Exploration of Heavy Metal Concentrations in Soil using Regression Analysis

[1] *Dr.G.Wiselin Jiji,*

[2]*Catherine Praiseye Vijayan,* [3]*Christy.G,* [4]*Alagumadathy.C,* [5]*Avadai Priya.L,* [6]*Anusha.P*

[1] **Principal, Dr.Sivanthi Aditanar College Of Engineering,Tiruchendur.**

[2,3,4,5]**UG Students,**[6]**PG Student, Dr.Sivanthi Aditanar College Of Engineering,Tiruchendur.**

ABSTRACT:

In this paper, we implemented and evaluated the proposed method for analysis of heavy metal concentration and classification using geotiff imagery (LANDSAT) data. Tiruppur district, province of Tamil Nadu, which is strongly polluted due to the increase in number of dyeing units in textile industries. This approach followed the standard scheme of processing: Soil samples collection, Image acquisition, Regression analysis and Classification. At first level, we collected the soil samples in 36 locations in Tiruppur District. In the second level, we studied the locations of the soil samples from geotiff imagery data. The heavy metal contamination in particular region of Tiruppur district was identified using quantitative analysis by Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES) and the result along with the LANDSAT images was calibrated using Regression Analysis and thus analyze the level of heavy metal concentration in the locations . Based on concentration, The region is clustered into highly polluted / acid polluted / not polluted.

KEYWORDS: Heavy metal concentration, Image acquisition, Regression Analysis, Classification.

INTRODUCTION:

Soil heavy metal contamination is a major environmental concern, and the ecological risk associated with heavy metals is increasing. In recent years, several studies have shown that the characteristics of the heavy metal contamination in urban soils and agricultural soils are different Heavy metals are found in most soils, and at high concentrations can cause various types of harm to humans, plants and animals. Most heavy metals are also enriched through the food chain, which can damage human health. Chronic exposure to Cd can have effects such as lung cancer, prostatic proliferative lesions, bone fractures, kidney dysfunction and hypertension. Chronic oral and

inhalation exposure to As can lead to skin lesions and lung cancer, respectively. Exposure to Pb may cause plumbism, anemia, nephropathy, gastrointestinal colic and central nervous system symptoms (Żukowska and Biziuk, 2008).

The high concentrations of heavy metals such as Cd in soil, the areas with low soil pH, and the presence of sensitive land uses such as agricultural and residential – are the elements that contribute to human health risk. The human health risk from heavy metal contamination is the study area is not only correlated with the soil metals concentrations, but is related to the overall environmental system. The land uses were considered when assessing the heavy metal health risks in this study. The size of the local population should also be considered in any further study, as with larger populations in a residential land there is a higher chance of adverse human health effects occurring. (Huarong Zhao et al.2012).

Manoj Kumar Tiwari [1] et.al suggested that the analysis of soil samples could be reached is that the concentrations of heavy metals in soil near to the dumping/disposal site is more and decreases as distance increases. Also in the depth wise analysis, it was observed that the in most of the soil samples, higher concentrations of selected heavy metals are observed near the surface of ground and magnesium which has highest concentration .The higher pH (alkaline) of the disposed industrial solid wastes may reduce the leach ate generation, so suitable alternatives can be used for the same at the time of industrial solid waste disposal or dumping near populated vicinity.

Chijioke Emenike[1] et.al suggested that the ultimate waste disposal option among most Asian nations is landfilling. Dumping of wastes in open places or landfills allows for generation of leach ate. Leach ate quality is influenced by the type of waste deposited at such disposal sites. This study had inferred that though the leach ate quality among Asian landfills may vary slightly, yet the characterized in-site compositions are similar. Hence, there is urgent need to revamp existing dumps and landfills, while at the same time sourcing for an alternative and greener waste disposal options that will reduce the use of landfill systems in Asia.

Nalawade P.M[1] et.al suggested that the assessment of the metal shows that the fly ash dumping ground and surrounding area contains availability of different heavy metals which can be well recover, reuse and recycled. The metal availability also suggested heavy metal contamination in the surrounding region of thermal power plant which is harmful to the human life living in that country. There is need to develop some technology for the eco-friendly bioremediation of metal pollutants, because it is one of the latest technique which is useful to remediate such kind of problem.

