WATER QUALITY ASSESSMENT WITH THE AID OF MACHINE LEARNING

Author: Kalpana Baheti

ABSTRACT

The chemical compounds and their concentration in a water body change with the environment and with the addition of effluents. With several contaminants and nutrients in the matrix, it becomes very tedious to analyze which set of interdependent reactions will occur followed by the newly developed contaminants and the concentration of old contaminants in the water. Such information is necessary to a certain extent at the beginning of an industrial venture or a project plan. This foresight is hence obtained with the aid of a repository of snapshots of chemical tests performed on the water sample in question, prior to the start of the project. These snapshots are analyzed and the effect of each individual contaminant on certain chemicals known to be present in the water are extrapolated. These predicted concentrations of known chemicals then act as analytical indicators and return the entire chemical information of the water body in question at given time in the future. The set of contaminants present in the effluents can vary, hence pattern matching algorithms map the real-time sets to the snapshots in the database to minimize computation. In this manner, quantity of effluent release and schedule of release may be re-adjusted during project planning stage, and with enough certainty, toxicity limits will not be crossed.

ANALYTICAL INDICATION

Chemical indicators are substances that give a visible sign, usually a color change, of the presence or absence of a threshold concentration of a chemical species. Their limitations, however, are that most of them are one-to-one indicators and cannot confirm several presences at once, concentrations are not provided by many of them and most importantly, they can only be used in real-time data analysis, not predictive analysis.

Analytical indicators on the other hand, observes the relationship between every nutrient with the other, determines the rate of change of one nutrient with respect to another. Consider the complexity issues:

If there were N nutrients being analyzed, the total amount of comparisons to be made would be ${}^{N}C_{2}$ which becomes considerably larger if N increases even a little. Hence, the search is filtered to analyzing sets of nutrients that have a correlation coefficient closer to 1 or -1 when compared to a single nutrient that has a strong dependence on time and physical environment and can be predicted with more certainty. To understand why we choose this arrangement, consider the diagrams below.

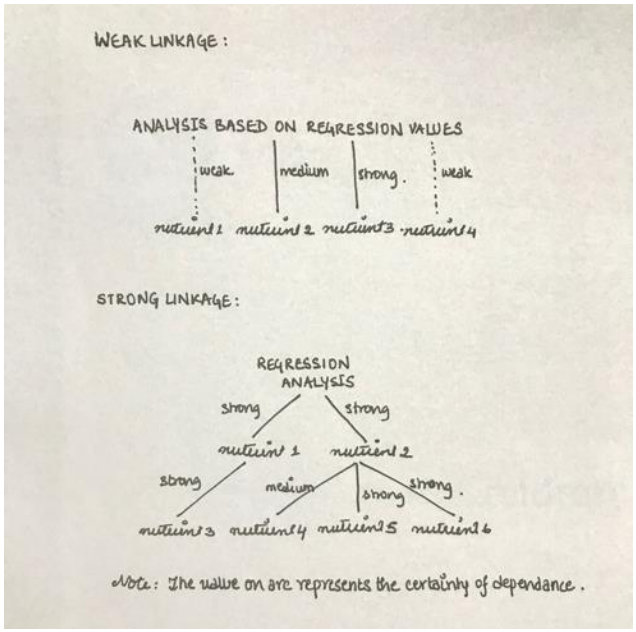


Diagram 2 has a much stronger logical linkage structure than the one in diagram 1. The correlation coeffecients near 1 and -1 (selected within a range 1 to 0.75 and -0.75 to 1 for most cases) show that the dependency is linear in nature and is strong. The sign bits + and – refer to direct dependency and inverse dependency respectively. The closer they are to 1 or -1, the stronger is the dependency. Given below are marked nutrients that have more certainty of prediction based on regression analysis, and nutrients having high linear dependency:

