Testing water pollution in a two layer aquifer

MANUEL GARCIA LEON[1] and JUE LIN[2]

[1]Undergraduate Student in Civil Engineering - ETSECCPB, Spain garcia.leon.m@gmail.com
[2]Undergraduate Student in Civil Engineering - ETSECCPB, Spain

Water bodies around urban areas may be polluted with chemical elements from urban or industrial activities. We study the case of underground water pollution. This is a serious problem, since underground water is high qualified drinkable water in a world where this natural resource is increasingly reduced. This study is focused on a two-layer aquifer. If the superficial layer is contaminated, the deeper layer could be spoiled as well. This contribution checks the equality of the mean or centered composition of the two layers, with the aim of inferring their possible hydraulic conectivity.

The data to be examined are different hydro-chemical components of water, such as nitrates and nitrites (related to nitrate/nitrite poisoning of animal stock), tensoactives (toxic to the ecosystem) or potassium (it promotes eutrophization of the water), represented in mg/l. As the data are compositional, we can group the pertinent elements and compare them applying ilr transformation. The ilr transformation is used for simplicity when comparing compositional vectors. MANOVA (Multivariate Analysis of Variance) is applied on the transformed data from the two layers. This provides a hypothesis test to discern whether the two aquifer layers can be considered a homogeneous continuum or, on the contrary, they should be considered as isolated layers. An illustrative example is presented. Used data sets, being synthetic, are inspired by a real case. These analyses suggest that the two aquifers are connected.

1 Introduction

Major cities worldwide have experienced rapid expansion of its metropolitan regions. This transformation is resulting in some environmental problems that directly affect the underground water. The analysis of the concentration ratios of NO_2^-/NO_3^- and NH_4^+ in the aquifers is important, as nitrite and nitrate ingestion carry a serious illness: the metahemoglobinemia. This implies an increase of metahemoglobin (a oxidized hemoglobin) in blood, uncapable of fixing oxigen and to transport it to the body tissues. The European Directive state that the maximum concentration of nitrite permitted in water is 0.5 mg/L and the concentration of nitrate is 50 mg/L. In the other hand, other contaminants can also carry on problems. Tensoactives, used as dishwashers and clothwashers, are also toxic to human body. Nutrients such as potassium can promote eutrophization of the water, making it more turbid. There is a need for determining a procedure that allows engineers and scientists to provide checking techniques for drinkable water. In order to test for these water components and other ones, they can be treated as compositional data because as compositions we will ignore dilution effects. This can be used to tell the hydraulic dependence of two layers of an aquifer. In this paper we will see how we can explore this.

2 Sampling

Our case study is inspired on an aquifer in a coastal environment. It is surrounded by agricultural lands with many different other uses: farming, industry and camping. For this analysis two aquifers have been considered: a deep aquifer A and a superficial aquifer B. The question is to determine if they are interconnected. An evidence of such connection could be similar concentration ratios of certain elements in the water. Note that there must be a certain criticism concerning the nature of the hydro-chemical elements involved. They must come from out of the system analyzed, be conservative and not react with the soil. For instance, if the two layers of the aquifer are embedded in the same type of soil, they can exhibit similar water characteristics even if they are disconnected.

As a starting point, a hypothetical set of fixed sampling points are determinated, in order to cover the whole study domain (Figure 1). The sampling design is assumed to cover all the possible

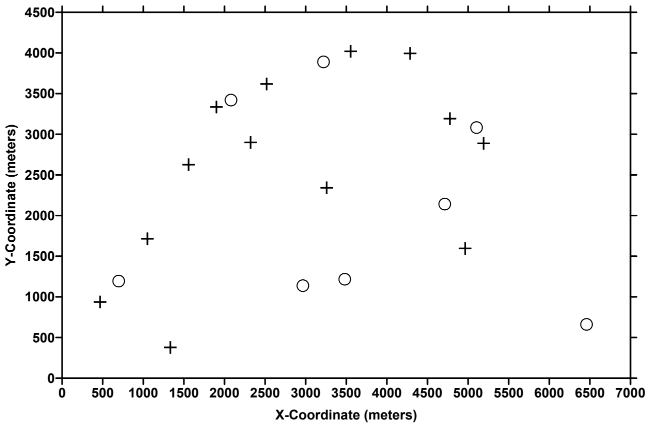

Figure 1: Distribution of the points where the water samples were supposed to be taken. Circle points are refered to aquifer A. Cross points are refered to aquifer B.