Akot[1]suggested that to determine total metals levels (Cu, Fe, Ni, Pb, Cd and Zn) in surface soils from the vicinity of a railway workshop. Soil pollution in the present study was assessed using enrichment factor and geo accumulation index values. The calculation of the enrichment values showed that Pb and Cu are enriched with 3.47 and 2.26 respectively. The results of geo accumulation index based on Muller's classification allows us to conclude that, for analyzed metals, the concentrations of Cd and Ni can be generally considered as background levels moderately polluted with Zn and highly to very highly polluted with Pb and Cu. Some of the elevated concentration of Pb, Cu and Zn are due to anthropogenic sources. The soils from the vicinity can be described as moderately to very highly polluted since all the samples collected showed a moderately high levels of contamination in at least one metal.

As suggested in Peters et al. (2003), large concentrations of heavy metals can be predicted using reflectance spectroscopy. However, it is much more difficult to determine small concentrations of heavy metals in soil as opposed to large concentrations in mines. The PLSR models for Pb, Zn , AS, Hg, Cr and Cd were also accurate in estimating the metal abundance for new (validation) samples. This research demonstrates the promise of using the PLSR modeling to determine Pb and other heavy metal contamination in urban areas (C. M. Pandit et al.2010).

This study integrates heavy metal content data, remote sensing images, topography data, other

ancillary data and GIS analyses for mapping the total Cu and Pb contents spatially. With 3D maps generated from multi-source data, the potential sources and transportation routes of Cu and Pb pollutants were inferred through visual inspections and field survey records. The inference was thereafter confirmed by mapping together the watersheds of the Dawu River, Jishui River and the mining area, the Cu/Pb content in soil, and sample sites. The areas in high risk of Cu or Pb contaminations were also suggested. The important roles that topography and streamflow play in the transportation of Cu and Pb easier understanding", this study demonstrated the arts of utilizing multi-source data in the mapping of environmental pollution as well as in the understanding of the role that topography plays in pollutant transport.

METHODOLOGY:

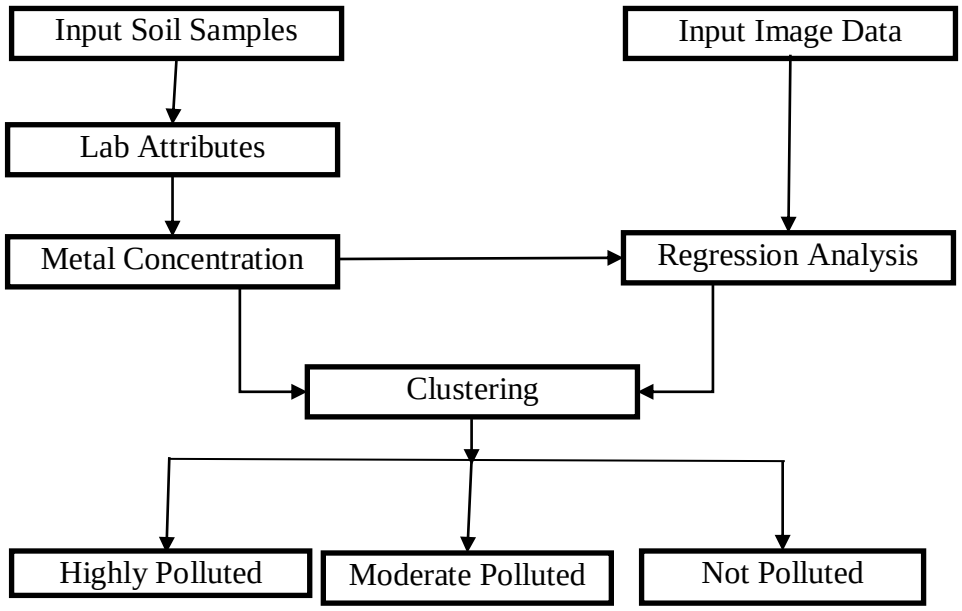

Fig 1 : Workflow of Proposed Method

The soil samples are collected and quantitative analysis by Inductively Coupled Plasma Atomic Emission Spectroscopy (**ICP**-AES) at IIT . The result of heavy metal concentration for 6 metals (AS, Cd, Hg, Zn, Pb and Cr) are obtained. Then, the geotiff images along with the heavy metal concentration are processed and calibrated using Partial Least Square Regression(PLSR) algorithm. From the PLSR algorithm we have obtained the results for 6 metals in 10 locations and the clustering is performed from this result. Here we use Hierarchical Clustering for classifing the values. The hierarchical clustering is performed based on the distance between the values and centroid of the cluster are obtained.Finally, the heavy metal concentration values are grouped into 3 clusters, based on the standard values for the 6 metals(AS, Cd, Hg, Zn, Pb and Cr) and the 3 clusters Highly polluted,Moderate polluted and not polluted are obtained for the 10 Locations for which the samples are analysed . Further, the Regression model is trained and the classification is done for other locations.

Partial Least-Squares Regression Modelling:-

Partial least-squares regression (PLSR) is a standard multivariate statistical technique that was developed by Herman Woldin 1966. The underlying assumption behind a PLSR model is that linear combinations of observed explanatory

variables, which are called components, factors, or latent variables, drive a system or process. PLSR is used to find eigenvectors of the explanatory variables that will produce score values. These score values must both summarize the variance of the explanatory variables well and highly correlate with the response variables. A simplified PLSR model consists of two outer relations and an inner relation.The two outer relations are a result of the Eigen structure decomposition of the X and Y matrices, respectively. The inner relation links the score for the X and Y matrices via a regression. X is an n x m matrix for the explanatory variables measured for n samples at m variables, and Y is an n x p matrix for the response variables measured for n samples at p variables. Decomposing X into a matrix T (the score matrix) and a matrix PT (the loadings matrix) plus an error matrix E derives the outer relation for the X matrix:

$$\mathbf{X} = \boldsymbol{TP^T + E}$$

In this way, Y is decomposed into the score matrix U and the loading matrix QT plus the error matrix F to create the outer relation for the Y matrix:

$$\mathbf{Y} = \boldsymbol{UQ^T + F}$$

T and P are sized n × k and m × k respectively; U and Q are sized n × l and p × l; k and l are the number of principal components for reconstructing X and Y; PT and QT represent the transpose of P and Q. Both E and F are residual matrices. E, the n×m residual matrix for the matrix X, results from using k principal components to summarize variance. F is the n × p residual matrix resulting from the subtraction of l principal components from the Y matrix.

The Purpose of PLSR is to minimize the norm of F while keeping the correlation between X and Y by the inner relation. B is an n × n matrix for the regression coefficient:

$$\mathbf{U = BT}$$

By deriving the orthogonal score matrix (T), PLSR remains unaffected by harmful collinearity in explanatory variables. The decomposition or reduction in the dimensionality of the explanatory variables allows the regression model to be composed of a small number of samples. PLSR has been applied to visible, near-infrared reflectance and mid-infrared reflectance to predict mineral abundance and biological activity in soil (Chang et al. 2005, Li 2006). A MATLAB program was implemented to conduct PLSR modeling on the samples. The spectral bands served as the explanatory variables, and six soil constituents (the 5 heavy metals and organic matter) were the response variables for PLSR modeling. X is an n × s matrix, and Y is a 6 × n matrix, with n as the number of samples and s as the number of spectral bands. Bands that are sensitive only to water content in the atmosphere were automatically determined by the program and left out during the regression, as they have no bearing on metal abundance predictions.

To create a PLSR model with high predictive power, an optimal number of PLSR components were selected for use in the model. Cross calibration of the PLSR model using an increasing number of components was utilized to determine the ideal number of components. One sample at a time was left out for the cross calibration of the response variables. For the n sets of reflectance data, a PLSR model was built using **n–1** spectra along with the corresponding abundance. The abundance of each metal and organic matter for the sample that was left out was predicted. This process was repeated n times until each sample had been left out once. The abundance predictions for the n soil samples were compared with the actual concentrations of the samples.

The resulting prediction error sum of squares (PRESS) was calculated:

$$\mathbf{PRESS} = \sum_{i=1}^{n} (\hat{y}_{(i)} - y_i)^2$$

where $\hat{y}_i$ is the estimate of a constituent value y_i for the left-out sample i.