Parameters	Free CO_2 r t	DO r t	BOD r t	COD r t	Sulphate r t	Phosphate r t	Nitrite r t	Iron r t	Sodium r t
Free CO_2	1	-	-	-	-	-	-	-	-
Dissolved Oxygen	0.349 1.598	1	-	-	-	-	-	-	-
BOD	-0.875 ✔ 16.321	-0.836 ✔ 18.325	1	-	-	-	-	-	-
COD	-0.651 ✔ 14.321	-0.8.5 ✔ 13.035	0.91 ✔ 18.356	1	-	-	-	-	-
Sulphate	-0.369 1.298	0.165 0.558	0.529 ✔ 3.1254	0.478 2.201	1	-	-	-	-
Phosphate	-0.497 2.207	-0.497 2.02	0.572 ✔ 3.489	0.405 2.202	0.94 ✔ 17.328	1	-	-	-
Nitrite	0.197 0.565	-0.774 ✔ 14.372	0.267 1.893	0.445 2.205	-0.121 0.584	-0.132 0.5109	1	-	-
Iron	-0.395 2.105	-0.586 ✔ 2.984	0.673 ✔ 12.251	-0.883 ✔ 17.32	0.282 0.598	0.221 0.631	0.7001 ✔ 14.05	1	-
Sodium	-0.228 1.526	-0.678 ✔ 8.325	0.622 ✔ 10.824	0.669 ✔ 5.324	0.3188 0.789	0.323 0.739	0.323 0.839	0.66 ✔ 10.235	1
Calcium	-0.394 0.897	-0.853 ✔ 14.367	0.447 2.201	0.332 1.245	0.192 0.987	0.406 0.795	0.552 ✔ 2.32	0.521 ✔ 2.291	0.512 ✔ 2.31
Magnesium	-0.25 1.784	-0.727 ✔ 13.254	0.647 ✔ 9.875	0.0689 12.146	0.435 1.478	0.452 0.821	0.882 ✔ 10.245	0.799 ✔ 14.201	0.923 ✔ 17.45
Potassium	-0.213 1.495	-0.609 ✔ 8.201	0.5479 ✔ 2.998	0.509 ✔ 2.229	0.234 0.586	0.272 0.521	0.234 0.512	0.799 ✔ 14.125	0.933 ✔ 15.264
Electrical conductivity	-0.257 0.228	-0.609 9.356	0.547 10.245	0.5006 2.23	0.234 1.125	0.276 0.218	-0.756 11.32	-0.714 8.125	-0.772 10.256

	Temp	Transparency	pH	DO	BOD	Alkalinity	Chloride	TDS	TH	Nitrate	Phosphate
Temp.	-	-0.61* ✔	-0.45	-0.55 ✔	0.29	-0.11	0.56 ✔	0.01	0.76* ✔	0.09	0.31
Transparency		-	0.30	0.16	-0.35	0.18	-0.23	0.53 ✔	-0.55 ✔	-0.17	0.04
pH			-	0.14	-0.37	-0.07	0.09	-0.07	-0.08	-0.10	-0.22
DO				-	-0.36	0.08	-0.24	-0.25	-0.69* ✔	0.01	-0.05
BOD					-	-0.20	0.04	-0.08	0.43	0.44	0.25
Alkalinity						-	-0.05	0.39	-0.56 ✔	0.36	0.43
Chloride							-	0.13	0.48	0.52 ✔	0.37
TDS								-	-0.20	0.11	0.09
TH									-	-0.005	0.06
Nitrate										-	0.58* ✔
Phosphate											-

*Significant at $P < 0.05$ level

This relationship between nutrients is considerably more reliable than basing the relationship of nutrients on physical environment and time factor alone. Hence, the aim is to keep the set of nutrients predicted by regression analysis to the minimum. The effect of this fundamental set on secondary nutrients is carefully studied. Several instances of the concentrations at different times and circumstances are obtained, plotted and analyzed. These instances are called snapshots, they give information on quality of water in terms of nutrients. From these snapshots, multivariant statistical models are created and the algorithm acts upon these models to get correlation coeffecients for pairs of nutrients. Then these nutrients are grouped using clustering.

