Study on the effect of temperature on the leaching of contaminants from Ballangen Tailings Deposit, Norway-A statistical method

F Yuan[1,2] and J Lu[1]

[1]Department of Engineering and Safety, UiT The Arctic University of Norway, N-9037 Tromsø, Norway

E-mail: yuan.fuqing@uit.no

Abstract. Temperature is one important factor that affects the leaching of contaminants from tailings deposit. However, randomness issue due to the precision of the instrument, the variation of ambient temperature, the individual skill of the lab technician etc., is inevitable and will affect the results when we evaluate the dependence of two parameters. In this study, a statistical method was developed to analyze the experimental data to reduce randomness. The experiment data from a laboratory batch leaching experiment on the tailings at Ballangen deposit, Northern Norway were used to analyze the dependency of concentrations of Cr, Cu, Ni and Zn in the leachate on temperature. The dependency of concentrations of Cr, Cu, Ni and Zn in the leachate is determined by testing the significance of the slope coefficient of regression analysis. The results show that the concentrations of Cr and Zn do not show dependency on the temperature. However, the temperature shows evident impact on the concentrations of Cu and Ni. High temperature will produce higher Cu and Ni concentration in the leachate.

1. Introduction

The extraction of metals and minerals can affect the natural environment to a significant extent both while active and after operations have ceased. Large amounts of waste rock and tailings are produced during mining operation. Proper handling of mine wastes is the most significant environmental issue associated with mining.

The leaching of contaminants from waste deposit will last for hundreds or even thousands of years, and degrade the environment significantly [1]. The contaminants will further transport through the environmental medium to the ecosystem and the human beings living in the area in the long-term [2,3]. Leaching of contaminants from waste deposit is affected by many factors, of which temperature is an important parameter [4].

Laboratory batch leaching experiment can be conducted to assess the effect of temperature on the leaching of contaminants. However, due to the precision of the instrument, the variation of ambient temperature, the individual skill of the lab technician etc., the resultant metal concentration shows variance from each other, even the values from the same sample. Randomness issue is inevitable. In the analysis, the randomness should be considered to avoid wrong decision-making. As randomness is not possible to eradicate, it is eminent to control the randomness to some acceptable levels. Statistical method should be introduced to measure and reduce the randomness.

In this study, a statistical method was developed to analyze the experimental data to reduce randomness. The experimental data were from a laboratory batch leaching experiment on the tailings at Ballangen deposit, Northern Norway [4]. The experiment was run at four different leaching temperatures and aimed at assessing the impact of temperature on the leaching of contaminants from waste deposit. The dependency of concentrations of Cr, Cu, Ni and Zn in the leachate on the leaching temperatures is analyzed by the developed statistical method.

2. Experiment design

In the experiment, the leaching test is carried out at four different temperature levels: 5℃, 10℃, 15℃, 20℃. 10 g of tailings was put into a 50ml centrifuge tube and added 20ml of deionized water. The mixture was set into four incubators that were set at 5℃, 10℃, 15℃, 20℃ respectively. The supernatant was collected from the top after 48 hours of mixing. A new identical 20 ml of deionized water were added again to the centrifuge tube. The procedure is repeated for six leaching cycles. The collected leachate were tested the concentrations of Cr, Cu, Ni and Zn. The detailed experiment procedure can be found in our another publication [4]. The four temperature levels are selected to test the dependency of the heavy metal concentrations on the temperature.

3. Analyzing method

Correlation analysis can assess the dependencies between two factors [5,6]. Correlation coefficient can be an indicator of the dependency. However, it cannot elaborate the dependency in detail. Regression analysis is an alternative to evaluate the dependency of two factors. One factor is designated as response variable. The other is as predictive variable. Regression can tell the numerical dependency among factors. The regression coefficient of predictive variable in the regression function is an indicator of the dependency. For given observed data, big coefficient value implies dependency that is more significant. However, it is not comparable among coefficients estimated for different predictive variable. High magnitude of coefficient could correspond to situations of little dependence or no dependence. Approach should be developed to address the significance of the dependence. The regression analysis is a basic data analysis method. This paper omits the discussion on regular regression issues, but only addresses on issues special for this paper.