Clustering (Hierarchical Clustering):

Hierarchical Clustering also known as *hierarchical cluster analysis*, is an algorithm that groups similar objects into groups called *clusters*. Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This continues until all the clusters are merged together. The metal content and the estimation results are made as input for the clustering and from that we are going to group the highly polluted, acid polluted and not polluted.

Hierarchical clustering can be performed with either a *distance matrix* or *raw data*. When raw

data is provided, the software will automatically compute a distance matrix in the background. The distance matrix below shows the distance between six objects.The *distance* between two clusters has been computed based on length of the straight line drawn from one cluster to another. The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for

the field of study. For example, if clustering crime sites in a city, city block distance may be appropriate (or, better yet, the time taken to travel between each location). Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in the physical world. After selecting a distance metric, it is necessary to determine from where distance is computed. As with *distance metrics*, the choice of linkage criteria should be made based on theoretical considerations from the domain of application. Where there are no clear theoretical justifications for choice of linkage criteria, *Ward's method* is the sensible default. This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster. This is often appropriate as this concept of distance matches the standard assumptions of how to compute differences between groups in statistics (e.g., *ANOVA*, *MANOVA*).

RESULT AND CONCLUSION:

The lab values for the soil heavy metal concentration are shown in table 1 as follows,

SAMPLE ID	As(mg/lit) (BDL)	Cd(mg/lit)	Hg(mg/lit) (BDL)	Zn(mg/lit)	Pb(mg/lit)	Cr(mg/lit)
S3	**-0.171**	**0.002**	**-2.054**	**3.997**	**0.068**	**0**
S5	**-0.232**	**0.000**	**-2.903**	**4.849**	**0.374**	**0**
S6	**-0.184**	**0.001**	**-2.833**	**3.281**	**0.263**	**0**
S16	**-0.227**	**0.004**	**-6.689**	**4.281**	**0.453**	**0**
S34	**-0.191**	**-0.005**	**-5.332**	**4.183**	**0.202**	**0**
S15	**-0.150**	**0.001**	**-16.66**	**29.38**	**0**	**2.157**
S18	**-0.071**	**0.002**	**-8.391**	**38.72**	**0**	**1.258**
S29	**-0.023**	**-0.001(BDL)**	**-6.362**	**0.174**	**0**	**1.008**
S22	**-0.063**	**0.010**	**-6.583**	**0.402**	**0**	**0.703**
S33	**-0.131**	**-0.002(BDL)**	**-7.188**	**33.67**	**0**	**0.988**

Table 1: Lab values of the Heavy Metal Concentration.

The heavy metal concentrations of the soil samples ranges from, Cd (-0.005 to 0.010mg/lit), Pb (0.068 to 0.453mg/lit), Hg(-2.054 to -16.66mg/lit), Cr(0.703 to 2.157mg/lit), As(-0.023 to -0.232 mg/lit) and Zn (0.174 to 38.72mg/lit) in the above 10 locations analyzed.

The result of regression algorithm for 6 metals in 10 locations are given below, In Location S3 , the regression generated for

Arsenic is -0.1428125,

Cadmium is 9.3749999999998E-4,

Mercury is -6.124062500000036,

Zinc is 10.699145833333432,

Lead is 0.13479166666666687,

Chromium is 0.5661458333333371

Similarly the matrix of locations and sample concentration regression generated as 10 6 matrix. From these values , the classification is done by Hierarchical Clustering.

The Hierarchical clustering is performed to cluster the values into 3 groups. Finally, the 3 groups are named as highly polluted, moderate polluted and not polluted based on the standard values for the 6 metals. In the figure 2, 3 clustered groups are differentiated with 3 colors (red, yellow, green).