heterogeneities due to the morfology of the terrain. In this set of points integrated water samples are taken for a certain period of time (six years, 3 samples per well every year). An integrated sample is defined as the sample obtained in a given time instant but in several nearby sampling points. In this case study, the set of points consists of 8 points for the aquifer A (circle points in Figure 1) and 12 points for the aquifer B (cross points in Figure 1). The chemical elements mesured can be found in Table 1:

Component	Formula or abbreviation
Chlorine	Cl^-
Sodium	Na^+
Potassium	K
Calcium	Ca
Magnesium	Mg
Carbonate	$CO_3^=$
Bicarbonate	HCO_3^-
Sulfate	SO_4^-
Tensoactives	TA
Ammonium	NH_4^+
Nitrite	NO_2^-
Nitrate	NO_3^-
Phosphate	PO_4^{3-}
Hydrogen ion	H_3O^+

Table 1: Chemical components present in the water and tested for their concentrations.

3 Data Analysis

Since $CO_3^=$ and NO_2^- have more than 10% of zeroes, it has been decided to not include them to avoid problems with log-ratios. This is reasonable because:

- The concentration of $CO_3^=$ can be low because of a high pH, which has made the equilibrium prone to HCO_3^-. If two aquifers have the same concentration of HCO_3^-, then they will have the same concentration of $CO_3^=$ given the same pH conditions. In other words, HCO_3^- and pH may be taken as surrogates of $CO_3^=$.
- If the redox conditions, and concentrations of NH_4^+ and NO_2^- are known, the concentration of NO_3^- in the water can be predicted. Certain microorganisms, present in the water, carry on the oxidation of NH_4^+ to NO_2^- to NO_3^-. Thus we can safely discard this variable for our goals.