3.1. Experimental data fusion

The experiment contains 4 unique temperature levels. For each temperature level, if analyzing the data using regression analysis, the regression coefficient is evaluated from 4 data sets. High uncertainty contains in the coefficient. As increasing the data size is costly, we can fuse test data of each leaching cycle to increase data size.

Most heavy metals' concentration depends on the leaching test cycles. But the explicit dependency expression is unknown. Let the metal concentration denoted by y, the leaching cycles denoted by s, temperature denoted by x. It assumes the concentration y depends on the leaching cycles as in formula (1).

$$y = bf(x) + f(s) \tag{1}$$

The concentration depends on the temperature in a function of $bf(x)$ and depends on the leaching cycles in a way as $f(s)$. The leaching cycles contribute to the concentration additively.

Table 1. Statistical model of each observation.

i \ j	j=1	…	j=j	…	j=m
i=1	$bf(x_1) + f(s_1) + \epsilon_{11}$	…	$bf(x_1) + f(s_j) + \epsilon_{12}$	…	$bf(x_1) + f(s_m) + \epsilon_{1m}$
i=2	$bf(x_2) + f(s_1) + \epsilon_{21}$	…	$bf(x_2) + f(s_j) + \epsilon_{22}$	…	$bf(x_2) + f(s_m) + \epsilon_{2m}$
…	…	…	…	…	…
i=n	$bf(x_n) + f(s_1) + \epsilon_{n1}$	…	$bf(x_n) + f(s_j) + \epsilon_{n2}$	…	$bf(x_n) + f(s_m) + \epsilon_{nm}$

Suppose the observed concentration is $y_{1j}, y_{2j}, \dots, y_{ij}, \dots y_{nj}$ at leaching cycle s_j. The n is number of temperature level. The m is number of leaching cycles. The statistical model for each observation is shown in table 1.

The ϵ_{ij} is residual follow same normal distribution with parameters $(0, \sigma^2)$. As each experiment measurement is independent, the ϵ_{ij} is independent from each other. To remove the effects of f(s), we can transform the data by letting $y_{ij}' = y_{ij} - \frac{1}{n}\sum_{i=1}^{i=n} y_{ij}$. The dependency of y on the temperature will not change after the transformation, i.e. the coefficient b will not change. The mathematical proof of it can refer to Appendix A. The experimental data of each leaching cycles can thus be fused to enlarge the sample size. In the rest of paper, the y_{ij} refers to the transformed y_{ij}' for simplicity.

3.2. Test on the dependence significance

The dependency of concentration on temperature is described as a regression formula (2).

$$y = bf(x) + a + \varepsilon \tag{2}$$

The null hypothesis and alternative hypothesis are

$$\begin{aligned} H_0&: b = 0; \\ H_1&: b \neq 0; \end{aligned} \tag{3}$$

A statistic can be established [7].

$$\frac{SS_{b=0} - SS_{b\neq0}}{SS_{b\neq0}/n - 2} = F \tag{4}$$

The F has an F-distribution with 1 and $mn - 2$ degree of freedom. The $SS_{b\neq0}$ is the squared residual of H_1. The $SS_{b=0}$ is the squared residual of H_0.

$$SS_{b\neq0} = \sum_{j=1}^{j=m} \sum_{i=1}^{i=n} (\widehat{y_{ij}} - y_{ij})^2 \tag{5}$$