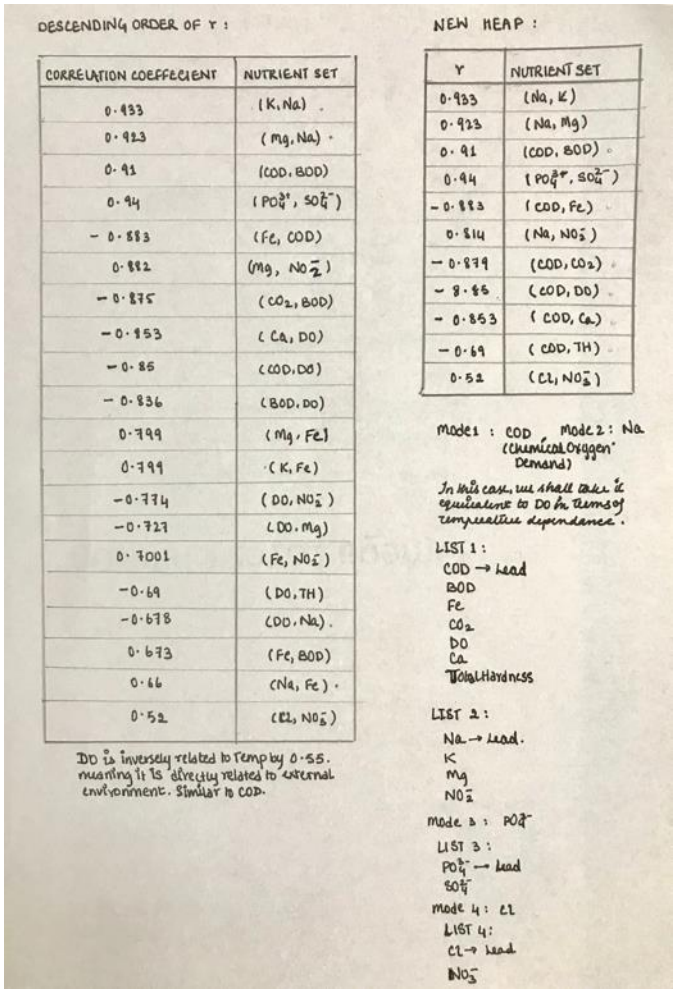
Examples of analytical indication would include the smooth inverse relation between concentration of dissolved oxygen and biological oxygen demand.

SNAPSHOTS AND CONCEPTS APPLIED TO THEM

The first step as mentioned before is to categorize secondary nutrients based on their indicator nutrient. This will be done using the following partial derivative of Djikstra's algorithm.

Algorithm:

1. For all positive selected correlation coefficients (r), group them in descending order of |r| along with the concerned nutrients.
2. Initialize the count of each nutrient to 0 and flag to untraversed (u)
   For all nutrient pairs, starting with highest coefficient;
   - If nutrient a. flag = u and nutrient b.flag = u
     - Push onto heap
     - a.flag = b.flag = traversed (t)
     - a.count + = 1
     - b.count + = 1
   - Else if nutrient a. flag = u xor nutrient b.flag = u
     - Consider if nutrient has already been traversed with nutrient c.
     - Create tuple ($r_{c,b} = r_{b,a} * r_{a,c}$, (nutrient c, nutrient a)) in favor of nutrient with Greater count.
     - Push tuple onto heap
     - b.flag = t
     - a.count + = 1
     - b.count + = 1
   - Else
     - Pass
3. Assess the heap and find mode of the total nutrient set m.
   This nutrient is an indicator nutrient. Push all tuples having m to another list and delete from the current heap.
   Repeat step 3 on the heap till heap is empty.
4. For every list created mark the indicator.



DESCENDING ORDER OF r :

CORRELATION COEFFECIENT	NUTRIENT SET
0·933	(K, Na)
0·923	(Mg, Na)
0·91	(COD, BOD)
0·94	(PO_4^{3-}, SO_4^{2-})
− 0·883	(Fe, COD)
0·882	(Mg, NO_3^-)
− 0·875	(CO_2, BOD)
− 0·853	(Ca, DO)
− 0·86	(COD, DO)
− 0·836	(BOD, DO)
0·799	(Mg, Fe)
0·799	(K, Fe)
− 0·774	(DO, NO_3^-)
− 0·727	(DO, Mg)
0·7001	(Fe, NO_3^-)
− 0·69	(DO, TH)
− 0·678	(DO, Na)
0·673	(Fe, BOD)
0·66	(Na, Fe)
0·52	(Cl, NO_3^-)

DO is inversely related to Temp by 0·55. meaning it is directly related to external environment. Similar to COD.

