Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, PCA, and Spatial Analysis. A Case Study

Supplementary Materials

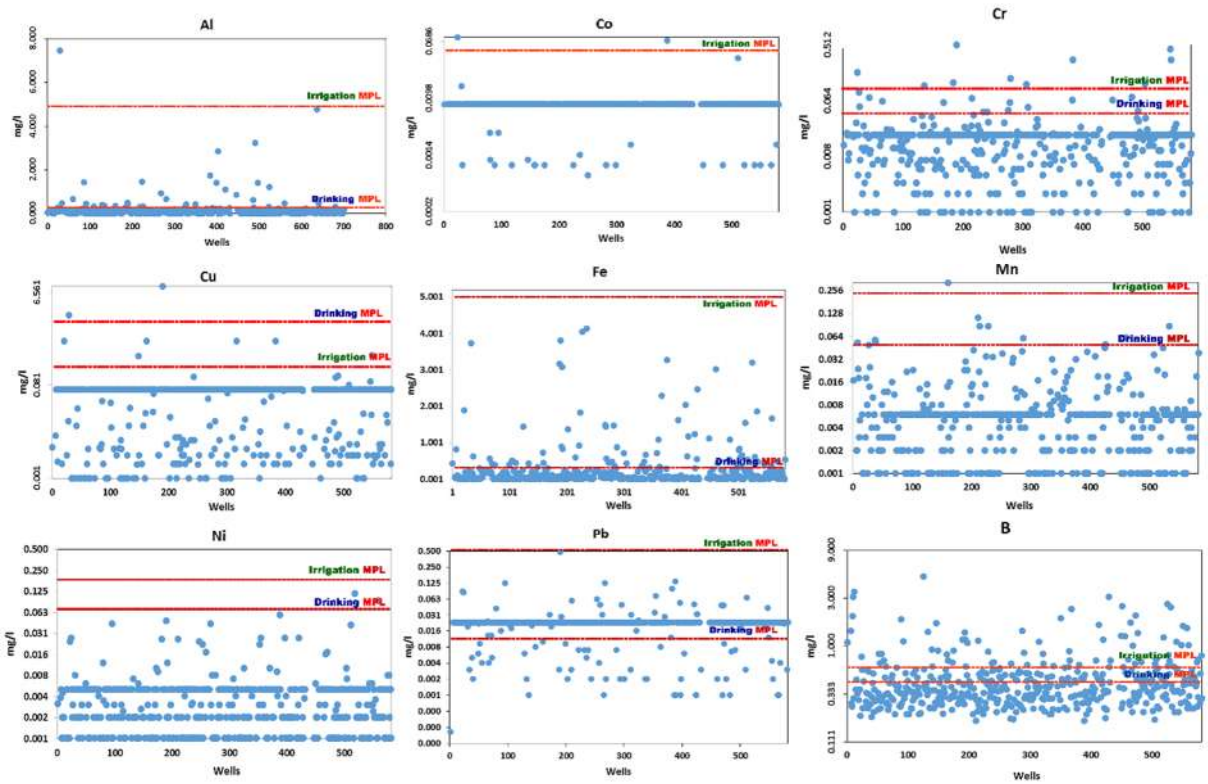

Figure S1. Comparison between the concentrations of ions and Maximum Permissible Limit (MPL) recommended by national standards for drinking water and irrigation water.