The $\widehat{y_{ij}}$ is the predicted value of regression. Sufficient large F value will lead to the rejection of H_0. In the regression analysis, when $b = 0$, there is no predictive variable. The unknown regression parameter is the constant. Applying the least square method, the corresponding mean square error is

$$MSE = \sum_{j=1}^{j=m} \sum_{i=1}^{i=n} (a - y_{ij})^2 \tag{6}$$

The minimal MSE is obtained at

$$\hat{a} = \frac{\sum_{j=1}^{j=m} \sum_{i=1}^{i=n} y_{ij}}{mn} \tag{7}$$

The proof of equation (5) is straightforward. Therefore

$$SS_{b=0} = \sum_{j=1}^{j=m} \sum_{i=1}^{i=n} (\hat{a} - y_{ij})^2 \tag{8}$$

The fusion of the data can improve the reliability of the hypothesis test. When no fused data considered, the test statistics F is with freedom (1, n). When fusion considered, the freedom becomes (1, mn). The F distribution with freedom (1, mn) is more concentrated than the (1, n), i.e. the statistics made on (1, mn) F distribution is higher reliable than the (1, n). The dependency relies on the coefficient b in (2). The significance level implied the dependency of the concentration on the temperature. When H_0 is accepted, the dependency is not significant.

3.3. *Model assumption validation*

The original experimental data has been converted to a regression with replications. The availability of the replication facilitates the model assumption validation, i.e, verifying the fitness of model (1) to the experimental data. The data representation with repeated data is shown in table 2.

Table 2. Experiment data with replication.

x \ y	Rep 1	Rep 1	Rep 1	…	Rep m
x_1	y_{11}	y_{12}	y_{13}	…	y_{1m}
x_2	y_{21}	y_{22}	y_{23}	…	y_{2m}
…	…	…	…	…	…
x_n	y_{n1}	y_{n2}	y_{n3}	…	y_{nm}

A hypothesis test can be proposed to test the fitness of data in table 2 to model (1). The null hypothesis and alternative hypothesis are

$$H_0:\ Y = \hat{b}X; \qquad H_1: Y \neq \hat{b}X. \tag{9}$$

The fitness test is named lack of fit test in statistics [8-11]. A statistics based on the F-distribution is proposed [12].

$$F_{LOF} = \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=m}(\bar{y_i} - \hat{y_i})^2}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=m}(y_{ij} - \bar{y_i})^2} \sim F(n-2, nm-n) \tag{10}$$

The $\hat{y_i}$ is the predicted value from regression. $\bar{y_i}$ is the average value row vector in table 2. Higher value leads to accept H_1 in equation (9), i.e. the model (1) does not fit the data properly.

3.4. *Variance reduction by data fusion*

The aim of fusing data is to increase the data size. As shown in Section 3.2, the performance of hypothesis test will be improved using fused data. This section proves the coefficient is unbiased and the variance of the regression coefficient reduced significantly. An unbiased estimate is the desired property for an estimation [13,14].

For a situation without replication, an estimates of slop b for least square estimation is [15,16].

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

When replication presents, the equation (11) can be rewritten as

$$\widehat{b_r} = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})\sum_{j=1}^{m}(y_{ij} - \bar{y})\right)}{m\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

where the

$$\bar{y} = \frac{1}{nm}\sum_{i=1}^{i=n}\sum_{j=1}^{j=m} y_{ij} \tag{13}$$

Let $\bar{y_i} = \frac{1}{m}\sum_{j=1}^{j=m} y_{ij}$ and $D_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. The $\widehat{b_r}$ can be rewritten as

$$\begin{aligned}\widehat{b_r} &= D_1\frac{\sum_{j=1}^{m}(y_{ij} - \bar{y})}{m} + D_2\frac{\sum_{j=1}^{m}(y_{ij} - \bar{y})}{m} + \cdots + D_n\frac{\sum_{j=1}^{m}(y_{nj} - \bar{y})}{m} \\ &= D_1(\overline{y_1} - \bar{y}) + D_2(\overline{y_2} - \bar{y}) + \cdots + D_n(\overline{y_n} - \bar{y}\end{aligned} \tag{14}$$

Since $\sum_{i=1}^{n} D_i = 0$, the equation (14) can be rewritten as

$$\widehat{b_r} = D_1\overline{y_1} + D_2\overline{y_2} + \cdots + D_n\overline{y_n} \tag{15}$$

As the observations are independent from each other, the $\overline{y_i}$ for $i = 1,2,..,n$ is then independent from each other. It is straightforward to conclude

$$\overline{y_i} \sim N(a + bf(T_i), \frac{\sigma^2}{m}) \tag{16}$$

where the expectation of $\overline{y_i}$ is $E(\overline{y_i}) = a + bf(T_i)$.

$$E(\widehat{b_r}) = \sum_{i=1}^{i=n} \frac{(x_i - \bar{x})(a + bx_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

It can be further rewritten as

$$E(\widehat{b_r}) = a\sum_{i=1}^{i=n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + b\sum_{i=1}^{i=n} \frac{x_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{18}$$

It is straightforward to find $\sum_{i=1}^{i=n} \frac{x_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1$ and $E(\widehat{b_r}) = b$. The least square estimator for the b is unbiased. The variance of $\widehat{b_r}$ is

$$Var(\widehat{b_r}) = \sum_{i=1}^{i=n} {D_i}^2 Var(\overline{y_i}) = \frac{1}{m} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{19}$$

The variance of $\hat{b}$ without replicated observation is

$$Var(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{20}$$

Therefore,

$$Var(\widehat{b_r}) = \frac{Var(\hat{b})}{m} \tag{21}$$

The introduction of replication can significantly reduce the randomness of the estimates.

3.5. Simulation study

The best way to test the performance of the dependency methods is to conduct more experiments. However, economically, more experiments are normally associated with the high cost and therefore are not practical in reality sometimes. A feasible way is to use simulation method. Experiment can be simulated using software such as Matlab. In this simulation, pseudo random values are drawn from the model

$$y = a + bx + \varepsilon \tag{22}$$

where ε following normal distribution with parameter (0,1). The intercepts are supposed to be a=0, 1, 2, 3, 4, 5 respectively. The α corresponds to f(s) in (1). From model (22), for each intercept, four random values are drawn. The x is selected as 5, 10, 15, 20, i.e. the m=6 and n= 4 corresponding to table 2. The intercepts are chosen as b=0, 0.01, 0.1, 0.5, 1. Total 5000 runs are conducted for each b. Applying the trend test approach proposed in Section 3.2, for replicated situation, the results are shown in table 3.

Table 3. Accuracy of replicated case.

	b=0	b=0.01	b=0.1	b=0.5	b=1
Accuracy	0.91	0.09	0.82	1	1

When non-replication is considered, only four observations are present. The accuracy is shown in table 4.

Table 4. Accuracy of non-replicated case.

	b=0	b=0.01	b=0.1	b=0.5	b=1
Accuracy	0.95	0.10	0.70	0.79	0.997

By comparing tables 3 and 4, it reveals the accuracy depends on the slope value b. When replications are present and the b is at a magnitude of 0.5 or above, the test method can achieve very high accuracy of almost 100%. When b is small at a magnitude of 0.1, for the replicated situation and non-replicated situation, it is only 0.82 and 0.7 respectively. When no trend presents, i.e. b=0, the accuracy is 0.91 for replicated situation. Surprisingly, the case with non-replication can even show higher accuracy than replicated situation when b=0. The improvement from the data fusion is obvious. Moreover, the simulation considers a situation with very few observations presented. The proposed method can have excellent performance when the slop b is very significant. When the slop b in the regression is near zero, the test performance is relatively weak, but it is still at acceptable level of accuracy. In the experimental data analysis, it is necessary to fuse experimental data.