NEW HEAP :

r	NUTRIENT SET
0·933	(Na, K)
0·923	(Na, Mg)
0·91	(COD, BOD)
0·94	(PO_4^{3-}, SO_4^{2-})
− 0·883	(COD, Fe)
0·814	(Na, NO_3^-)
− 0·879	(COD, CO_2)
− 8·86	(COD, DO)
− 0·853	(COD, Ca)
− 0·69	(COD, TH)
0·52	(Cl, NO_3^-)

mode1 : COD (Chemical Oxygen Demand)

mode 2 : Na

In this case, we shall take it equivalent to DO in terms of temperature dependence.

LIST 1 :
COD → lead
BOD
Fe
CO_2
DO
Ca
Total Hardness

LIST 2 :
Na → lead.
K
Mg
NO_3^-

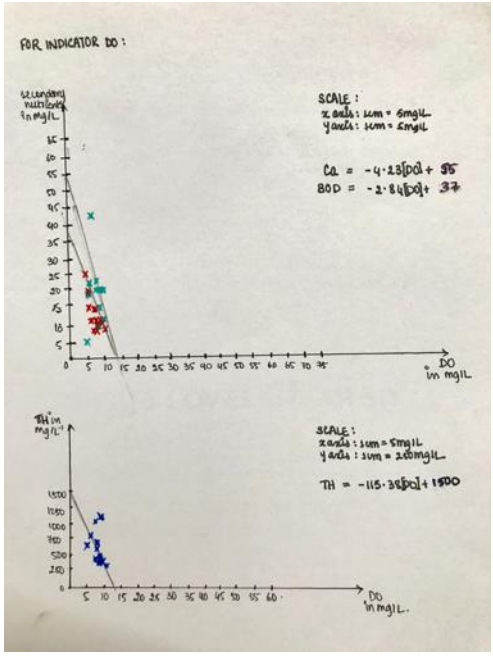
mode 3 : PO_4^{3-}

LIST 3 :
PO_4^{3-} → lead
SO_4^{2-}

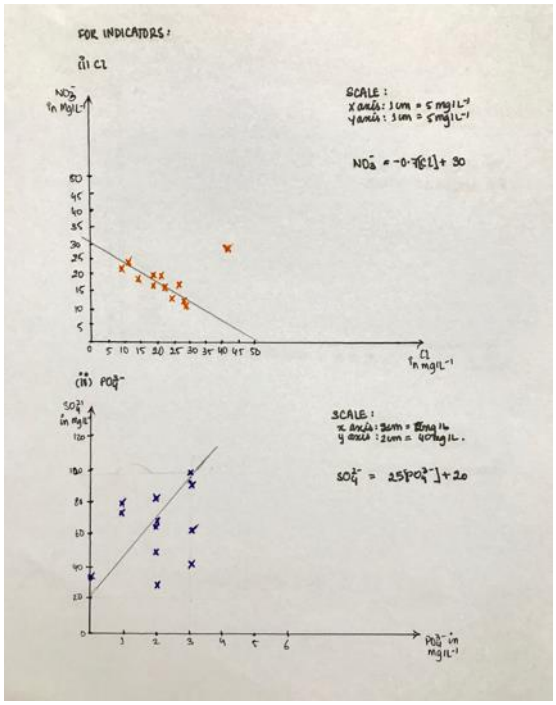
mode 4 : Cl

LIST 4 :
Cl → lead
NO_3^-

The next step involves sample data analysis with the use of snapshot algorithms based on variation of indicators which has now reduced our computation. Arrange the indicator samples in ascending order of concentration. Since the rest of chemicals have strong linear dependency, the graphs plotted, with the help of curve fitting, will give gradients of the dependency, that is, relative rate of change.