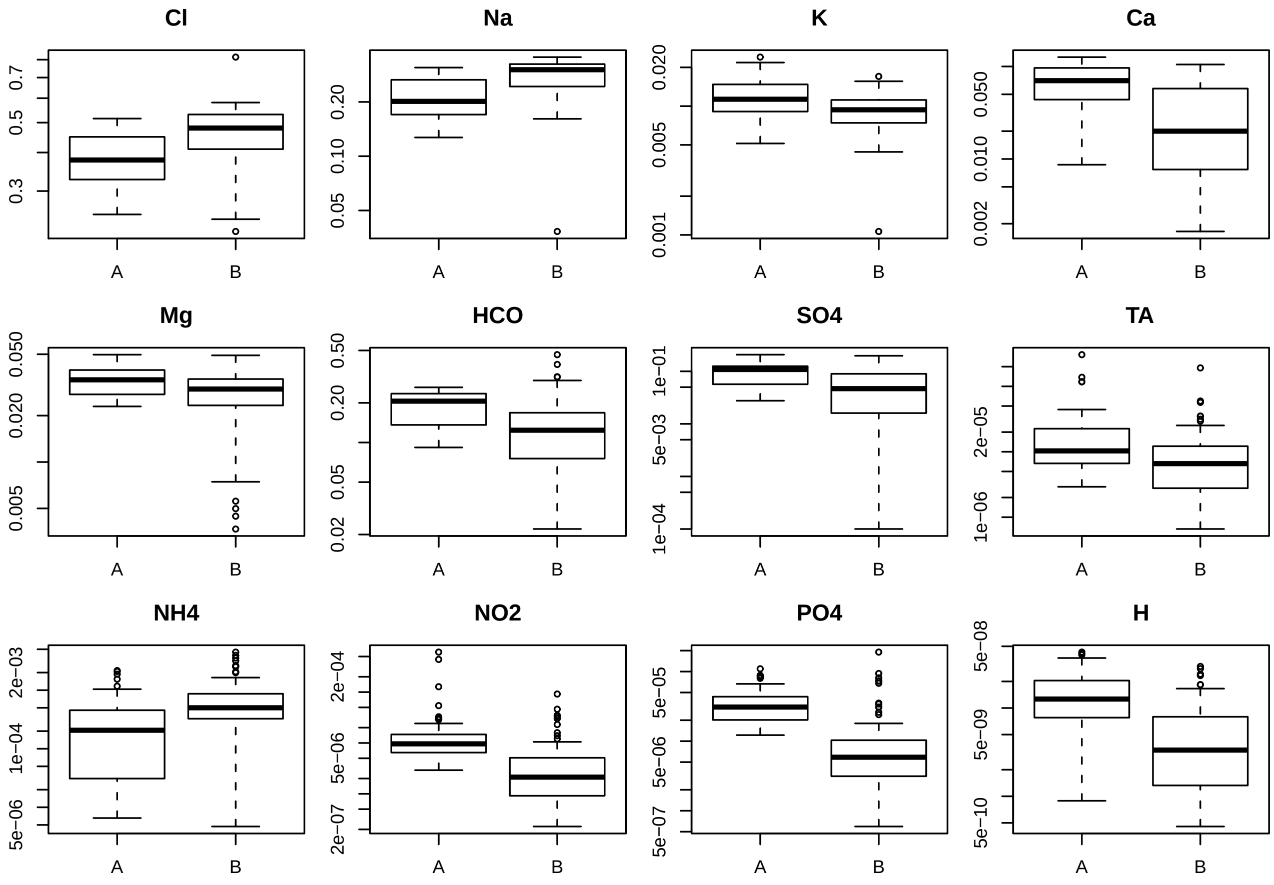

Figure 2: Boxplot of all elements and for aquifers A and B. Logarithmic scale is used for the concentrations.

After removing these elements, data have been purged. Rows where zeroes appeared have been removed, as well as rows with data that are obvious outliers. Special attention has been paid to aquifer B, and to some elements of the aquifer A (e.g. ammonium), since they seem to have a large dispersion (Fig. 2).

The water components are considered compositional (Buccianti et al., 2006). Data is converted into compositions with total equal to 1 and boxplots are used to display their differences between aquifers. Figure 2 shows that some elements are likely to have different behaviour in both aquifers.

The data are converted into clr coefficients and a biplot analysis is applied. Figure 3 represents the 72.6% of the total variance. From the biplot, it is evident that the two aquifers are quite different, although there is a certain overlap in the 3rd quadrant. The elements that discriminate mostly are NH_4^+ and the complex Cl^-, Na^+ and Mg. There may be also two clusters for the aquifer A and, three, for the aquifer B. This has not been explored.

If we only consider elements with the following properties:

- The elements involved come from out of the analyzed system
- They are not related to soil
- And they are expected to be mostly conservative, thus not react strongly within the framework with other dissolved species.

We can work with Cl^-, Na^+, TA, and PO_4^{3-}. In order to apply MANOVA on the data, a sequential binary partition table is constructed to convert the compositions into ilr coordinates (Egozcue and Pawlowsky-Glahn, 2005). In reference to the partitions (Table 2, Figure 4):

1. TA is separated from the other elements, as dishwashers and clothwashers involve more complex molecules, such as perfumes.

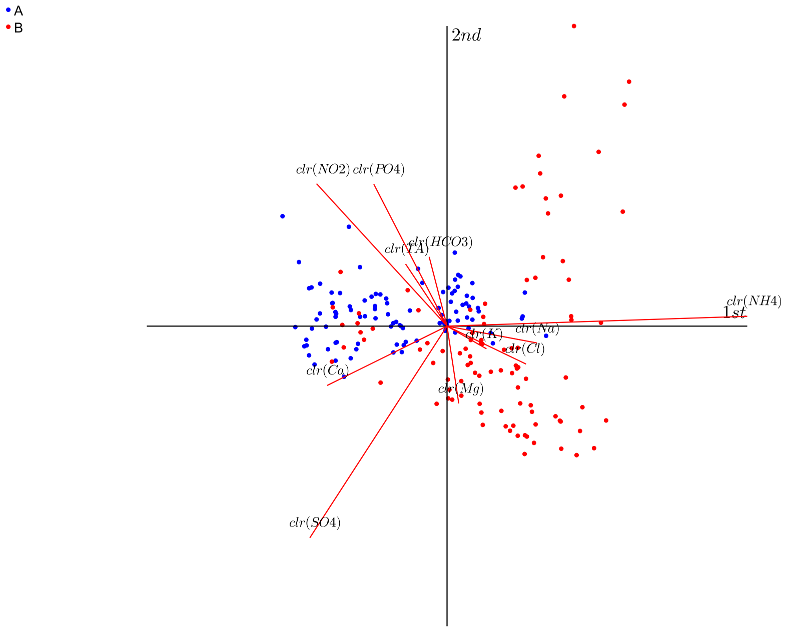

Figure 3: Biplot for the two aquifers. Blue is for aquifer A. Red is for aquifer B.

2. PO_4^{3-} is separated from the $ClNa$ complex.

3. Cl^- is separated from Na^+.

$balance$	Cl	Na	TA	$PO4$	r	s
A	1	1	-1	1	3	1
B	1	1	0	-1	2	1
C	1	-1	0	0	1	1

Table 2: Partition table

With this basis, the dendrogram of Figure 4 shows:

1. There are the same $Cl^-{*}Na^+{*}PO_4^{3-}/TA$ ratios. Althogh with high dispersity for both aquifers.

2. There are the same $Cl^-{*}Na^+/PO_4^{3-}$ ratios. Although with high dispercity for aquifer B.

3. There are the same Cl^-/Na^+ ratios, and with almost none dispercity.

These are evidences of contaminant transport between the aquifers.

4 Introduction to MANOVA

The main goal of MANOVA (Raykov and Marcoulides, 2008; Rencher, 2002) is the examination of mean differences across several groups when more than one dependent variables (DVs) are considered simultaneously. That is, MANOVA is essentially an analysis of variance (ANOVA) with $p \geq 1$ response (dependent) variables.

The assumptions of MANOVA are:

1. The observations are independent.

2. The population variance matrices for the p dependent variables are equal.

3. The mean observations on the dependent variables follow a multivariate normal distribution in each group.

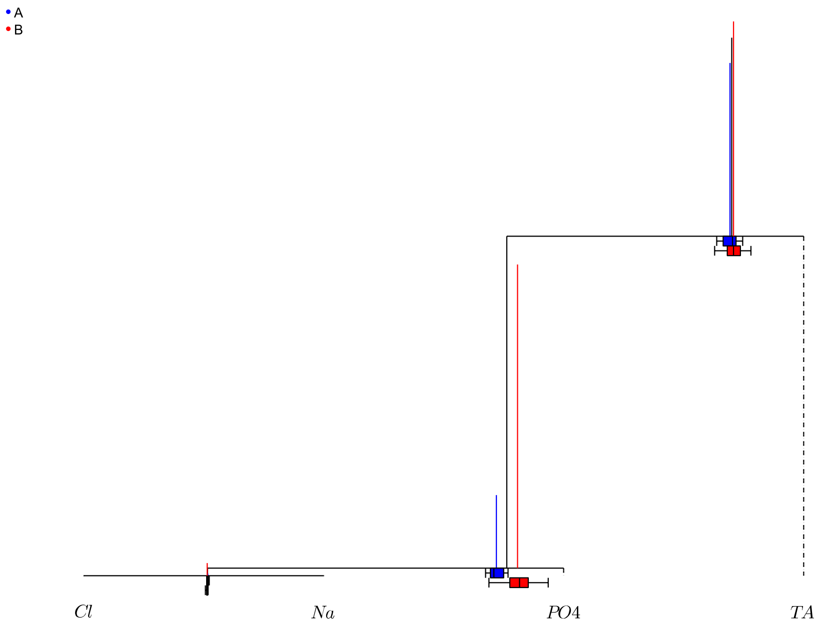

Figure 4: Dendrograms for Cl^-, Na^+, TA and PO_4^{3-} . Blue is for aquifer A and red is for aquifer B.

The independence and the equal variance are necessary to obtain an F-test of known degrees of freedom. In this paper, the following conventions will be used: D is the number of parts, i.e. D=11 chemical elements which are analised with the water samples. The dimension is D-1, ν_H are the degrees of freedom for the hypothesis, ν_E are the degrees of freedom for the error, and α is the type I error rate.

The mean vectors of the k samples are to be compared for significant differences. The hypothesis is, thus, $\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_k$ vs. $\mathcal{H}_1$: at least two $\boldsymbol{\mu}$'s are unequal.

Two matrixes are defined: **H** and **E**. (Rencher, 2002).

$$\mathbf{H} = n\sum_{i=1}^{k}\left(\overline{\mathbf{y_i}} - \overline{\mathbf{y}}_{..}\right)\left(\overline{\mathbf{y_i}} - \overline{\mathbf{y}}_{..}\right)^{'} = \sum_{i=1}^{k}\frac{1}{n}\mathbf{y}_i\mathbf{y}_i^{'} - \frac{1}{kn}\mathbf{y}_i\mathbf{y}_i^{'} \tag{1}$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_{\mathbf{i}}\right)\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_{\mathbf{i}*}\right)^{'} = \sum_{ij}\mathbf{y}_{ij}\mathbf{y}_{ij}^{'} - \sum_{i}\frac{1}{n}\mathbf{y}_i\mathbf{y}_{i*}^{'} \tag{2}$$

Note that **y** represents ilr-transformed compositions. The $p \times p$ matrix **H** (Eq. 1) is called "hypothesis" matrix. If it is assumed that no linear dependencies exists in the variables, the rank of **H** is the smallest of p and ν_H, min(p,ν_H). Therefore, **H** can be singular. On the other hand, the $p \times p$ "error" matrix **E** (Eq. 2) has a within group sum of squares for each variable on the diagonal, with analogous sums of cross products off diagonal. The rank of **E** is min(p, ν_E), usually p.