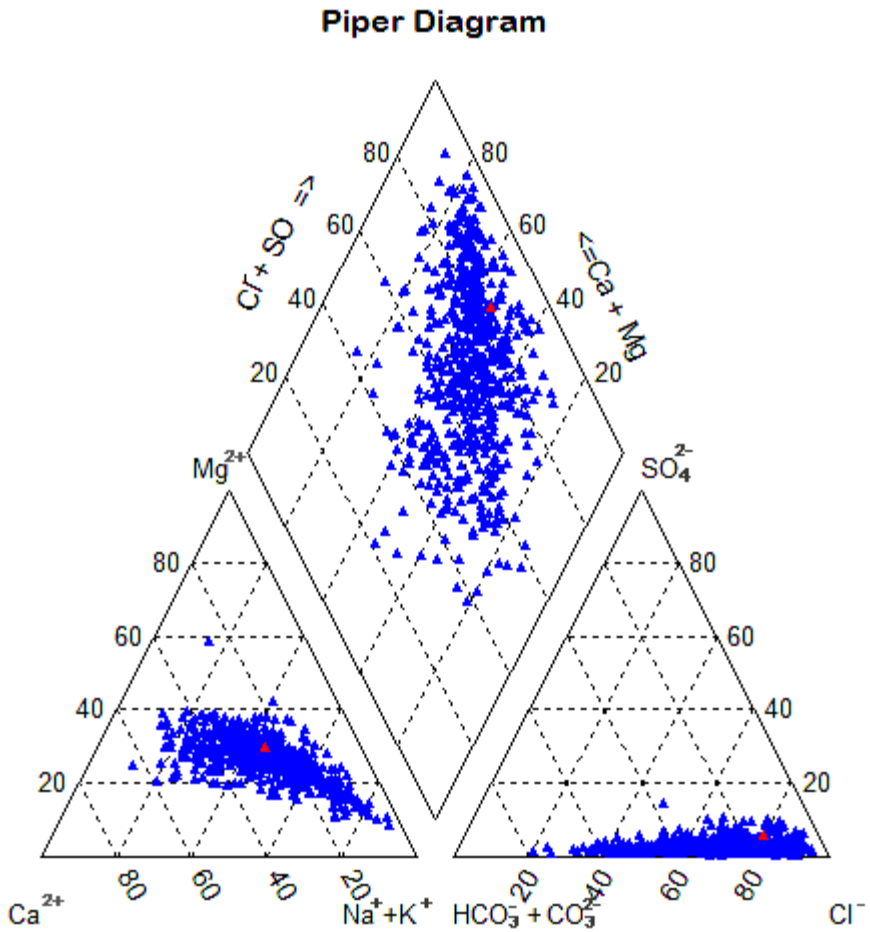

Figure S2. Groundwater types plotted on a Piper trilinear diagram.

Table S1. Indexes provided by R NbClust Package [1].

Index	Equation
Hartigan [2]	$\mathbf{Hartigan} = \left(\frac{\mathbf{trace}\left(\mathbf{W_q}\right)}{\mathbf{trace}\left(\mathbf{W_{q+1}}\right)} - \mathbf{1}\right)(\mathbf{n} - \mathbf{q} - \mathbf{1})$
Scott [3]	$\mathbf{Scott} = \mathbf{nlog}\frac{\mathbf{det}(\mathbf{T})}{\mathbf{det}(\mathbf{W_q})}$
Marriot [4]	$\mathbf{Marriot} = \mathbf{q^2det}(\mathbf{W_q})$
Trcovw [5]	$\mathbf{Trcovw} = \mathbf{trace}\left(\mathbf{COV}\left(\mathbf{W_q}\right)\right)$
Tracew [6]	$\mathbf{Tracew} = \mathbf{trace}(\mathbf{W_q})$
Friedman [7]	$\mathbf{Friedman} = \mathbf{trace}(\mathbf{W_q^{-1}B_q})$
Rubin [8]	$\mathbf{Rubin} = \frac{\mathbf{det}(\mathbf{T})}{\mathbf{det}(\mathbf{W_q})}$
Cindex [9]	$\mathbf{Cindex} = \frac{\mathbf{S_w} - \mathbf{S_{min}}}{\mathbf{S_{max}} - \mathbf{S_{min}}}$
Ball [10]	$\mathbf{Ball} = \frac{\mathbf{W_q}}{\mathbf{q}}$
Ptbiserial [11]	$\mathbf{Ptbiserial} = \frac{[\overline{\mathbf{S_b}} - \overline{\mathbf{S_w}}][\mathbf{N_w N_b}/\mathbf{N_t^2}]}{\mathbf{s_d}}$
KL [12]	$\mathbf{KL}(\mathbf{q}) = \left\lvert\frac{\mathbf{DIFF_q}}{\mathbf{DIFF_{q+1}}}\right\rvert$
CH [13]	$\mathbf{CH}(\mathbf{q}) = \frac{\mathbf{trace}(\mathbf{B_q})/(\mathbf{q}-\mathbf{1})}{\mathbf{trace}(\mathbf{W_q})/(\mathbf{n}-\mathbf{q})}$
Duda [14]	$\mathbf{Duda} = \frac{\mathbf{W_k} + \mathbf{W_l}}{\mathbf{W_m}}$
Pseudot2 [14]	$\mathbf{Pseudot2} = \frac{\mathbf{V_{kl}}}{\frac{\mathbf{W_k} + \mathbf{W_l}}{\mathbf{n_k} + \mathbf{n_l} - \mathbf{2}}}$
Beale [15]	$\mathbf{Beale} = \mathbf{F} \equiv \frac{\left(\frac{\mathbf{V_{kl}}}{\mathbf{W_k} + \mathbf{W_l}}\right)}{\left(\left(\frac{\mathbf{n_m} - \mathbf{1}}{\mathbf{n_m} - \mathbf{2}}\right)\mathbf{2}^{\frac{\mathbf{2}}{\mathbf{p}}} - \mathbf{1}\right)}$