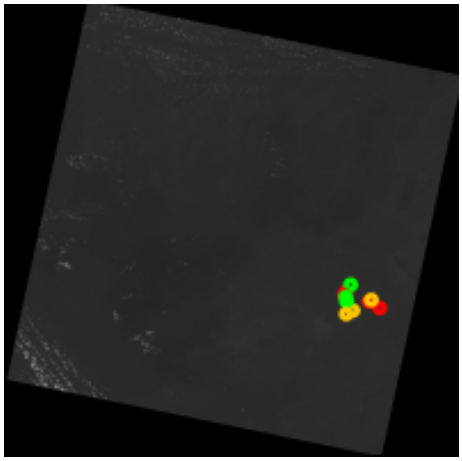

Fig 2: Clustered Regions

CONCLUSION:

This paper presents a new approach for soil heavy metal concentration based on the use of LANDSAT images. In this study, our major contributions is that affected areas is classified in to 3 groups based on the pollution standard values of the heavy metals. High concentrations of Cd, Pb, Cr, Hg and Zn were observed in the study area indicating that industrial activities had effect on heavy metal distributions. Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal component regression instead of finding hyper planes of maximum variance between the response and independent variables. Clustering analysis can be considered as a quick and low-cost method. These approaches can be used for providing reliable information regarding the spatial distribution of soil heavy metals and their pollutant sources.

REFERENCES:-

[1] Kortegast A. P. et al. 2007, Leachate Generation and treatment at the Buk Tagar landfill, Malaysia, Sardinia 2007, Eleventh International Waste Management and Landfill Symposium.

[2] Chao et al., 2014, A review on heavy metal contamination in the soil worldwide: Situation, impact and remediation techniques, Environmental Skeptics and Critics, 2014, 3(2): 24-38.

[3] Mandal A. and Sengupta D., 2002, Characterisation of coal and fly ash from coalfired thermal power plant at Kolaghat—possible environmental hazards.

[4] T. Kemperand S. Sommer. ,“Mapping and monitoring of residual heavy metal contamination and acidification risk after the Aznalcóllar mining accident (Andalusia, Spain) using field and airborne hyperspectral data”.

[5] Pan, Ting-Tiao, Da-Wen Sun, Jun-Hu Cheng, and Hongbin Pu. "Regression algorithms in hyperspectral data analysis for meat quality detection and evaluation."

[6] Żukowska J, Biziuk M. Methodological evaluation of method for dietary heavy metal intake.

[7] Patz JA, Daszak P, Tabor GM, Aguirre AA, Pearl M, Epstein J, et al. Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence.

[8] Huarong Zhao , Beicheng Xia , Chen Fan , Peng Zhao , Shili Shen. Human health risk from soil heavy metal contamination under different land uses near Dabaoshan Mine, Southern China.

[9] Arman Nadari & Mohmmad Amir Delavar & Babak Kaboudin &Mohammad Sagdegh Askari. Assessment of spatial distribution of soil heavy metals using ANN-GA, MSLR and satellite imagery

[10] Yiyun Chen , Yaolin Liu , Yanfang Liu , Aiwen Lin , Xuesong Kong , Dianfeng Liu , Xiran Li , Yang Zhang , Yin Gao and Dun Wang. Mapping of Cu and Pb Contaminations in Soil Using Combined Geochemistry, Topography, and Remote Sensing: A Case Study in the Le’an River Floodplain, China.

[11] Thomas Kemper And Stefans Ommer. Estimate of Heavy Metal Contamination in Soils after a Mining Accident Using ReflectanceSpectroscopy.

[12] Saraswat P. K. and Chaudhary K., 2014, Effect of Fly Ash (FA) to improving soil quality and increase the efficiency of crop productivity.

[13] Cai, L.M., Huang, L.C., Zhou, Y.Z., Xu, Z.C., Peng, X.C., Yao, L.A., Zhou, Y., Peng, P.A.,2010. Heavy metal concentrations of agricultural soils and vegetables from Dongguan, Guangdong.

[14] Peters, D.C., Hauff, P.L., Peppin, W.A., Prosh, E.C., Borstad, G.A., Henderson Iii, F.B., Dillenbeck, E.D., Closs, L.G., Lee, K., Thiros, J.L., Coulter, D.W., Robbins,D.A. And Wangerud, K.W., 2001, Waste identification and characterization for the international smelter area using CASI/SFSI hyperspectral remote sensing data.

[15] Caaminee M. Pandit, Gabriel M. Filippelli And Lin Li.,"Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression".