif==2) # proporción de acuíferos no sobre-explotados
      dpc[i]=nrow(datosAcuif[cmres$cluster==i & datosAcuif$POR_DISP_DEF>0,])/
          nrow(datosAcuif[cmres$cluster==i,])
  }
  mi.dpc=data.frame("dpc"=dpc,"cluster"=(1:nG))
  mi.dpc=mi.dpc[order(mi.dpc$dpc),]
  # la siguiente línea crea la columna "Grupo"
  datosAcuif$Grupo=rep(-1,nrow(datos))
  for (i in 1:nG) {
    datosAcuif[cmres$cluster==mi.dpc$cluster[i],"Grupo"]=i
  }
  return(datosAcuif)
}
```

El algoritmo Fuzzy cMeans Clustering➔

Materiales y métodos (4)

1) Especifica de antemano el número de grupos (conglomerados) requeridos (parámetro *nG* en la función *calificaAcuiferos*), de tal forma que *nG* $\in$ [2,n) donde n es el número máximo de acuíferos (653). Inicializa el parámetro de ponderación m > 0 (usualmente 2), la cota de error ε > 0 y el número máximo de iteraciones $maxT$.
2) Proporciona pesos de forma aleatoria $\mu_{ij}^{(0)} \sim U(0,1)$. μ_{ij} es el valor de pertenencia del i-ésimo dato al j -ésimo grupo (clúster).
3) Inicializa el contador de iteraciones ($t = 1$)
4) Calcula los centroides de los nG *clústers* usando la Ecuación (2):

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \qquad (j = 1,2,\ldots,nG) \tag{2}$$

5) Actualiza el valor de pertenencia μ_{ij} con v_j:

$$\mu_{ij} = \left(\sum_{k=1}^{nG} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \qquad (i = 1,2,\ldots,n;\ j = 1,2,\ldots,nG) \tag{3}$$

Donde x_i representa el vector de características del objeto i (acuífero i).

6) Calcula $e = \left\| \mu^{(t)} - \mu^{(t-1)} \right\|$
7) Si $e < \varepsilon$ o $t > maxT$ detén el algoritmo, en caso contrario $t = t + 1$ y regresa al paso 4

Resultados (1)

- El algoritmo fue ejecutado tomando como base la información del estado de disponibilidad de los acuíferos asociada al año 2009 y sus características.

- Esto representa una simulación de lo que hubiera pasado si se hubiera utilizado el algoritmo con la información disponible en 2009.

- La clasificación por *nAcuifDef* fue comparada contra los resultados de una predicción simple basada en el ordenamiento ascendente del porcentaje de disponibilidad de los acuíferos que no estaban en déficit en dicho año
  - Se utilizó el porcentaje de disponibilidad de esos mismos acuíferos publicado en los años subsecuentes

Resultados (2)

- Ejemplo: primeros 20 acuíferos ordenados por porcentaje de disponibilidad (año 2009) y los 20 acuíferos seleccionados por *nAcuifDef*

Clave	Porcentaje de disponibilidad (%)					En déficit
	2009	2013	2015	2018	2020	
0844	0.021	23.874	23.874	19.876	-14.383	1
0843	0.045	0.045	0.045	20.959	20.672	0
0846	0.070	0.358	0.358	23.751	22.961	0
0209	0.178	-0.222	-8.811	8.252	5.887	1
2625	0.284	25.485	26.867	24.059	3.344	0
1505	0.390	0.372	0.304	-0.436	-0.438	1
0306	0.445	1.023	0.979	-20.579	-20.746	1
2632	0.859	41.871	41.993	16.862	11.251	0
2641	1.176	1.588	2.261	0.061	-0.085	1
2105	1.285	14.375	14.428	10.933	11.421	0
1804	1.499	23.908	17.091	17.272	13.439	0
0601	1.557	5.256	3.403	2.071	3.245	0
2412	1.623	-0.002	-0.002	0.728	-1.542	1
2624	1.812	8.070	8.199	7.694	4.325	0
2025	1.833	10.480	9.766	8.212	5.150	0
1917	1.841	-0.930	-1.261	-17.044	-40.118	1