"$\boldsymbol{n}$" means number of observations, "$\boldsymbol{p}$" means number of variables, and "$\boldsymbol{q}$" means number of clusters.
$\boldsymbol{X} = \{\boldsymbol{X_{ij}}\}$, where $\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}$ and $\boldsymbol{j} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{p}$; so $\boldsymbol{X} = \boldsymbol{n} \times \boldsymbol{p}$, means data matrix of $\boldsymbol{p}$ variables measured on $\boldsymbol{n}$ independent observations.
$\overline{\boldsymbol{X}} = \boldsymbol{q} \times \boldsymbol{p}$ matrix of cluster means; $\overline{\boldsymbol{x}}$ means centroid of data matrix $\boldsymbol{X}$;
$\boldsymbol{n_k}$ means number of objects in cluster $\boldsymbol{C_k}$, and $\boldsymbol{c_k}$ means centroid of cluster $\boldsymbol{C_k}$;
$\boldsymbol{x_i}$ means $\boldsymbol{p}$ p-dimensional vector of observations of the $\boldsymbol{i}$th object in cluster $\boldsymbol{C_k}$ and $||\boldsymbol{x}|| = (\boldsymbol{x}\mathbf{T}\mathbf{x})^{1/2}$
$\boldsymbol{W_q} = \sum_{k=1}^{q}\sum_{i \in C_k}(\boldsymbol{x_i} - \boldsymbol{c_k})(\boldsymbol{x_i} - \boldsymbol{c_k})^{\mathbf{T}}$, where $\boldsymbol{W_q}$ is the within-group dispersion matrix for data clustered into $\boldsymbol{q}$ clusters;
$\boldsymbol{B_q} = \sum_{k=1}^{q}\boldsymbol{nk}(\boldsymbol{c_k} - \overline{\boldsymbol{x}})\,(\boldsymbol{c_k} - \overline{\boldsymbol{x}})^{\mathbf{T}}$, where $\boldsymbol{B_q}$ is the between-group dispersion matrix for data clustered into $\boldsymbol{q}$ clusters;
$\boldsymbol{N_t} = \frac{\boldsymbol{n}(\boldsymbol{n}-\mathbf{1})}{\mathbf{2}}$, where $\boldsymbol{N_t}$ means total number of pairs of observations in the data set;
$\boldsymbol{N_w} = \sum_{k=1}^{q}\frac{\boldsymbol{n_k}(\boldsymbol{n_k}-\mathbf{1})}{\mathbf{2}}$, where $\boldsymbol{N_w}$ means total number of pairs of observations belonging to the same cluster;
$\boldsymbol{N_b} = \boldsymbol{N_t} - \boldsymbol{N_w}$, where $\boldsymbol{N_b}$ is the total number of pairs of observations belonging to different clusters;
$\boldsymbol{S_w} = \sum_{k=1}^{q}\sum_{i,j \in C_k i<j}\boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{x_j})$, where $\boldsymbol{S_w}$ means sum of the within-cluster distances and
$\boldsymbol{S_b} = \sum_{k=1}^{q-1}\sum_{l=k+1}^{q}\sum_{i \in C_k j \in C_l}\boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{x_j})$, where $\boldsymbol{S_b}$ is the sum of the between-cluster distances.

Ratkowsky [16]	$\mathbf{Ratkowsky} = \frac{\bar{S}}{q^{1/2}}$
CCC [17]	$\mathbf{CCC} = \ln\left[\frac{1 - E(R^2)}{1 - R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{\left(0.001 + E(R^2)\right)^{1.2}}$
Silhouette [18]	$\mathbf{Silhouette} = \frac{\sum_{i=1}^{n} S(i)}{n}$
McClain [19]	$\mathbf{McClain} = \frac{\overline{S_w}}{\overline{S_b}} = \frac{S_w/N_w}{S_b/N_b}$
SDindex [20]	$\mathbf{SDindex} = \frac{\frac{1}{q}\sum_{k=1}^{q}\lVert\sigma^{(k)}\rVert}{\lVert\sigma\rVert} + \frac{D_{max}}{D_{min}}\sum_{k=1}^{q}\left(\sum_{z=1}^{q}\lVert c_k - c_z\rVert\right)^{-1}$
Dunn [21]	$\mathbf{Dunn} = \frac{\min_{1\le i<j\le q} d(C_i, C_j)}{\max_{1\le k\le q} \operatorname{diam}(C_k)}$
SDbw [22]	$\mathbf{SDbw} = \frac{\frac{1}{q}\sum_{k=1}^{q}\lVert\sigma^{(k)}\rVert}{\lVert\sigma\rVert} + \frac{1}{q(q-1)}\sum_{i=1}^{q}\left(\sum_{j=1, i\neq j}^{q}\frac{\mathrm{density}(u_{ij})}{\max\left(\mathrm{density}(c_i), \mathrm{density}(c_j)\right)}\right)$
DB [23]	$\mathbf{DB}(q) = \frac{1}{q}\sum_{k=1}^{q}\max_{k \neq l}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right)$
Frey [24]	$\mathbf{Frey} = \frac{\bar{S}_{b_{j+1}} - \bar{S}_{b_j}}{\bar{S}_{w_{j+1}} - \bar{S}_{w_j}}$

Table S2. Comparison of the groundwater quality in relation to drinking water and irritations standards.

Ítem	NOM-127-SSA1-1994	EPA	WHO	EU	FAO	Max Value
pH	8.5	8.5	7.0 - 8.0	---	8.4	8.78
EC	0.46	0	0.25	0.25	7	8.76
TDS	---	---	1000	---	2000	5606.4
HCO_3^-	---	---	120	---	---	683.2
Cl^-	250	7	250	7	350	2613.37
SO_4^{2-}	400	5.3	250	5.3	n/a	325.73
Ca^{2+}			75*			731.26
Mg^{2+}	30	n/a	30	n/a	12	238.50
K^+	---	---	12	---	---	24.24
Na^+	200	n/a	200	9	3	1051.1
NO_3^-	10	10	50	11.25	5-30	30.3
P	---	---	---	---	---	4.27
B	n/a	n/a	0.5	1.0	0.7	5
Fe	0.3	0.3	0.3	0.3	5	5.06
Mn	0.15	0.05	0.05	0.05	0.2	0.33
F	1.5	2	1.5	1.5	1	0.95
Br	n/a	0.01	0.01	0.01	n/a	5.67
Li	n/a	n/a	n/a	n/a	2.5	0.14
Cu	2	1	2	2	0.2	8.83
Zn	5	5	3	n/a	2	0.53
Al	0.2	0.2	0.2	0.2	5	7.47
Cr	0.05	0.1	0.05	0.05	0.1	0.53
Ni	n/a	n/a	0.07	0.02	0.2	0.22
Pb	0.01	0.015	0.01	0.01	5	0.49
Co	n/a	n/a	n/a	n/a	0.05	0.08
Cd	0.005	0.005	0.003	0.005	0.01	0.003