4. Experimental data analysis

The concentrations of Cr, Cu, Ni, Zn in the leachate are measured at each leaching cycle. In order to investigate the dependency, several functions f(T) are defined, considering both the linear and nonlinear situations. Figure 1 plot the concentrations of Cu and Zn against temperature separately. Each connected line corresponds a leaching test cycle in the figure. For both Cu and Zn, the original data doesn’t exhibit evident trend, neither increasing nor deceasing. The Zn concentration against temperature is evidently random. After fusing the six leaching cycle data, as shown in figure 2, the Zn is still evidently random. The Cu shows slightly increasing trend. The visual checking of the trend is subjective. The test method proposed in Section 3 is used to assess the dependency.

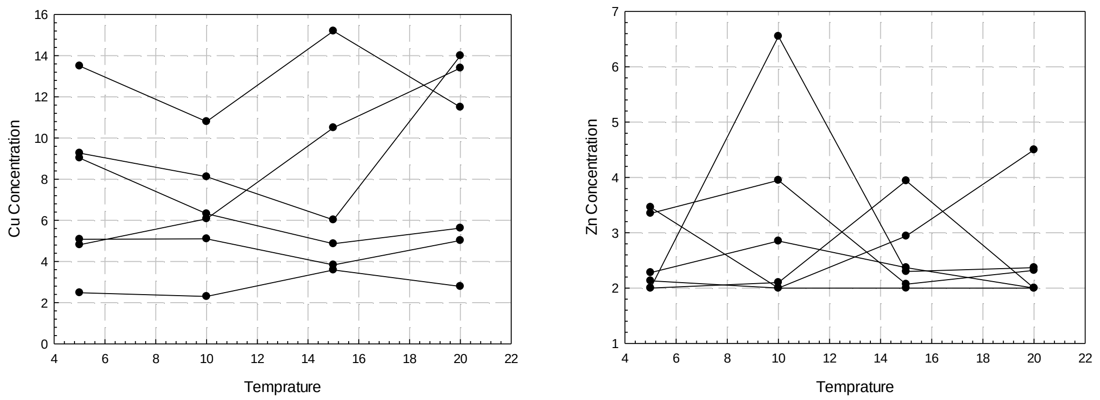
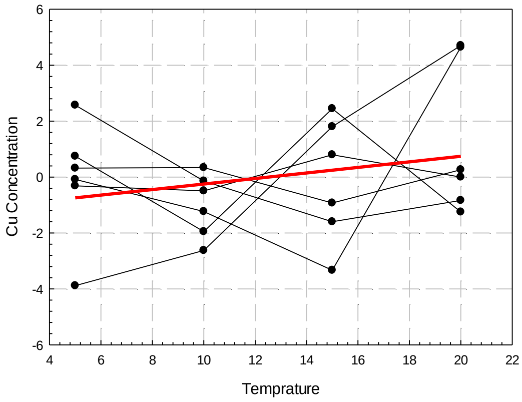


Figure 1. Concentration Cu (Left) and Zn (Right) concentration vs temperature.



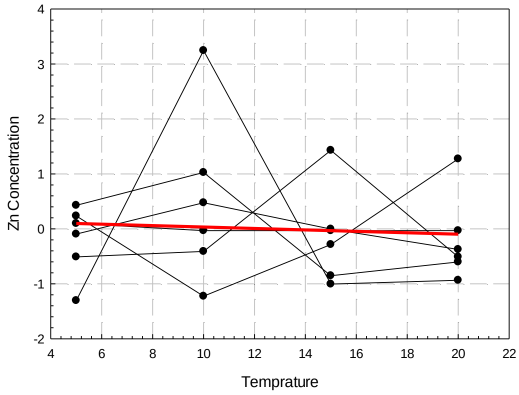


Figure 2. Transformed Cu (Left) and Zn (Right) concentration.