2205	1.855	4.785	6.732	1.108	-2.373	1
0825	1.867	1.867	1.867	49.524	-54.962	1
1321	2.087	0.836	0.827	0.788	0.559	0
0815	2.204	2.204	2.204	0.265	-0.128	1
	0	**3**	**3**	**3**	**9**	**10**

Clave	Porcentaje de disponibilidad (%)					En déficit
	2009	2013	2015	2018	2020	
0844	0.021	23.874	23.874	19.876	-14.383	**1**
0306	0.445	1.023	0.979	-20.579	-20.746	**1**
2412	1.623	-0.002	-0.002	0.728	-1.542	**1**
1917	1.841	-0.930	-1.261	-17.044	-40.118	**1**
2205	1.855	4.785	6.732	1.108	-2.373	**1**
0825	1.867	1.867	1.867	49.524	-54.962	**1**
0201	2.753	45.618	45.633	1.659	1.608	**0**
0312	3.515	11.493	11.493	6.467	-2.944	**1**
0506	4.066	-7.815	-10.786	-21.629	-25.655	**1**
0501	4.080	3.988	14.847	2.882	-7.094	**1**
1622	7.267	3.613	2.643	1.391	-0.013	**1**
1920	7.819	7.819	7.819	3.638	19.717	**0**
0248	8.481	13.385	9.630	7.081	5.071	**0**
1101	9.997	10.323	9.964	4.418	9.572	**0**
3216	10.954	10.409	8.804	24.356	7.138	**0**
0834	11.855	13.261	13.154	-45.211	-127.909	**1**
2601	12.245	-3.179	-3.179	17.540	-39.424	**1**
3209	23.603	13.777	6.260	6.875	0.055	**0**
1017	79.983	79.983	79.983	62.577	-257.440	**1**
1103	97.469	71.497	71.497	-14.981	71.046	**1**
	0	**4**	**4**	**5**	**13**	**14**

Resultados (3)

Rango de acuíferos	Acuíferos en déficit en el futuro de acuerdo con el orden de 2009	Acuíferos en déficit correctamente pronosticados por nAcuifDef por cada nG (número de clústers) utilizado					
		nG=10	nG=20	nG=30	nG=40	nG=50	nG=60
10	5	5	5	6*	5	6*	6*
20	10	10	11	12	14	14	16*
30	15	14	14	17	17	19	21*
40	19	18	25	19	18	22	26*
50	22	22	25	26	28*	27	28*
60	28	25	29	30	33*	30	32
70	28	26	31	43*	41	38	32
80	33	32	35	39	44*	41	42
90	41	34	37	44	47	49*	47
100	46	37	40	50	54*	52	50

Conclusiones

- Se presenta el algoritmo nAcuifDef basado en la técnica de clustering difuso (*Fuzzy CMeans, FCM*) para detectar los próximos acuíferos a caer en déficit
- Las pruebas indican necesario un proceso de calibración pero obtiene mejores resultados comparado con el ordenamiento basado en el porcentaje de disponibilidad de los acuíferos.
- Esto es de gran utilidad para monitorear acuíferos que de momento parecen no correr riesgo, pero que por sus características podrían caer en déficit pronto.
- Dada la gravedad del problema de la sobreexplotación de los recursos hídricos actuales y la velocidad con la que este crece, se requieren de técnicas que respondan de la misma forma, en lo que el aprendizaje no supervisado tiene mucho que aportar.
- Aún quedan parámetros por mejorar en el algoritmo, como la inclusión de otro tipo de atributos para clasificar los acuíferos, la determinación automática del número de grupos requerido y la métrica para calificar los acuíferos.
- Se considera que el algoritmo presentado es un ejemplo de la forma que el aprendizaje automático se pueden aplicar para lidiar con problemas relacionados con el aprovechamiento de los recursos hídricos de naturaleza difícil o impredecible.

"El saber de mis hijos hará mi grandeza"

Sexto Congreso Nacional de Riego, Drenaje y Biosistemas

COMEII- 2021 / Hermosillo, Sonora

¡GRACIAS!

Alberto González Sánchez

Instituto Mexicano de Tecnología del Agua

alberto_gonzalez@tlaloc.imta.mx