Units: Ion concentration (mg/L), pH (Standard Units), EC (∂S/m), TDS (mg/L), --- Not available.NOM-127-SSA1-1994: Mexican Official Standard. EPA: Environmental Protection Agency. EU: European Union. WHO: World Health Organization. FAO: Food and Agriculture Organization.

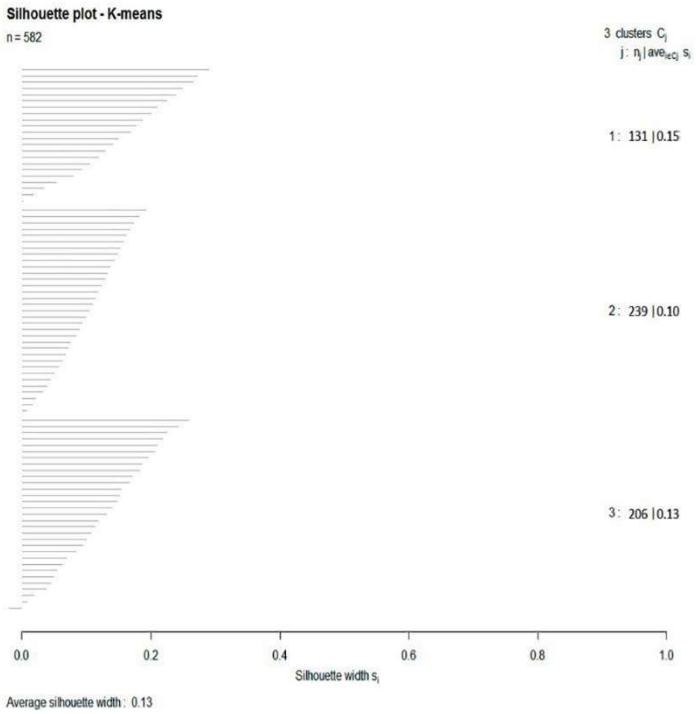

Figure S3. Silhouette plot of the clustering with high-dimensional dataset.

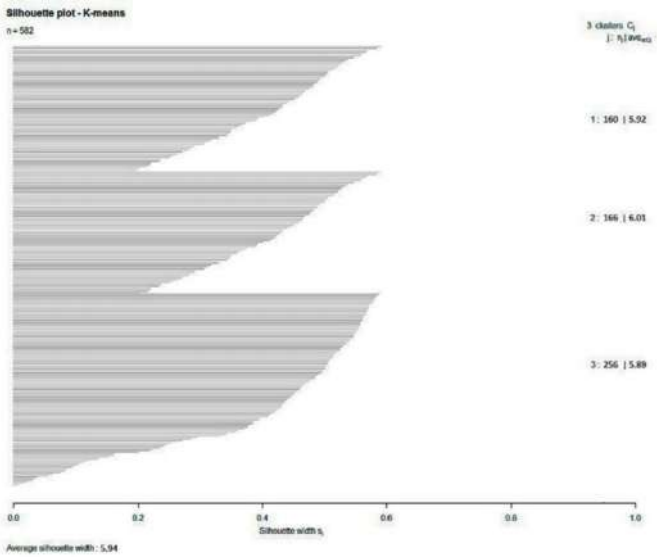

Figure S4. Silhouette plot of the clustering with low-dimensional dataset.