The number of sample m is 6 and the m is 4 in the example for the statistical model. As the explicit expression of the dependency of concentration on temperature is unclear, this paper tried the f(T) in a form of T, $T^{0.5}$, T^2 , T^3, T^4 and *log(T)*. Tables 5 and 6 show the hypothesis test results for all the f(T). The "P values" for slope show the degree of the dependency. High P-value implies high dependency. "Slope" shows the regression coefficient b of (1). "LOF" describes the fitness of model to the data. The p value of "LOF" show the function can fit the data or not. Higher LOF p value implies higher possibility not fitting the data. "Dependency" describes if the concentration depends on the temperature in a form of f(T). The higher p value implies higher possibility of dependency. The classical statistics choose the confidence level at least 0.95. In this case, the p value should at least 0.95 to claim the dependency is evident.

Table 5. Test on the concentration of Cr and Cu.

	Test on the concentration of Cr				Test on the concentration of Cu			
f(T)	P Value	Slope	LOF	Dependency	P Value	Slope	LOF	Dependency
T	0.4130	-0.0203	0.2483	No	0.7884	0.0992	0.1437	Yes
$T^{0.5}$	*0.5186*	*-0.1758*	*0.9956*	*x*	*0.7318*	*0.5922*	*0.9999*	*x*
T^2	0.21	-3.92e-4	0.2740	No	0.8577	0.0046	0.4227	Yes
T^3	0.060	-5.2e-6	0.2681	No	0.8891	0.000223	0.3974	Yes
T^4	0.0327	-1.4e-7	0.2678	No	0.9024	0.0000115	0.3698	Yes
log(T)	*0.6138*	*-0.3455*	*1*	*x*	0.6596	0.8194	1	*x*

Table 6. Test on the concentration of Ni and Zn and Ni.

	Test on the concentration of Ni				Test on the concentration of Zn			
f(T)	P Value	Slope	LOF	Dependency	P value	Slope	LOF	Dependency
T	0.6955	5.8	0.0558	Yes	0.2720	-0.0131	0.1476	No
$T^{0.5}$	*0.6576*	*36*	*0.9995*	*x*	0.1905	-0.0603	0.5717	No
T^2	0.7405	0.2502	0.2522	Yes	0.3871	-7.47e-4	0.1822	No
T^3	0.7559	0.012	0.2254	Yes	0.4436	-4.0324	0.1810	No
T^4	0.7560	5.79e-4	0.2024	Yes	0.4637	-2.0e-6	0.1764	No
log(T)	*0.6092*	*52*	*1*	*x*	0.0984	-0.0499	0.5077	No

For Element Cr, the slope for all the dependency functions f(T) are negative. It implies Cr concentration decreases with the increasing temperature. However, the dependency is not significant. The p values for the function f(T) is small. In the LOF test, the $T^{0.5}$ and *log(T)* are not able to fit the data, as the p values for the LOF is above 0.95 and is thus strongly significantly. However, for the fitted function f(T), the p-values of slope are very low. The data imply the Cr concentration is not dependent on the temperature. For element Cu, the slop or the dependency is positive. The concentration of Cu increases with temperature. The same as element Cr, in the LOF test, the $T^{0.5}$ and *log(T)* are not able to fit the data. The other function fit the data. Among the fitted model, the linear dependency f(T)=T fit best, as it has the lowest p value. However, the p value for the f(T)=T is 0.7884. The dependence of concentration of Cu on the temperature is not significant.

For element Ni, the slope is positive. The concentration of Ni increases with temperature. The same as element Cr and Cu, in the LOF test, the $T^{0.5}$ and *log(T)* are not able to fit the data. The other models fit the data. The linear function f(T)=T is the best. The fitting is significant, as the p value is as small as 0.0558. The confident level of the fitting is near 0.95. The p value for the linear dependent is 0.6955. The dependence is not significant. For element Zn, the slope is negative. Figure 2 shows the regression function (the red line). The concentration of Zn decreases with temperature. In the LOF test, all the models fit the data. For the significance of slope, the p values are all below 0.5. The dependence of concentration on the temperature is not significant.