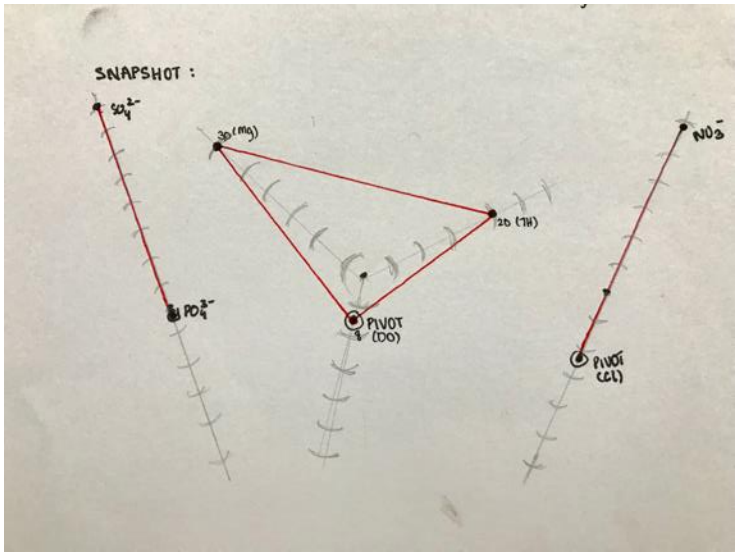


FOR INDICATOR DO :



The Snapshot Algorithm:

1. Obtain future values of indicator nutrients with the aid of regression analysis and plot the points on the set of concentric circles representing concentrations. Then join all points to form a polygon.
2. Name the indicator's vertex as pivot. Attach the rate relative gradient (b) of all the other vertices as an attribute. Perform this for all indicator graphs.
3. Now study the effect of a set of contaminants from an effluent on the indicator set of nutrients, since these are of smaller size, they may be handled reaction by reaction.
4. Deduce new values of indicator concentrations.
5. Push the pivots of each graph to new concentrations.
    a. Nutrient concentration = (pivot concentration)*b + C
    b. Move the vertices to the new locations on graph.



RESULTS AND PROOF OF CONCEPT

After sampling these results against a set of contaminants (Heptachlor, Dieldrin, Aldrin, Diazinon), that are periodically drained in the lake, the propagated values of the nutrients are:

1. Chloride = 120 ppm
2. Dissolved Oxygen = 8 ppm
3. Phosphate = 0.5 ppm
4. Sulphate = 32.5 ppm
5. Nitrate = 0.5 ppm
6. Calcium = 21.16 ppm

Comparing it with the table given below, it is proven that the method works with required accuracy.

STRENGTHS AND DRAWBACKS

1. This method of computation does not require in depth knowledge of chemical redox reactions that occur in water, nor does it need a reaction-by-reaction computation caliber. It is based on observations and data analysis, and the understanding of how to maximize use of strong dependences, eliminate weak dependences and return more accurate results.
2. The drawback is that several times perfect analytical indicators are not found. Also, there may always be a possibility of error as all predictions are subjected to, due to inclusion of outliers while calculating gradients, incorrect correlation coefficients and such misfortunes.

SCOPE IN ENVIRONMENTAL CHANGE

A slightly more accurate method can result in better planning of projects involving effluent release, while simultaneously ensuring their obligation to prescribed environmental water standards. This may be generated as a template and handle interrelated resources' chemical quality in a reinforced manner. The method may be applied to set of contaminants too, in such a manner that when multiple effluent loads are released, even though each one individually might tip the environmental standards, the net effect of the aggregated release is considerably harmless to the environment. This will also help project planners decide when to release which effluents and how to schedule their production year.

BIBLIOGRAPHY

1. D. Kamal, A.N. Khan, M.A. Rahman and F. Ahamed, 2007. Study on the Physicochemical Properties of Water of Mouri River, Khulna, Bangladesh, *Pakistan Journal of Biological Sciences, 10: 710-717.*
   DOI: 10.3923/pjbs.2007.710.717
   URL: https://scialert.net/abstract/?doi=pjbs.2007.710.717 for correlation coefficient data Udaisagar Lake, Udaipur.
2. http://www.moef.nic.in/sites/default/files/nlcp/P%20-%20World%20Case%20Studies/P-30.pdf for comparison with actual data.
3. http://www.environmentaljournal.org/1-3/ujert-1-3-8.pdf for correlation coeffecient data and training data sets of Udaisagar Lake, Udaipur.
4. https://www.researchgate.net/figure/Water-chemistry-of-Lake-Udaisagar-values-are-mean-n-24-AE1-SE_fig3_321575458 for training data sets of Udaisagar Lake, Udaipur.
5. Oxidation of Diazinon by Aqueous Chlorine: Kinetics, Mechanisms, and Product Studies
   Qi Zhang and and Simo O. Pehkonen*

Journal of Agricultural and Food Chemistry **1999** *47* (4), 1760-1766
DOI: 10.1021/jf981004e to study effect of contaminants on nutrients.