Table S3. Number of clusters suggested by each index.

Índex	Number of clusters suggested
Frey	**1**
CCC, Scott, McClain, SDindex	**2**
Hartigan, Silhouette, Marriot, TrCovW, TraceW, Friedman, Rubin, Cindex, Ball, PtBiserial	**3**
KL, CH, Duda, Pseudot2, Beale, Ratkowsky	**6**
DB	**8**
Dunn, SDbw	**10**

References

1. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **2014**, *61*, 1–36.
2. Hartigan, J.A. *Clustering Algorithms*. John Wiley & Sons: New York; NY, USA, 1975.
3. Scott, A. J.; Symons, M. J. Clustering methods based on likelihood ratio criteria. *Biometrics* **1971**, 27, 387–397, doi:10.2307/2529003.
4. Marriott, F.H.C. Practical Problems in a Method of Cluster Analysis. *Biometrics* **1971**, 27, 501–514, doi:10.2307/2528592.
5. Milligan, G.W.; Cooper, M.C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **1985**, *50*, 159–179, doi:10.1007/bf02294245.
6. Friedman, H.P.; Rubin, J. On some invariant criteria for grouping data. *J. Am. Statist. Assoc.* **1967**, *62*, 1159–1178, doi:10.1080/01621459.1967.10500923.
7. Fukunaga, K.; Koontz, W.L.G. A criterion and an algorithm for grouping data. *IEEE Trans. Comput.* **1970**, *C-19*, 917–923, doi:10.1109/T-C.1970.222799.
8. Hubert, L.J.; Levin, J.R. A general statistical framework for assessing categorical clustering in free recall. *Psychol. Bull.* **1976**, *83*, 1072–1080, doi:doi.org/10.1037/0033-2909.83.6.1072.
9. Ball, G.H.; Hall, D.J. Isodata, a Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute: Menlo Park, CA, USA, 1965.
10. Milligan, G.W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* **1980**, *45*, 325–342, doi:10.1007/bf02293907.
11. Milligan, G.W. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* **1981**, *46*, 187–199, doi:10.1007/bf02293899.
12. Krzanowski, W.J.; Lai, Y.T. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* **1988**, *44*, 23–34, doi:10.2307/2531893.
13. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27, doi:10.1080/03610927408827101.
14. Duda, R.O. The Ho-Kashyap procedurss. In *Pattern Classification and Scene Analysis*; John Wiley & Sons: New York; NY, USA, 1973, pp. 159–166.
15. Beale, E. *Cluster Analysis*; Scientific Control Systems Ltd: London, UK, 1969.
16. Ratkowsky, D.; Lance, G. Criterion for determining the number of groups in a classification. *Aust. Comput. J.* **1978**, *10*, 115–117.
17. Sarle, W. *SAS Technical Report A-108, Cubic Clustering Criterion*; SAS Institute Inc.: Cary, NC, USA, 1983.
18. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65, doi:https://doi.org/10.1016/0377-0427(87)90125-7.
19. McClain, J.O.; Rao, V.R. CLUSTISZ: A program to test for the quality of clustering of a set of objects. *J. Mark. Res.* **1975**, *12*, 456–460.
20. Halkidi, M.; Vazirgiannis, M.; Batistakis, Y. *Quality Scheme Assessment in the Clustering Process*; Springer: Heidelberg, Germany, 2000.
21. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104, doi:10.1080/01969727408546059.
22. Halkidi, M.; Vazirgiannis, M. Clustering validity assessment: finding the optimal partitioning of a data set. In Proceedings of 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001.
23. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227, doi:10.1109/TPAMI.1979.4766909.
24. Frey, T.; van Groenewoud, H. A cluster analysis of the D² matrix of white spruce stands in saskatchewan based on the maximum-minimum principle. *J. Ecol.* **1972**, *60*, 873–886, doi:10.2307/2258571.