The dependency of heavy metal concentration on the temperature is evaluated by testing the significance of the slope in the two dimensional regression analyses. For all the four elements Cr, Cu, Ni and Zn, from statistical perspective, the individual element's dependency on the temperature is not significant. However, it also cannot conclude the dependency is not significant. The decision based on the classical hypothesis test assuming a test is significant when it is with high probability such as 0.95 or above. The threshold of the decision with this level is too high. The decision based this threshold is biased, as it favor the hypothesis of concentration not dependent on the temperature. For example, for the case of Cu, as shown in figure 2, the data exhibit the concentration depends on the temperature to some extent. If the confidence interval defined as 0.95, the concentration will be claimed not dependent on the temperature, as the p value is $0.7884<0.95$. However, the 0.7884 essentially favors the concentration depends on temperature as the p value is over 0.5.

If we define the threshold as 0.5 so as the two hypotheses are equal. P value of slope below 0.5 is considered as concentration not dependent on the temperature; otherwise, dependent. For element Cr, as shown in table 5, all the order of nonlinear regression is below 0.5. The logarithm dependency is out of consideration as it failed in the fitness test. The concentration of Cr is not dependent on the temperature. For element Cu, as shown in table 6, the P values are above 0.5. The Cu is dependent on the temperature. For element Ni, as shown in 6, the P value is also above 0.5, the concentration is thus considered dependent on the temperature. The Ni concentration tends to increase with the temperature. For the element Zn, as shown in table 6, the P values are below 0.5. The concentration of Zn is not considered dependent on the temperature.

5. Conclusions

The dependency of the heavy metal concentration on temperature is essentially not random from physical and chemical perspective. However, the observed data is random due to the measurement error, the dissimilarity of the sample etc. This paper fuses the experimental data from several leaching test cycles to reduce the randomness to lower the possibility of wrong conclusion making. The dependency of concentrations of Cr, Cu, Ni and Zn in the leachate is determined by testing the significance of the slope coefficient of regression analysis. By analyzing the data, the concentration of Cr and Zn do not show dependency on the temperature. However, the temperature shows evident impact on the concentrations of Cu and Ni. High temperature will produce higher Cu and Ni concentration in the leachate. The dependency implies further that the Cu and Ni that are leached out from the waste would vary in a year, as the temperature varies in a year in the area of the sample collection. Especially under the global climate change context, the effect of temperature on the leaching of contaminants should be investigated further. Analyzing the data from a statistical perspective can reduce the risk of wrong decision making. Other than using the subjective and rough visual checking approach, the using of hypothesis test and data fusion is more objective and more convincing than methods used in some of the literature papers.

Appendix A

The concentration y depends on the leaching cycles in a form as $y_{ij} = bf(x_i) + f(s_j)$+$\epsilon_{ij}$, as shown in table 1. Suppose the $\epsilon_{ij} \in N(0, \sigma^2)$, i.e. all the residual follows a same Normal distribution $N(0, \sigma^2)$. For leaching test j, taking the average of all the temperature $i = 1,2, \ldots, n$, we transform the y_{ij} into

$$y_{ij}' = y_{ij} - \frac{\sum_{i=1}^{n} y_{ij}}{n} \tag{A.1}$$

Substituting the full expression of y_{ij} into the (A.1),

$$y_{ij}' = bf(x_i) + f(s_j) + \epsilon_{ij} - \frac{b\sum_{i=1}^{n} f(x_i) + nf\big(s_j\big) + \sum_{i=1}^{n} \varepsilon_{ij}}{n} \tag{A.2}$$

It can be reduced to

$$y_{ij}' = b\left[f(x_i) - \frac{\sum_{i=1}^{n} f(x_i)}{n}\right] + \epsilon_{ij} - \frac{\sum_{i=1}^{n} \varepsilon_{ij}}{n} \tag{A.3}$$