There are many statistic tests in the literature. The most common tests are Hotelling-Lawley and Wilks. When $\mathcal{H}_0$ is true, α is the same one to reject it. However, when $\mathcal{H}_0$ is false, they have different probabilities of rejecting it. R uses an approximation of the F distribution for the statistics.

Lawley-Hotelling. The Lawley-Hotelling statistic is defined as:

$$U^{(s)} = tr\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{s}\lambda_i \tag{3}$$

$\mathcal{H}_0$ is rejected for large values of the test statistic. The tolerance for the value of the test statistic is previously chosen, in each test. For our tests, we choose a tolerance of α =0.05.

Wilks. In this case, the likelihood ratio test is:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \tag{4}$$

Wilks' Λ compares the within sum of squares and "error matrix" $\mathbf{E}$ to the total sum of squares and products matrix $\mathbf{E}+\mathbf{H}$ (in a similar way of the univariate F-statistic). $\mathcal{H}_0$ is rejected if $\Lambda \leq \Lambda_{\alpha,p,\nu_H,\nu_E}$.

The asymptotic distribution of both statistics is the Chi-square distribution (Gupta et al., 2008).

First of all, MANOVA tests are carried out comparing all 3 ilr components for the two aquifers. The ilr components can be considered independent and the variances must be checked for hypothesis test on means. On the other hand, to obtain normal distributions, bootstrap must be applied to the data. Bootstrap simulates as much data as needed to make the whole sample to assimilate a normal distribution.The individual hypothesis test and the bootstrap are not applied in this case. Thus, all three hypotheses are considered to be true, *a priori*. Secondly, ANOVA test is carried out for each of the components. Third, MANOVA is carried out on the ilr components in groups of two.

5 Results and discussions

If we see Chart 3, we can see that the $Cl^{-}*Na^{+}/PO_4^{3-}$ ratio is the same for both aquifers. This means that we have the same $Cl^{-}*Na^{+}$ concentration and/or PO_4^{3-} concentration. Thus, there is marine water pollution flowing upside down or bottom-up. In the other hand, there could also be PO_4^{3-} transport.

balance	*variables*	$p-value$
A	Cl*Na*PO4/TA	5.282e-10
B	Cl*Na/PO4	0.2817
C	Cl/Na	0.01399
A-B-C		2.478e-09
A-B		4.381e-09
B-C		0.02451
A-C		4.382e-10

Table 3: Results of the MANOVA tests

6 Conclusions

We test for Cl^{-}, Na^{+}, TA and PO_4^{3-}. Chart 3 shows that we have the same $Cl^{-}*Na^{+}$ concentration and/or PO_4^{3-} concentration. Thus, either there is marine water pollution flowing upside down or bottom-up or there is PO_4^{3-} transport. We can conclude that the two aquifers are probably connected.

7 Acknowledgements

This article has been write under the supervision of Prof. Dr. J.J. Egozcue, Ms. Maribel Ortego, Dr. Vera Pawlosky-Glahn and Dr. Raimon Tolosana-Delgado. The authors would like to thank them all for their constant dedication, patience and cheers when we were taking our first steps in the field of the compositional data analysis. We would also thank them for providing us the possibility of participating in the CodaWork 2011.

References

Buccianti, A., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2006). *Compositional Data Analysis: from theory to practice.* Number 264 in Special Publications. London, UK: The Geological Society.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 795–828.

Gupta, A., S. Harrar, and Y. Fujikoshi (2008). Manova for large hypothesis degrees of freedom under non-normality. *TEST 17*, 120–137.

Raykov, T. and G. A. Marcoulides (2008). *An Introduction to Applied Multivariate Analysis.* Routledge. LLC, New York: Taylor & Francis Group.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd edition ed.). Wiley. New York: John Wiley & Sons.