The b is the desired parameter. Obviously, the b in (A.3) is the same as b in the (1). The transformation of the data does not change the b, but it can remove the unexpected influence of $f(s_j)$. The entry $\epsilon_{ij} - \frac{\sum_{i=1}^{n} \varepsilon_{ij}}{n}$ is a linear combination of random variable ϵ_{ij}. Each variable follows Normal distribution $(0, \sigma^2)$. The linear combination of Normal distribution variable is still Normal distribution [17]. The parameter of the new Normal distribution is $\epsilon_{ij}' \sim (0, (1 - \frac{1}{n})\sigma^2)$. The derivation of the variance $(1 - \frac{1}{n})\sigma^2$ is omitted. Conclusively, after transformation, the regression is in a form of $y = bf(x) + \varepsilon'$. As desired, the b retains in the new transformed data. We can test the significance of the b in the transformed data to test the dependency of concentration on temperature.

References

[1] Sun Y, Xie Z M, Li J, Xu J M, Chen Z L and Naidu R 2006 Assessment of toxicity of heavy metal contaminated soils by the toxicity characteristic leaching procedure *Environ. Geochem. Hlth.* **28** 73-8

[2] da Silva E F, Mlayah A, Gomes C, Noronha F, Charef A, Sequeira C, Esteves V and Marques A R F 2010 Heavy elements in the phosphorite from Kalaat Khasba mine (North-western Tunisia): Potential implications on the environment and human health *J. Hazard. Mater.* **182** 232-45

[3] Bissen F F H 2003 Arsenic – a review. Part 1: Occurrence, toxicity, speciation, mobility *Acta Hydrochim. Hydrobiol.* **31** 9-18

[4] Lu J and Yuan F 2019 The effect of temperature and precipitation on the leaching of contaminants from Ballangen tailings deposit, Norway (Accept) *WIT Transactions on Ecology and the Environment* **231** 75-89

[5] Taylor R 1990 Interpretation of the correlation-coefficient - a basic review *J. Diagn. Med. Sonog.* **6** 35-9

[6] Nagelkerke N J D 1991 A note on a general definition of the coefficient of determination *Biometrika* **78** 691-2

[7] Kendall M G, O'Hagan A and Forster J 2003 *Kendall's Advanced Theory of Statistics* 2nd ed.

ISBN: 9780470685693 (New Jersey, USA: John Wiley & Sons Ltd.)
[8] Martinez A, Riu J and Rius F X 2000 Lack of fit in linear regression considering errors in both axes *Chemometr. Intell. Lab.* **54** 61-73
[9] Miller F R and Neill J W 2016 Lack of fit tests for linear regression models with many predictor variables using minimal weighted maximal matchings *J. Multivariate Anal.* **150** 14-26
[10] Rose M S and Fick G H 1997 Assessment of lack of fit in simple linear regression: An application to serologic response to treatment for syphilis *Stat. Med.* **16** 373-84
[11] Wang D X and Conerly M D 2003 Evaluation of three lack of fit tests in linear regression models *J. Appl. Stat.* **30** 683-96
[12] Johnson R A and Wichern D W 2007 *Applied Multivariate Statistical Analysis* 6th ed. (N.J.: Pearson Prentice Hall, Upper Saddle River)
[13] Lehmann E L and Casella G 1998 *Theory of Point Estimation* 2nd ed. (New York: Springer)
[14] Freedman D, Pisani R and Purves R 2007 *Statistics* 4th ed. (New York: W W Norton & Co.)
[15] Rao C R 1973 *Linear Statistical Inference and Its Applications* 2d ed. (New York: J. Wiley)
[16] Myers R H 1990 *Classical and Modern Regression with Applications* 2nd ed. (Boston: PWS-KENT)
[17] Beckman R J 1973 *Introductory Engineering Statistics* 2nd Edi – I Guttman, S S Wilks and S Hunter *J. Am. Stat. Assoc.* **68** 247